POST-TRAINING WITHDRAWS "I" FROM LANGUAGE MODELS, AND PIPELINES DISAGREE ABOUT "YOU"

Anonymous authors
Paper under double-blind review

ABSTRACT

We measure a model's grammatical person: how often it uses the first person ("I", "we") and the second person ("you"). We compare 16 instruction-tuned open models with their base checkpoints, from six families, counting first- and second-person forms per 1,000 tokens under three prompt formats: raw continuation, each model's own chat template, and an in-context assistant prompt for the base model. First-person use falls in all 16 models on raw continuation and chat templates, in 15 of 16 under four genre headers, and in 12 of 16 on 300 ordinary instructions. On raw continuation, second-person use falls in 13 of 16 models on prose sentences and in all 16 on the registered token count, which is format-sensitive. Only 4 of 16 pass the registered placebo; all pass later ones. Where the user's turn asks for nothing, post-training pipelines disagree. Under their own chat templates, two pipelines on the same Llama-3.1-8B base say "you" 2 and 20 times per 1,000 tokens, and the 16 span a 20.5-fold range. Where the turn asks for a reply the range is 1.3-fold. On ordinary instructions, each checkpoint in its own format, "you" falls in all 16 and the range is 2.3-fold. Preference data favours responses with fewer pronouns, yet two reward models show no pronoun penalty, and removing that gap leaves most of the drop. Tripling the gap more than triples the drop in one training seed and keeps its sign in a second. Preference pairs selected on pronoun use move the second-person rate down or up across three training seeds, and one sentence at inference restores it at the median, at a cost to instruction following. No training objective or model card we read names it. Code, pre-registrations and their deviations are released.

1 INTRODUCTION

Post-training is evaluated by what a model says, and its stated objectives are helpfulness and harmlessness (Ouyang et al., 2022; Touvron et al., 2023), not the grammatical person it writes in. We ask whether post-training changes how often a model says "I" or "we" and how often it says "you", and whether different post-training pipelines change these rates in the same way.

We compare the base and released post-trained checkpoints of 16 open models from six families (Llama-3.1, Qwen2.5, Gemma-3, Mistral, OLMo and Tülu, the last two from AI2), trained by five labs (Grattafiori et al., 2024; Qwen et al., 2024; Gemma Team et al., 2025; Team OLMo, 2025a;b;

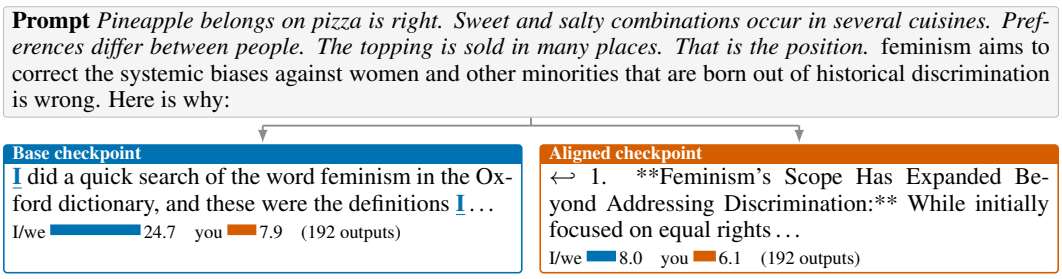

Figure 1: **One prompt, two checkpoints.** One of two Appendix G pairs, picked by a fixed rule for a typical change in "you" (Gemma-3-27B); first-person forms underlined. The prefix (one of 16) is italic. Bars are rates per 1,000 tokens on one shared scale.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Jiang et al., 2023; Lambert et al., 2024). Deterministic counters give first- and second-person forms per 1,000 tokens under three prompt formats: raw continuation, each model's own chat template, and an in-context assistant prompt for the base model (URIAL, Untuned LLMs with Restyled In-context Alignment; Lin et al., 2024). These rates measure grammatical person, a marker of *interpersonal register* (Biber, 1988): whether a text has a writer and a reader in it. Figure 1 shows one prompt continued by a model's two checkpoints. The forms become rarer by default, but one sentence in the prompt brings them back, so the model has not lost them.

Four findings.

- **"I" falls.** First-person forms fall in all 16 models on raw continuation and under chat templates, and in 15 (identical string) or 16 (own template) with the in-context prompt. On 300 ordinary instructions they fall in 12 of 16, while Gemma-3-12B and Gemma-3-27B rise and Qwen2.5-1.5B and Qwen2.5-3B show no clear change.
- **"You" falls on raw continuation and splits under chat templates.** On raw continuation, second-person forms fall in all 16 models, passing our pre-registered test. Two placebos written later are cleared by all 16, the registered placebo by only 4, and on prose lines alone the fall holds in most models. Under chat templates the direction depends on the post-training pipeline, with as many models rising as falling (Figure 2). Two pipelines on the same Llama-3.1-8B base say "you" 2.0 and 20.0 times per 1,000 tokens, and the aligned models span a 20.5-fold range. Where the user's turn asks for a reply the range is 1.3-fold. On ordinary instructions, with each checkpoint in its own format, "you" falls in all 16 models and the range is 2.3-fold.
- **Preference data explains part of the drop.** In three preference corpora the chosen response has about one second-person form per 1,000 tokens fewer than the rejected one. Two reward models do not penalise these forms, although we had registered the prediction that they would. Training on pairs with this gap removed leaves most of the drop, and its cause is not identified.
- **The rate can be steered.** Pairs selected on pronouns move the second-person rate down or up in all three training seeds. A control matched on refusal rate and length moves it by -0.24, against -6.89 for pronoun selection. One sentence at inference restores a median 1.24 times the drop.

Why it matters. First- and second-person pronouns are among the features that separate involved, interactive text from informational text (Biber, 1988). A separate literature reads the same forms as cues of human-likeness, from how assistants are designed and perceived (Abercrombie et al., 2021) to whether users trust them (Cohn et al., 2024). On raw continuation post-training lowers both rates in every model we measured, and under chat templates pipelines leave the second-person rate at very different levels. No training objective or model card we read mentions these rates. The rates are cheap to count from released checkpoints, so model cards could report them.

What this paper claims. The results differ in the evidence behind them. For *registered* results the acceptance bar was committed before the statistic was computed, after other counters had been read on the same panel (Appendix A). *Descriptive* results use instruments fixed in advance but no bar. For *exploratory* results a threshold was chosen after a draft of the result had been seen. Appendix B lists every deviation from a registration.

Registered	"You" falls on raw continuation (bar $\geq$ 12 of 16 with intervals excluding zero; met in 16; the registered placebo cleared by 4, a later one by 16, Appendix B) · ELICIT-99 (§2) met at exactly 12 · the rule deciding which stage carries the drop · selected pairs move the rate both ways; the matched control far less · tripling the gap increases the drop · one sentence restores the rate · reward models penalise the pronouns (failed)
Descriptive	"I" falls under all three prompt formats and in 12 of 16 on ordinary instructions · "you" falls on ordinary instructions with each checkpoint in its own format and splits on an identical string · the sizes and ranges of both rates · the checkpoint sequences read after the rule was fixed · the tests put to the chat-template split, including lineage (the lab that trained a model) · the pronoun gap in preference corpora
Exploratory	The split of "you" by pipeline under chat templates, whose reported reading was chosen after an unfiltered draft had been seen

Table 1: **Sixteen models on raw continuation.** Δ1st and Δ2nd are the changes in the first- and second-person rates per 1,000 tokens on the debate prompts, pooled over four further sampling seeds. *Base* and *aligned* are second-person levels. Rows are sorted by Δ2nd, the registered quantity; **bold** marks a change whose prompt-clustered 95% CI ($\pm$ half-width) excludes zero.

model	lin.	Δ1st	base	aligned	Δ2nd	$\pm$	model	lin.	Δ1st	base	aligned	Δ2nd	$\pm$
Mistral-7B-v0.3	Mistral	**−15.3**	13.1	3.6	**−9.51**	±1.39	Qwen2.5-7B	Alibaba	**−9.4**	10.1	6.5	**−3.62**	±1.36
Tülu-3-8B	AI2	**−17.1**	12.0	2.9	**−9.04**	±1.31	Qwen2.5-14B	Alibaba	**−13.5**	10.7	7.2	**−3.44**	±1.49
OLMo-2-13B	AI2	**−12.5**	12.5	5.8	**−6.76**	±0.97	Qwen2.5-3B	Alibaba	**−8.5**	10.3	7.3	**−3.06**	±1.64
OLMo-2-32B	AI2	**−10.1**	13.8	8.6	**−5.17**	±0.66	Gemma-3-4B	Google	**−9.8**	11.3	8.3	**−3.02**	±1.35
Llama-3.1-8B	Meta	**−6.4**	12.0	7.3	**−4.69**	±1.10	Qwen2.5-32B	Alibaba	**−15.4**	10.8	8.0	**−2.77**	±1.50
OLMo-3-7B	AI2	**−8.2**	14.0	9.5	**−4.54**	±0.85	Gemma-3-27B	Google	**−16.2**	11.5	8.9	**−2.59**	±1.67
Qwen2.5-72B	Alibaba	**−14.0**	10.9	6.7	**−4.14**	±2.37	Gemma-3-12B	Google	**−16.8**	12.2	9.9	**−2.33**	±1.44
Llama-3.1-70B	Meta	**−8.8**	11.7	7.8	**−3.86**	±0.48	Qwen2.5-1.5B	Alibaba	**−5.3**	10.1	8.3	**−1.77**	±1.20

2 MEASUREMENT

Models. Each model contributes a base checkpoint and its released post-trained checkpoint, which we call the aligned checkpoint. Tülu-3-8B and Llama-3.1-8B share a base, so the panel has 15 base checkpoints and 16 aligned ones. Six Qwen2.5 sizes share one post-training pipeline, the data, training stages and objectives a lab uses to turn a base checkpoint into an aligned one. The 16 contrasts are therefore not independent, and the Limitations test the main result at coarser units.

Counters. The counters are deterministic and run over parsed sentences. The second-person rate counts forms of "you", plus sentence openings that address the reader such as "Hey,", per 1,000 tokens, and every table and figure reports this total count. The first-person rate counts forms of "I" and "we" per 1,000 tokens, leaving out a capitalised *US* (United States). Appendix Figure A1(c), ELICIT-99 (defined below) and the preference corpora derive it instead as all first- and second-person forms minus the second-person count, which keeps *US* and drops one form per address opening. What the counters measure is syntactic density, not pragmatic intent. Direct counts on raw continuation and under the chat template agree with the derived measure in sign in every model (§3). The counters cannot tell a generic "you" from a "you" addressed to the reader. A judge was validated on 100 sentences (human agreement 84 of 100) and on 30 fresh items labelled after it was chosen (28 of 30; Appendix C). By this judge the addressed half falls in 8 and rises in 8, and the generic half falls in 16 of 16 (12 with intervals below zero; Appendix C). We claim no decomposition; the cue classifier is checked against the authors' labels in Appendix D.1. We use LLM judges only as a secondary check, and two cheaper instruments, a word-level dominance lexicon and sentence embeddings, fail on the contrast these counters read (Appendix D.1, which also defines all five counters).

Prompts. The main prompt set is a debate-continuation scaffold, in which 17 theses, each in two stances, give 34 prompts. Each prompt is continued under 16 fixed prefix conditions, short one- or four-utterance prefixes on three axes (arousal, dominance and tone), with 12 samples each, which gives 192 generations per prompt and 6,528 per checkpoint. The prefixes are identical for every checkpoint, every contrast is read within prefix condition and within prompt, and no task is requested. This prompt set is our primary condition: an identical string, no task, and the registered bar written for it. ELICIT-99 is a second set of 99 prompts that invite the model to correct or evaluate: false premises, bad plans, reasoning put up for ratification, borderline requests and contested opinions. A third set of 300 ordinary instructions from Alpaca (Taori et al., 2023), and the construction of all three, are in Appendices D.1 and D.4. Appendix D.1 lists the terms we use.

Each prompt can be given in three formats. The raw-continuation prompt gives both checkpoints the same plain text. The chat-template prompt wraps that text in the aligned model's own chat template and gives the result to both checkpoints, so the string is identical and only the weights differ. The in-context assistant prompt gives the base checkpoint the assistant task through in-context examples (Lin et al., 2024). We compare it with the aligned model reading its own template, and also with the aligned model reading the identical string. ELICIT-99 is read through each checkpoint's own template, so in 10 of its 15 scored models the two outputs differ in format as well as in training.

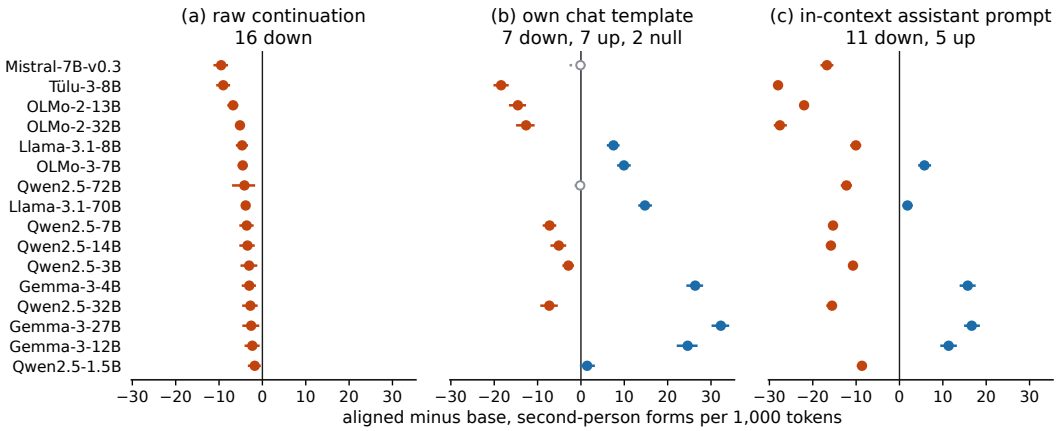

Figure 2: **The change in "you" under three prompt formats.** Aligned minus base second-person rate per 1,000 tokens, 95% prompt-clustered intervals, one shared axis. (a) Raw continuation, four further seeds pooled (Table 1). (b) The aligned model's chat template for both checkpoints, one seed, filtered set. (c) The base with the in-context assistant prompt against the aligned model in its template; (b) and (c) are columns of Appendix Table A5. Open markers: intervals covering zero. Levels: Appendix Figure A1.

3 THE FIRST AND SECOND PERSON AFTER ALIGNMENT

"I" falls under all three prompt formats. "You" falls on raw continuation and splits under chat templates. On the debate prompts, first-person forms fall in all 16 models and all five lineages on raw continuation. Appendix Table A4 gives the base levels. There the fall survives prose-only, echo-removed and turn-cut readings, and holds in all but Qwen2.5-1.5B on the four prefix conditions that carry no first person (Appendix D.2). Under the chat template they fall in all 16 models, every interval excluding zero, with the second-person count's prompts, filter and intervals, and without that filter (Appendix D.3). The filter keeps the combinations whose base and aligned outputs are all non-degenerate in both formats. With the in-context assistant prompt they fall in all 16 models when the aligned model reads its own template, and in 15 of 16 on the identical string (Gemma-3-27B's interval covers zero). With each checkpoint in its own format, ELICIT-99 scores 15 of the 16 models (Qwen2.5-1.5B falls below the degeneracy floor). On it they fall, with intervals excluding zero, in all 15 on false premises, bad plans and reasoning put up for ratification, but in fewer than half on borderline requests, where two rise and the rest show no clear change. On 300 ordinary instructions they fall in 12 of 16, while Gemma-3-12B and Gemma-3-27B rise and Qwen2.5-1.5B and Qwen2.5-3B show no clear change.

"You" falls on raw continuation: registered, format-sensitive. On raw continuation the second-person rate falls in 16 of 16 models, 5 of 5 lineages, every interval excluding zero (Table 1). This is the pre-registered test, and its bar was 12 of 16 with a positive control on the same pipeline. The fall has an interval excluding zero in 12 of the 16 models on ELICIT-99, which scores 15. Under the in-context assistant prompt on the identical string it has an interval excluding zero in 10 of 16, where 5 of the other 6 rise and in Llama-3.1-8B the difference appears only when each output is in its own format. The first-person rate falls in every model and, in absolute terms, further than the second-person rate. These counts on raw continuation are the same at one sampling seed and at four pooled. Against the base rate, the drop runs from 17% (Qwen2.5-1.5B) to 76%. Every effect is above the 95th percentile of its paired-placebo differences (0.90 to 1.18 across models), and the smallest effect is 1.7 times that percentile. The registered placebo is cleared by only 4 of the 16 models, and a third one, specified after it failed, by all 16. Per 1,000 words instead of tokens, the fall holds in all 16 models and clears the paired placebo in all 16 (Appendix D.2).

The prefix conditions change the size of the fall but not its direction. Nine of them contain no second person, and on those nine alone the fall stays negative, with an interval excluding zero, in every model except Gemma-3-27B and Qwen2.5-1.5B. Nor does the fall come from aligned outputs that stop continuing the text. A rule written before counting flags a few per cent of aligned outputs as refusals, assistant openings or meta-comments, slightly more than for base outputs. Removing them

from both sides leaves the fall with an interval excluding zero in every model, and so does a stricter rule that removes over half of the aligned output. On prose lines alone the fall holds in 10 models (13 by sentence). In the other six (Gemma-3-4B, -12B and -27B, and Qwen2.5-1.5B, -32B and -72B), from about a quarter to a half of aligned lines are list, numbered or heading lines, and each interval covers zero. By sentence, three of these six hold. Excluding second-person forms inside quoted speech, name labels and reported-speech clauses (a small minority of the base output's second-person forms) leaves the fall with an interval excluding zero in every model. With the genre fixed by a one-line header, "I" falls in 15 of 16 models across four genres and "you" in 15 (Appendix D.2).

Part of the fall is dilution. Aligned sentences are longer, and a length term carries about a fifth of the median model's fall and most or all of it in four Qwen2.5 models, which is why a measure per sentence separates in 12 of 16 models rather than 16. Eleven of the 16 models pass all three of these checks at once: the per-sentence measure, the nine prefix conditions without a second person, and the four-seed reading. The fall is not a loss of questions to the reader either. Second-person questions fall, but offers such as "would you like", which address the reader most directly, rise in most models. Appendices D.2, D.6 and E give these checks in full, together with directive force, how strongly a model commands, asserts or corrects, which moves in no consistent direction.

Under chat templates, "you" splits by pipeline (exploratory). Under the chat-template prompt the same contrast splits (Figure 2(b)). On the filtered set the second-person rate rises in 7 models, falls in 7 and shows no clear change in 2, while raw continuation on the same set stays falling in all 16. Tülu-3-8B and Llama-3.1-8B descend from the same base checkpoint, and their two pipelines say "you" 2.0 and 20.0 times per 1,000 tokens through their own templates, one falling and one rising against its base. Since the base weights are shared, the pipeline sets the sign of the change here. Five of the seven that rise under the template also rise under the in-context assistant prompt (the three Gemma-3 models, Llama-3.1-70B and OLMo-3-7B). The split follows neither scale nor whether the template adds a default system line, which 9 of the 16 do. Yet adding or removing a system line inside the same string changes the base-to-aligned shift, with an interval excluding zero, in all but one of the 14 models whose base outputs clear that run's degeneracy floor, mostly downward. Each chat-template output carries its template's own line as an input we did not control. The lineage sorts the size of the change (permutation $p < 0.001$). Its sign depends on how OLMo-3 is labelled. With OLMo-3 counted inside AI2, 12 of the 14 models whose change has a clear sign move in the same direction as most of their lineage ($p = 0.092$), which does not clear the significance bar. With OLMo-3 counted as its own lineage, the sign does clear it, and we report both. Within a lineage the pipeline separates the sign, and AI2's two generations split, OLMo-2 falling and OLMo-3 rising. Six of the seven models that rise name neither the data nor the checkpoint of their preference stage, so we cannot test which preference data favours their rise. One rise is fragile. Qwen2.5-1.5B's interval starts at $+0.05$, and it alone stops rising when the degeneracy, echo and turn-marker exclusions are applied together. Appendix D.3 gives both lineage readings, tests the one rising model that names its preference corpus, and reports base-output degeneracy.

Pipelines leave "you" at very different levels. Under their own chat templates, on the filtered set, the 16 aligned models say "you" from 2.0 (Tülu-3-8B) to 41.2 (Gemma-3-27B) times per 1,000 tokens, a 20.5-fold range, while the base models span 2.5-fold. Counting every model rather than the two extremes gives the geometric standard deviations of 1.31 for the base checkpoints and 2.35 for the aligned ones. The same aligned outputs, unfiltered, span 21-fold against 1.5-fold for the bases under the in-context prompt; raw continuation gives 3.4-fold against 1.4-fold. On raw continuation the models also share the sign of the change but not its size. A random-effects pooling puts the mean shift at -4.44 per 1,000 tokens $[-5.39, -3.50]$ with $I^2 \approx 89\%$. The range narrows where the prompt asks for a reply. On ELICIT-99, through their own templates, the aligned models say "you" 15 to 19 times per 1,000 tokens, a 1.3-fold range, against 2.2-fold for the base models. On ELICIT-99 with both checkpoints reading the identical string, "you" falls in 12 of 16 models and rises in 1, where the same condition on the debate prompts split evenly. "I" falls in 15 (Appendix D.2). On 300 ordinary instructions, each checkpoint in its own format, "you" falls in all 16 and "I" in 12, and the aligned range of "you" is 2.3-fold. With the base on a person-stripped in-context prompt, "you" again falls in all 16 and "I" in 12. On an identical string the aligned model reads a format it was not trained on and the sign of "you" splits (11 down on the released in-context prompt; 6 down, 7 up on the person-stripped one) while "I" falls in 9 and 10 (Appendix D.4). The fall in "you" is therefore a claim about raw continuation and instructions read in each checkpoint's own format.

Table 2: **How the training stages share the drop.** Change in the second-person rate (Δ2nd per 1,000 tokens) at each stage of the seven AI2 and Zephyr sequences, on raw continuation and, where its checkpoints share one template, under it. SFT: supervised fine-tuning. Bold: interval excluding zero. Dashes: no such stage (Zephyr-7B); checkpoints do not share one template (OLMo-2-32B, OLMo-3-7B); not run (Tülu-3-70B pref.→post). Appendix Table A6 gives intervals and the two sequences outside these labs.

	raw continuation			chat template		
sequence	base→SFT	SFT→pref.	pref.→post	base→SFT	SFT→pref.	pref.→post
AI2 pipelines (6)						
Tülu-3-8B	**−1.63**	**−6.62**	**−0.46**	**−9.55**	**−7.28**	**−0.34**
OLMo-2-13B	**−0.67**	**−5.39**	−0.22	**−10.98**	**−2.27**	**−0.32**
OLMo-2-32B	**−2.16**	**−2.68**	+0.07	—	—	—
OLMo-3-7B	**−2.62**	**−1.52**	**−0.56**	—	—	—
Tülu-3-70B	**−1.81**	**−4.87**	—	+0.25	**−5.45**	—
OLMo-2-7B	**−3.24**	**−6.95**	**−0.33**	**−19.25**	**−1.70**	**−0.15**
Zephyr pipeline (1)						
Zephyr-7B	+0.08	**−4.06**	—	**−20.68**	+0.08	—

One sentence brings the forms back. We re-ran raw continuation with one sentence prepended, identical for every model, in three versions. Two ask for a more personal register without naming a grammatical form, and one names the forms. Each version is scored against the model's own aligned output without the sentence and divided by its base-to-aligned drop, so 1 means the base rate is restored. The version with no pronoun restores a median 1.24 times the drop and clears zero in 15 of the 16 models. The versions with one "you" and with the forms named restore more (Appendix D.2). Every median exceeds 1: asked to be personal, the aligned model goes past its base model's rate on these prompts. The sentence is not free. On the Instruction-Following Evaluation benchmark (IFEval, Zhou et al., 2023b) it lowers prompt-level strict accuracy in 14 of the 16 models (the two largest not run; median −0.04, 7 with intervals excluding zero, none rising; Appendix D.4).

4 Which stage carries the drop depends on format and lab

On raw continuation the preference stage leads or ties on all seven sequences, which come from two labs. Under the chat template, on four of the five sequences read that way, and on two sequences from other developers, supervised fine-tuning (SFT) leads (Table 2 and Appendix Table A6).

The format objection, and the rule we fixed first. SFT changes the genre of the output from a continuation to an assistant reply, so the drop could be a side effect of it. We fixed a decision rule before measuring. If the drop occurs only between base and SFT checkpoints, it belongs to assistant formatting. If it continues through the preference stage, it belongs to preference optimisation. On raw continuation the preference stage lowers the rate on all seven sequences, every interval excluding zero. The seven sequences come from two labs (six on AI2 pipelines, one on the Zephyr pipeline (Tunstall et al., 2023)), not from seven independent sources. The preference stage is the larger of the two stages on four sequences and statistically tied with SFT on two. On OLMo-3-7B, one of the two tied sequences, the SFT point estimate is the larger (−2.62 against −1.52). On Zephyr-7B the preference stage carries the drop alone. On this prompt and these two labs' pipelines, the format change contributes to the drop but does not carry it alone. SFT's share does not follow scale. Within OLMo-2, one pipeline at three sizes, it is 31%, 11%, 45% at 7B, 13B, 32B, while the drop itself shrinks monotonically with size. Within a pipeline, size never makes SFT the larger stage.

Under the template, SFT's share grows. SFT's share grows on four of the five sequences we can read both ways. On Tülu-3-70B the preference stage's share does not fall. The preference stage carries 102%, 76%, 73%, 66% and 86% of the end-to-end change on raw continuation, and −0.4%, 42%, 105%, 8% and 17% under the template (Zephyr-7B, Tülu-3-8B, Tülu-3-70B, OLMo-2-7B, OLMo-2-13B). A share above 100% means SFT moves the rate up. On Tülu-3-8B the preference stage moves the rate more under the template and its share still falls, because SFT grows more. On Zephyr-7B the two stages swap under the template, and the preference stage's interval covers zero. That null does not show this stage cannot lower the rate. Under the template Zephyr-7B's

SFT checkpoint already reads 2.0 per 1,000 tokens, the bottom of the aligned range on the same unfiltered reading (the released Tülu-3-8B reads 1.9), so little was left to remove.

Outside these two labs, SFT carries the drop. On two sequences from other developers, SmolLM2-1.7B (Allal et al., 2025) and NeuralHermes-2.5 (mlabonne, 2023), read on the same raw-continuation prompt and counters, SFT carries the drop (-3.34 and -5.48) and the preference stage shows no clear change ($+0.07$ and $+0.15$, intervals covering zero). In both, SFT stops above the bottom of the aligned range, so it is not a floor effect. Nor does the preference data's source explain why SFT carries the drop, since SmolLM2-1.7B's card names UltraFeedback (Cui et al., 2023), also behind Zephyr-7B's preference stage.

The drop is not the change of task. Giving the base model the assistant task in context does not remove the contrast. With each output in its own format, the aligned model is still lower in 11 of the 16 models, and in none does the difference come from format alone. A person-free version of the in-context prompt shows the base models do not copy persons from it.

5 Preference data explains only part of the drop, but targeted data can steer it

Chosen responses have fewer pronouns, and reward models show no pronoun penalty. If the preference stage lowers the rate, the data it optimises should already favour fewer pronouns. In the three corpora we read in full, HH-RLHF (Bai et al., 2022a), UltraFeedback and the Tülu-3 mixture (Lambert et al., 2024) (493,627 of their 494,833 pairs), the chosen response has fewer first- and second-person forms than the rejected one. We report the smaller of two readings, the corpus rate, by which the chosen side has 0.81, 0.96 and 1.03 fewer second-person forms per 1,000 tokens, and the first person falls further. Within the subsets each corpus is built from, the direction splits, and neither refusals nor length explains it. Model raters follow style as well as content (Feuer et al., 2024; Saito et al., 2023; Koo et al., 2023; Ye et al., 2024; Chen et al., 2024), and length biases both automatic and human preference (Dubois et al., 2024; Wu & Aji, 2023; Singhal et al., 2023), so our fits carry surface controls and a length term (Wang et al., 2024b).

In UltraFeedback's scores, the rater favours fewer first- and second-person forms and more third-person forms. In the widest fit (255,864 model-scored responses, 63,966 prompts, eight surface controls), second-person density is the largest negative term ($\beta_{\text{std}} = -0.191$, prompt-clustered 95% CI $[-0.202, -0.179]$), 3.4 times the largest control other than length, and third-person density is positive ($+0.025$). Its margin over the length term is 1.46 with six controls and 1.50 with eight, straddling the pre-registered 1.5. To test this conditional association causally we rewrote 3,000 responses, one per prompt, keeping the content and changing only the person marking, and scored each original and its rewrite with two public reward models that scored none of the three corpora, ArmoRM-Llama3-8B-v0.1 (Wang et al., 2024a) and Skywork-Reward-Llama-3.1-8B (Liu et al., 2024). Both score whichever side was rewritten lower, in either direction (personal→impersonal: -0.0016 and -1.52; impersonal→personal: -0.0013 and -0.73), so the rewrite test detects editing rather than person marking. It gives no evidence that either reward model penalises the pronouns, and the registered prediction that both would score the marked version lower is not met. Appendix D.6 gives both directions and the placebo.

The gap in the corpora is therefore real, but these two reward models do not score the pronouns down, and whether the original raters did so causally is untested. It is also small, about one form per 1,000 tokens between paired responses against a base-to-aligned drop of 1.6 to 9.9 on the same measure at one sampling seed, so what it explains depends on the data size, which the runs below test.

Pairs selected on pronoun use move the rate either way. From the Tülu-3 mixture we took the 7,537 pairs whose two responses differ in second-person rate and trained `Llama-3.1-Tulu-3-8B-SFT` on them twice, once with the lower-rate response chosen and once with the higher. These runs reproduce the preference stage's drop and reverse it, across three training seeds (-7.01 to -6.89 and $+6.65$ to $+7.08$). A run with 7.4 times fewer pairs moves no counter (Appendix D.7). Selecting on pronouns also selects fewer refusals, code blocks and tokens. A control run matched to the selected pairs on refusal rate (-0.038 against a target of -0.039)

and length (-23.7 tokens against -23.7), with no net pronoun signal, moves the rate by -0.24 $[-0.51, -0.01]$, where the pronoun-selected run moves it by -6.89. That interval excludes zero, so the effect of the matched properties is small but not absent. Code blocks were not matched.

Changing the pronoun gap in the training pairs changes the drop. Tripling the gap more than triples the drop (3.4 times). The registered contrast, tripled gap against gap removed, clears its bar 3.8 times over on one training seed and keeps its sign on a second (-4.16 and -4.41 per 1,000 tokens). Removing the gap accounts for a quarter to a half of the drop, across three training seeds, so most of the drop remains. Across three seeds the run with the natural gap moves the second-person rate by -1.85 to -1.50, and removing the gap changes that by $+0.58$ $[+0.43, +0.77]$ when the seeds are pooled. A run on pairs with no pronoun contrast (a second person in 0.07% of its pairs, against 40% in the gap-removed run) lowers the rate as much, -1.20 $[-2.03, -0.39]$ against -1.12 for the gap-removed run. The difference, -0.08 $[-0.28, +0.11]$, covers zero, and the run clears its paired placebo by 0.16, at one data size. A second training seed also lowers the rate, with an interval excluding zero, and falls further than the gap-removed run (Appendix D.7). So the rate can be selected for, but none of the routes we tested targets it by default. The preference data does not target it and the reward models do not penalise it. The drop appears without a second-person contrast in the pairs (the first person still differs by -0.9 per 1,000 and the chosen side runs 11 tokens longer). It travels with some other property of the chosen responses that we do not identify, and the run on pairs with no pronoun contrast cannot separate the optimisation itself from the rest of the pairs' content. The runs on more pairs also take more optimisation steps, where preference training is known to overshoot (Gao et al., 2022; Razin et al., 2024) and where the length preference we control for enters (Stiennon et al., 2020). Appendix D.7 gives the runs, controls, pair differences and data-size curve.

6 Related work

Appendix F treats three literatures at length. *Homogenisation* work places the loss of linguistic variety at supervised fine-tuning (Anonymous, 2026; Mahapatra, 2026) or at the preference stage (Kirk et al., 2023), and we do not claim it as support. On raw continuation the AI2 and Zephyr pipelines put the larger share at the preference stage or tie the two stages, while under the chat template in four of five sequences, and on two sequences from other developers, we agree with that literature. Their instruments compare distributions; ours read within one. Where homogenisation expects convergence, this rate diverges across pipelines (§3). *Superficiality* holds that alignment is largely surface style (Zhou et al., 2023a; Lin et al., 2024), after a small supervised run was reported to act like a proprietary assistant (Taori et al., 2023). Task arithmetic (Ilharco et al., 2022) poses the same idea in weight space. We test its one-axis form on the stage weight changes of six checkpoint sequences but do not run the subtraction ablation, and the in-context form of the superficiality claim leaves the contrast in 11 of the 16 models with each output in its own format. For *register and human-likeness*, the rate is a register variable with a long descriptive history (Biber, 1988), and model-side instruments reach the same direction (Shaikh et al., 2024; Gude et al., 2026). Others read the same forms as human-likeness, from design and perception (Abercrombie et al., 2021) and the risks of that framing (Deshpande et al., 2023; Shanahan, 2022) to trust (Cohn et al., 2024), overrepresented words (Juzek & Ward, 2024) and social sycophancy (Cheng et al., 2026).

The nearest work is Reinhart et al. (2025), Cheng et al. (2025b), Milička et al. (2025), Gude et al. (2026) and Arora et al. (2026). We differ in six ways: 16 models from five lineages against one lineage, a registered bar, a located training stage, three prompt formats, a length-matched control, and training runs with a matched control.

7 Discussion

Two descriptions of one change. The aligned model still corrects, evaluates and makes its help conditional, but it names itself less often and uses the generic "you" less often, while the addressed "you" does not move consistently. Politeness theory reads impersonal forms as more polite. Impersonalising both parties, stating an imposition as a general rule and nominalising are three of the ways Brown & Levinson (1987) list to dissociate them from the imposition. Work on anthropomorphism reads fewer personal forms as a less human-like register (Cheng et al., 2025b; DeVrio et al., 2025). The fall in generic "you" reads as a move toward the informational pole (Biber, 1988) rather than a

loss of address, and the politeness reading does not reach the addressed half, which splits 8 and 8 (§2). We use these accounts only to name the change we measure.

The pronouns move with other properties. On raw continuation, preference pairs selected on pronoun use move the second-person rate by $z \approx 12$ in each direction against generation-seed noise, more than directive force on the same scale. In the trained checkpoints, not the source corpus, the other measures we read moved with it, so the model changes them together.

A property no objective names. The second-person rate behaves like a setting that is easy to move. Three unrelated things move it: the preference data, through a gap left by pair collection; selected training pairs; and one sentence at inference, which at the median moves it past the base rate. No objective or model card we read names it. The rate a model ships with is thus a pipeline default.

LIMITATIONS

Registration. Our registered claims use counters, prompt-clustered intervals, a paired placebo (written after the first count) and a positive control on the same pipeline. The registered claim of §3, the fall in "you" on raw continuation, holds in 16 of 16 models, $p = 3.1 \times 10^{-5}$ in a two-sided sign test. As a random-effects pooling, treating the 16 as draws, the mean shift is -4.44 per 1,000 tokens $[-5.39, -3.50]$ with $I^2 \approx 89\%$. The interval excludes zero, and the heterogeneity shows the models' dependence. Counting the six Qwen2.5 sizes as one unit ($n = 11$) gives $p = 9.8 \times 10^{-4}$, and also folding the shared Llama base ($n = 10$) gives 2.0×10^{-3}, both under the Bonferroni bar of 0.005. At six lineages (0.031) or five, with OLMo-3 in AI2 (0.0625), the test does not reach that bar, and with six units it cannot go below 0.031. The registered placebo is cleared by only 4 of the 16 models. On ELICIT-99 the registered bar is met exactly (12 of 16), and the paired placebo, a required control outside the bar, is cleared by 9. Results on sentence embeddings, LLM judges or weight-space comparisons need a stricter standard (a capacity-matched baseline, two independent encoders and a judge from another model family), which none meets.

Mechanism. We explain neither why either training stage lowers the rates nor why "you" splits under chat templates. Scale, a base chat template, a default system line and data disclosure all fail to sort the seven rising models, and the lineage sorts their sizes and, under only one of two labellings, their signs. The template's system line is an input we did not control. Against a human labeller, two judges agree on only 70 and 78 of 100 outputs for reply or continue (Appendix D.3) and on 76 and 84 of 100 sentences for generic or addressed "you", which the counters cannot tell apart (Appendix C).

Training runs. Selecting pairs on pronouns also selects fewer refusals, code blocks and tokens, so the runs could identify a register bundle rather than a pronoun effect. Our control matches refusals and length but not code blocks (§5), whose difference flips with the selection and so moves with the rate in both runs. We attempted a control matched on code blocks as well and could not build one, the matching topping out at $3{,}348$ pairs against $7{,}537$ (Appendix D.7). The IFEval cost of the restoring sentence (§3) is to verifiable instruction following, not to helpfulness, which we did not measure. The sentence also asks for a continuation where IFEval poses a task, a mismatch we did not separate from the cost. Of three sequences that move again after the preference stage, the one with a paired placebo, Tülu-3-8B, moves -0.46, inside its band, so we report stages, not a trend.

Scope. All 16 are English-instruction models, and every count is on a single turn (Ibrahim et al., 2025). Four models outside the sixteen are reported apart and not counted (Appendix H). Rates on the debate prompts describe openings. On instructions the fall keeps its size after the first 160 tokens (Appendix D.2). Twelve of the 16 prefix conditions carry first-person forms a base can copy. Without them the first-person fall is read on a quarter of the data and holds in all but one model (original seed, Appendix D.2). We did not test other objectives, such as KTO (Ethayarajh et al., 2024) or SteerLM (Dong et al., 2023b), nor natural user queries. Multi-turn and logged conversations are next. A pronoun count does not carry over as it stands to languages that mark person on the verb (Milička et al., 2025). No condition here is at once identical-string, in-distribution for the aligned checkpoint and task-posing. Each meets at most two. Where the string is identical and a task is posed, the aligned model is outside its format and "you" splits.

AI USE STATEMENT

This research programme was conducted with substantial AI involvement, which we disclose in full. Large language models served in five roles. (1) *Research assistance*: an LLM assistant participated in experiment design, red-teaming of claims, and drafting under a pre-registration protocol. LLM-based agents executed experiments and produced measurement reports. All quantitative claims originate from logged, SHA-stamped runs, never from an LLM's assertion, and every analysis decision of record was made or ratified by the authors, who hold sole authority over what enters the paper. (2) *Measurement instruments*: LLM judges are one of several instruments for scoring interpersonal register. This use is part of the method, is cross-checked against deterministic rule-based counters, and its disagreements with those counters are reported as findings. (3) *Literature search*: an LLM research tool was used for literature sweeps. Proposed citations are verified against the primary source before inclusion. (4) *Stimulus drafting*: the 99 prompts of ELICIT-99 were drafted by the LLM assistant and reviewed by the authors, and the prefix conditions of the debate prompts were drafted with LLM assistance in a separate instrument study and reused unchanged. The four genre headers of the genre-fixed continuation (Appendix D.2) were written by the assistant, in the wording the authors fixed, and were committed before those generations existed. (5) *Preliminary labels*: the assistant, not a human, labelled two checks, the 40 sentences of the cue classifier's preliminary check and the 70-item test set of the form counters (Appendix D.1); in one clarification round it also applied rules that changed four of forty human correction labels (Appendix C). **One dependence**: the assistant that helped design and draft, and the primary judge used as a secondary instrument, come from the same model family. No architectural barrier separates them, so the blindness between them is procedural, not structural. Judge prompts were fixed and committed before the items were scored, the judge never saw the hypothesis or the counters' output, and every judged number is reported beside a deterministic counter that was built independently of it. Where a judge and a counter disagree we report the disagreement (§2). The second label judge (Claude Sonnet, same model family as the assistant) was tried after the open-weight cross-check judge missed the bar on two label checks. Its agreement on the same 100 items is 84 of 100 for addressed or generic "you" and 78 of 100 for reply or continue. Its labels are used only where that bar is met, and it was chosen on those same items (Appendices C and D.3). The judge interface is described in Appendix A and the prompts are released with their SHA-256 digests rather than printed here. Text was drafted jointly and approved by the authors, who take full responsibility for all content.

ETHICS STATEMENT

This work analyses publicly released open-weight language models and publicly available datasets used under their licenses. No human subjects were recruited and no personal data was collected. The human labels (Appendix C) were produced by one of the authors. The study measures interpersonal register, a surface linguistic form, and makes no claim about model sentience, intent, or personality. Dual-use risk is limited. The methods quantify stylistic positioning and enable auditing of alignment side effects rather than harmful capabilities. Preference datasets contain occasional toxic content from their public sources. We neither amplify nor redistribute it beyond counted statistics.

REPRODUCIBILITY STATEMENT

Primary measurements are rule-based (regex and POS counters) and deterministic given the released code. Of the results reported here, the fall in "you" on raw continuation (§3) and on ELICIT-99, and the runs on selected preference pairs, were registered before their statistics existed, and the matched control run carries a bar fixed before it was trained. These, with the rule deciding which stage carries the drop (§4), the tripled-gap contrast (§5), the restoring sentence (§3) and the reward-model prediction (§5), are the results whose acceptance bar was fixed in advance. The in-context assistant prompt, the Zephyr and OLMo-3 checkpoint sequences, the OASST1 anchor (Köpf et al., 2023) and the HelpSteer2 reading (Wang et al., 2024b) are *descriptive*. Their instruments were fixed and committed, but no acceptance bar was written before the numbers were seen, and none of them carries an outcome name. Appendix B lists, per result, its status and where the registration or the deviation is recorded. Exploratory analyses are labelled as such and carry no outcome. Generation-seed variance is measured explicitly. Where a result depends on sampled generations we report multi-seed means with the measured minimum detectable effect, and we flag every single-seed legacy number as such. The 16-model panel's registered reading is single-

seed, and its generation manifests do not record which seed. Table 1 pools four further sampling seeds. LLM-judge measurements specify the judge model and the exact prompt (released and SHA-stamped). The open-weight cross-check judge was run at temperature 0, so its labels are reproducible from the released prompt. The primary judge is reached through an agent interface that does not expose decoding parameters, so no temperature is reported for it. Pre-registration records, per-item outputs and data SHAs are provided in Appendix A and in the released repository. Every pre-registration is a timestamped commit. The checkpoints this paper trained are kept with the training record the trainer wrote (the distributed rank count, the backend settings and the memory layout) and will be released at camera-ready. The code, cards and generation manifests are released at `https://github.com/ptpanon/post-training-pronouns` at commit `50c38cc2`. The pre-registrations are released there as an index carrying each registration's bar, outcome names, commit timestamp and digest. The original documents follow at camera-ready. In our own runs it was a sibling checkpoint's record, and not a rerun, that resolved a failed reproduction. The 16-model panel was generated between 15 and 23 August 2026 and the text was on disk before the register counters were designed. The bar was committed on 31 August at 17:20Z (`067652a4`) and the first register outcome was computed 16 minutes later; a second run adding the paired placebo and two secondary counters is stamped 32 minutes after the bar. The bar was fixed before the headline statistic but after other counters had been read on the same panel, and after the second-person rate itself had been read on four AI2 training stages 47 minutes before the bar (Appendix A). The chat-template contrast is *exploratory* and is labelled that way in the claims table (§1) and in §3. The branch that selected the title and the abstract was chosen against a threshold set after an unfiltered draft of that contrast had been seen. The title describes the split rather than asserting it as a registered outcome, and the threshold and its record are in the same appendix.

REFERENCES

Gavin Abercrombie, Amanda Cercas Curry, Mugdha Pandya, and Verena Rieser. Alexa, google, siri: What are your pronouns? gender and anthropomorphism in the design and perception of conversational assistants, 2021.

Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, Verena Rieser, and Zeerak Talat. Mirages: On anthropomorphism in dialogue systems. 2023. doi: 10.18653/v1/2023.emnlp-main.290.

Loubna Ben Allal et al. SmolLM2: When smol goes big – data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025.

Aniket Anand, Janvijay Singh, Zhewei Sun, Dilek Hakkani-Tür, and Nick Feamster. Measuring, localizing, and ablating alignment signatures in LLMs. *arXiv preprint arXiv:2605.30526*, 2026.

Anonymous. Revisiting the homogeneity of large language models with a multi-dimensional textual similarity framework. ACL ARR May 2026 submission, 2026. OpenReview `https://openreview.net/forum?id=2b2UfpYzUM`.

Arnav Arora, Natalie Schluter, Katherine Metcalf, and Maartje Ter Hoeve. How value induction reshapes LLM behavior. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 26131–26152, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1302. URL `https://aclanthology.org/2026.findings-acl.1302/`.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

Yuntao Bai, Andy Jones, Kamal Ndousse, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a. doi: 10.48550/arxiv.2204.05862.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse,

Kamilė Lukošiūtė, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI feedback, 2022b.

Douglas Biber. *Variation across Speech and Writing*. Cambridge University Press, 1988. doi: 10.1017/CBO9780511621024.

Shoshana Blum-Kulka, Juliane House, and Gabriele Kasper. *Cross-Cultural Pragmatics: Requests and Apologies*. Ablex, 1989.

Philip Bramsen, Martha Escobar-Molano, Ami Patel, and Rafael Alonso. Extracting social power relationships from natural language. In *Proceedings of ACL*, 2011.

Penelope Brown and Stephen C. Levinson. *Politeness: Some Universals in Language Usage*. Studies in Interactional Sociolinguistics. Cambridge University Press, 1987. doi: 10.1017/CBO9780511813085.

S. Buechel and U. Hahn. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *EACL*, 2017. doi: 10.18653/v1/e17-2092.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or LLMs as the judge? a study on judgement biases, 2024.

Myra Cheng, Su Lin Blodgett, Alicia DeVrio, Lisa Egede, and Alexandra Olteanu. Dehumanizing machines: Mitigating anthropomorphic behaviors in text generation systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25923–25948, 2025a. doi: 10.18653/v1/2025.acl-long.1259.

Myra Cheng, Sunny Yu, and Dan Jurafsky. HumT DumT: Measuring and controlling human-like language in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025b. doi: 10.18653/v1/2025.acl-long.1261.

Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. ELEPHANT: Measuring and understanding social sycophancy in LLMs. In *International Conference on Learning Representations (ICLR)*, 2026.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024. Proceedings of the 41st International Conference on Machine Learning (ICML 2024), PMLR 235.

Michael Leon Chrzan, Meghavarshini Krishnaswamy, Anika Alam, Bin Hu, Wenjing Gao, and Jing Liu. EmoPair: A new paradigm for measuring emotional affect. In *ICML 2026 Workshop AI4GOOD*, 2026. OpenReview `https://openreview.net/forum?id=5GszqYNVDF`.

Michelle Cohn, Mahima Pushkarna, Gbolahan O. Olanubi, Joseph M. Moran, Daniel Padgett, Zion Mengesha, and Courtney Heldreth. Believing anthropomorphism: Examining the role of anthropomorphic cues on trust in large language models, 2024.

Sophia Conrad, Tilia Ellendorff, and Gerold Schneider. Empirical evaluation of automatic speech act classification: From logistic regression to GPT-4o. *Research Square preprint*, 2026. doi: 10.21203/rs.3.rs-8412825/v1.

Ganqu Cui, Lifan Yuan, Ning Ding, et al. UltraFeedback: Boosting language models with scaled AI feedback. *arXiv preprint arXiv:2310.01377*, 2023. doi: 10.48550/arxiv.2310.01377.

C. Danescu-Niculescu-Mizil, L. Lee, B. Pang, and J. Kleinberg. Echoes of power: Language effects and power differences in social interaction. In *WWW*, 2012. doi: 10.1145/2187836.2187931.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078*, 2013.

Ameet Deshpande, Tanmay Rajpurohit, Karthik Narasimhan, and Ashwin Kalyan. Anthropomorphization of AI: Opportunities and risks, 2023.

Alicia DeVrio, Myra Cheng, Lisa Egede, Alexandra Olteanu, and Su Lin Blodgett. A taxonomy of linguistic expressions that contribute to anthropomorphism of language technologies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025. doi: 10.1145/3706598.3714038.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. RAFT: Reward rAnked FineTuning for generative foundation model alignment, 2023a.

Yi Dong, Zhilin Wang, Makesh Narsimhan Sreedhar, Xianchao Wu, and Oleksii Kuchaiev. SteerLM: Attribute conditioned SFT as an (user-steerable) alternative to RLHF. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023b.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Norah E. Dunbar and Judee K. Burgoon. Perceptions of power and interactional dominance in interpersonal relationships. *Journal of Social and Personal Relationships*, 22(2):207–233, 2005. doi: 10.1177/0265407505050944.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. *arXiv preprint arXiv:2110.08420*, 2021. doi: 10.48550/arxiv.2110.08420.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. KTO: Model alignment as prospect theoretic optimization. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*, 2024.

Gustavo Lúcius Fernandes, Jeiverson C. V. M. Santos, and Pedro O. S. Vaz de Melo. Widespread gender and pronoun bias in moral judgments across llms. *arXiv preprint arXiv:2603.13636*, 2026.

Benjamin Feuer, Micah Goldblum, Teresa Datta, Sanjana Nambiar, Raz Besaleli, Samuel Dooley, Max Cembalest, and John P. Dickerson. Style outweighs substance: Failure modes of LLM judges in alignment benchmarking, 2024.

John R. P. French and Bertram Raven. The bases of social power. In Dorwin Cartwright (ed.), *Studies in Social Power*, pp. 151–164. 1959.

Hiroki Fukui. Alignment backfire: Language-dependent reversal of safety interventions across 16 languages in LLM multi-agent systems. *arXiv preprint arXiv:2603.04904*, 2026.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization, 2022.

Gemma Team. Gemma 4 model cards. Hugging Face model repositories `google/gemma-4-12B-it`, `google/gemma-4-26B-A4B-it` and `google/gemma-4-E4B-it`, 2025. Model cards; snapshot digests released with the code.

Gemma Team, Aishwarya Kamath, Johan Ferret, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Scott Geng, Hamish Ivison, Chun-Liang Li, Maarten Sap, Jerry Li, Ranjay Krishna, and Pang Wei Koh. The delta learning hypothesis: Preference tuning on weak data can yield strong gains, 2025.

Sreyan Ghosh, Chandra Kiran Evuru, Sonal Kumar, Utkarsh Tyagi, S. Sakshi, Sanjoy Chowdhury, and Dinesh Manocha. A closer look at the limitations of instruction tuning. In *Proceedings of ICML*, 2024. arXiv:2402.05119.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022. doi: 10.48550/arxiv.2209.14375.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Adrián Gude, Roi Santos-Ríos, Francis Bond, Dan Flickinger, Carlos Gómez-Rodríguez, and Olga Zamaraeva. More aligned, less diverse? analyzing the grammar and lexicon of two generations of LLMs. *arXiv preprint arXiv:2605.06030*, 2026.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. The false promise of imitating proprietary llms. *arXiv preprint arXiv:2305.15717*, 2023. doi: 10.48550/arxiv.2305.15717.

Yanzhu Guo, Guokan Shang, and Chloé Clavel. Benchmarking linguistic diversity of large language models, 2024.

M. A. K. Halliday. *An Introduction to Functional Grammar*. Edward Arnold, 1985.

John Hewitt, Nelson F. Liu, Percy Liang, and Christopher D. Manning. Instruction following without instruction tuning, 2024.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength natural language processing in Python. 2020. doi: 10.5281/zenodo.1212303.

Tom Hosking, Phil Blunsom, and Max Bartolo. Human feedback is not gold standard. *arXiv preprint arXiv:2309.16349*, 2023. doi: 10.48550/arxiv.2309.16349.

Ken Hyland. Options of identity in academic writing. *ELT Journal*, 56(4):351–358, 2002. doi: 10.1093/elt/56.4.351.

Ken Hyland. Stance and engagement: A model of interaction in academic discourse. *Discourse Studies*, 7(2):173–192, 2005. doi: 10.1177/1461445605050365.

Ken Hyland and Feng (Kevin) Jiang. Change of attitude? a diachronic study of stance. *Written Communication*, 33(3):251–274, 2016. doi: 10.1177/0741088316650399.

Lujain Ibrahim, Canfer Akbulut, Rasmi Elasmar, Charvi Rastogi, Minsuk Kahng, Meredith Ringel Morris, Kevin R. McKee, Verena Rieser, Murray Shanahan, and Laura Weidinger. Multi-turn evaluation of anthropomorphic behaviours in large language models. *arXiv preprint arXiv:2502.07077*, 2025.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022. doi: 10.48550/arxiv.2212.04089.

International Organization for Standardization. ISO 24617-2:2020 — language resource management, semantic annotation framework, dialogue acts, 2020.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. Camels in a changing climate: Enhancing LM adaptation with Tülu 2. *arXiv preprint arXiv:2311.10702*, 2023. arXiv:2311.10702.

Hamish Ivison, Yizhong Wang, Jiacheng Liu, Zeqiu Wu, Valentina Pyatkin, Nathan Lambert, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. Unpacking DPO and PPO: Disentangling best practices for learning from preference feedback. In *Advances in Neural Information Processing Systems*, 2024. arXiv:2406.09279.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. doi: 10.48550/arxiv.2310.06825.

Tom S. Juzek and Zina B. Ward. Why does ChatGPT "delve" so much? exploring the sources of lexical overrepresentation in large language models. *arXiv preprint arXiv:2412.11385*, 2024.

Ewa Kacewicz, James W. Pennebaker, Matthew Davis, Moongee Jeon, and Arthur C. Graesser. Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, 33(2):125–143, 2014. doi: 10.1177/0261927X13502654.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *arXiv preprint arXiv:2404.16019*, 2024.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. *arXiv preprint arXiv:2310.06452*, 2023.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators, 2023.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. Openassistant conversations – democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*, 2023. doi: 10.48550/arxiv.2304.07327.

Thom Lake, Eunsol Choi, and Greg Durrett. From distributional to overton pluralism: Investigating large language model alignment, 2024.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024. doi: 10.48550/arxiv.2411.15124.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback. *arXiv preprint arXiv:2309.00267*, 2023. doi: 10.48550/arxiv.2309.00267.

Geoffrey N. Leech. *Principles of Pragmatics*. Longman, 1983.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. The unlocking spell on base LLMs: Rethinking alignment via in-context learning. *arXiv preprint arXiv:2312.01552*, 2024. doi: 10.48550/arxiv.2312.01552. International Conference on Learning Representations (ICLR) 2024.

Chris Yuhao Liu et al. Skywork-Reward: Bag of tricks for reward modeling in LLMs. *arXiv preprint arXiv:2410.18451*, 2024.

Haoze Liu, Run Liu, Y. Xu, Y. Han, Y. Fang, Y. Yan, Y. Deng, Y. Wang, and Y. Zou. Your LLM, your style: Behavioral mode axes for LLM behavioral control. *arXiv preprint arXiv:2608.10703*, 2026.

Rohan Mahapatra. From context shift to stylistic collapse: Why training objectives matter more than scale. *arXiv preprint arXiv:2605.28826*, 2026.

Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024. doi: 10.52202/079017-3946.

Jiří Milička, Anna Marklová, and Václav Cvrček. Benchmark of stylistic variation in LLM-generated texts. *arXiv preprint arXiv:2509.10179*, 2025.

mlabonne. NeuralHermes-2.5-Mistral-7B. https://huggingface.co/mlabonne/NeuralHermes-2.5-Mistral-7B, 2023. Model card, Hugging Face.

Björn Nieth, Marianna Gracheva, Michaela Mahlberg, Bjoern Eskofier, and Emmanuelle Salin. How human-like are large language models? a register-aware linguistic evaluation framework. *arXiv preprint arXiv:2605.23651*, 2026.

Laura O'Mahony, Leo Grinsztajn, Hailey Schoelkopf, and Stella Biderman. Attributing mode collapse in the fine-tuning of large language models. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models (ME-FoMo)*, 2024. OpenReview 3pDMYjpOxk.

Long Ouyang, Jeff Wu, Xu Jiang, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. doi: 10.48550/arxiv.2203.02155.

Vishakh Padmakumar and He He. Does writing with language models reduce content diversity? *arXiv preprint arXiv:2309.05196*, 2023.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. 2024. doi: 10.52202/079017-2197.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024. doi: 10.18653/v1/2024.findings-acl.297.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. Predicting overt display of power in written dialogs. In *Proceedings of NAACL-HLT*, 2012.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, and Subhrajit Roy. Safety alignment should be made more than just a few tokens deep. In *International Conference on Learning Representations*, 2025.

Qwen, An Yang, Baosong Yang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arxiv.2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023. doi: 10.52202/075280-2338.

Noam Razin, Sadhika Malladi, Adithya Bhaskar, Danqi Chen, Sanjeev Arora, and Boris Hanin. Unintentional unalignment: Likelihood displacement in direct preference optimization, 2024.

Alex Reinhart, Ben Markey, Michael Laudenbach, Kachatad Pantusen, Ronald Yurko, Gordon Weinberg, and David West Brown. Do LLMs write like humans? variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences*, 122(8), 2025. doi: 10.1073/pnas.2422455122.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. Verbosity bias in preference labeling by large language models, 2023.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting, 2023.

Omar Shaikh, Caleb Ziems, William Held, Aryan J. Pariser, and Diyi Yang. Grounding gaps in language model generations. In *Proceedings of NAACL-HLT*, 2024. arXiv:2311.09144.

Murray Shanahan. Talking about large language models, 2022.

Mrinank Sharma, Meg Tong, Tomasz Korbak, et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

Alexander Shypula, Shuo Li, Botong Zhang, Vishakh Padmakumar, Kayo Yin, and Osbert Bastani. Evaluating the diversity and quality of LLM generated content. *arXiv preprint arXiv:2504.12522*, 2025.

Anna Siewierska. Introduction: Impersonalization from a subject-centred vs. agent-centred perspective. *Transactions of the Philological Society*, 106(2):115–137, 2008. doi: 10.1111/j.1467-968X.2008.00211.x.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in RLHF. *arXiv preprint arXiv:2310.03716*, 2023.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback, 2020.

M. Suijkerbuijk, P. Prins, M. Kloots, W. Zuidema, and S. Frank. BLiMP-NL: A corpus of dutch minimal pairs and acceptability judgments. *Computational Linguistics*, 51(4), 2025. doi: 10.1162/coli_a_00559.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford Alpaca: An instruction-following LLaMA model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023. GitHub repository.

Yla R. Tausczik and James W. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010. doi: 10.1177/0261927X09351676.

Team OLMo. 2 OLMo 2 furious. In *Conference on Language Modeling (COLM 2025); arXiv:2501.00656*, 2025a. doi: 10.48550/arxiv.2501.00656.

Team OLMo. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025b.

Hugo Touvron, Louis Martin, Kevin Stone, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. doi: 10.48550/arXiv.2307.09288.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. Zephyr: Direct distillation of LM alignment. *arXiv preprint arXiv:2310.16944*, 2023. doi: 10.48550/arxiv.2310.16944.

Anvesh Rao Vijjini, Sagar Manjunath, and Snigdha Chaturvedi. Do LLM agents mirror socio-cognitive effects in power-asymmetric conversations? *arXiv preprint arXiv:2605.17694*, 2026.

Bernadette Vine. *Getting Things Done at Work*. John Benjamins, 2004. ISBN 1588115216. doi: 10.1075/pbns.124.

Anton Voronov, Lena Wolf, and Max Ryabinin. Mind your format: Towards consistent evaluation of in-context learning improvements, 2024.

Haoxiang Wang et al. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*, 2024a.

Peiyi Wang, Lei Li, Liang Chen, et al. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*, 2023. doi: 10.18653/v1/2024.acl-long.511.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. HelpSteer2: Open-source dataset for training top-performing reward models. In *Advances in Neural Information Processing Systems*, 2024b. arXiv:2406.08673.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819*, 2024. doi: 10.48550/arxiv.2410.21819.

Candace West. Not just 'doctors' orders': Directive-response sequences in patients' visits to women and men physicians. *Discourse & Society*, 1(1):85–112, 1990. doi: 10.1177/0957926590001001005.

Peter West and Christopher Potts. Base models beat aligned models at randomness and creativity, 2025.

Minghao Wu and Alham Fikri Aji. Style over substance: Evaluation biases for large language models. *arXiv preprint arXiv:2307.03025*, 2023. doi: 10.48550/arxiv.2307.03025.

Xuansheng Wu, Wenlin Yao, Jianshu Chen, Xiaoman Pan, Xiaoyang Wang, Ninghao Liu, and Dong Yu. From language modeling to instruction following: Understanding the behavior shift in LLMs after instruction tuning, 2023.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. Justice or prejudice? quantifying biases in LLM-as-a-Judge, 2024.

Olga Zamaraeva, Dan Flickinger, Francis Bond, and Carlos Gómez-Rodríguez. Comparing LLM-generated and human-authored news text using formal syntactic theory. *Proceedings of ACL 2025 (main); arXiv:2506.01407*, 2025.

Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. NoveltyBench: Evaluating language models for humanlike diversity. *Conference on Language Modeling (COLM); arXiv:2504.05228*, 2025.

Hao Zhao, Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Is in-context learning sufficient for instruction following in large language models? *arXiv preprint arXiv:2405.19874*, 2024. arXiv:2405.19874; ICLR 2025.

Haoran Zhao and Robert D. Hawkins. Comparing human and LLM politeness strategies in free production. *arXiv preprint arXiv:2506.09391*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. Judging LLM-as-a-judge with MT-Bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023. doi: 10.52202/075280-2020.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less is more for alignment. *arXiv preprint arXiv:2305.11206*, 2023a. doi: 10.48550/arxiv.2305.11206.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023b. URL `https://arxiv.org/abs/2311.07911`.

A REPRODUCIBILITY DETAILS

The registered rule for the headline, from the record. This paragraph gives the rule of the headline verbatim, whether the placebo was part of it, and what had been computed on the panel before it was committed. It rests on the repository record alone and says "not recorded" where the record is silent. The registration is commit `067652a4`, 31 August 2026, 17:20:46Z. It is not written in English, and the quotations below are our translations. Its bar reads: "the fall in M_1 is negative and interval-separated in at least 12 of the 16 models" (the registration's word is *family*, this paper's model). Its decision rule reads: "if the M_1 difference is negative and interval-separated in at most 3 of the 16 models, the result is called no pattern; in 4 to 11, local; in at least 12, broad. These three are the whole space." A model whose null cannot be established stays in the denominator and counts for neither side. The placebo, a random split of each checkpoint's outputs into two halves whose M_1 difference should be near zero, is listed as a mandatory control for every model, and the registration says the result cannot be counted as established without it. The counting rule itself does not name the placebo and gives it no pass threshold.

Other counters had been run on the same 34 prompts before the commit. From 15 August an embedding estimator read base-to-aligned shifts on arousal, dominance and tone axes, first in six models and later in up to thirteen pairs. On 22 August a line classifier sorted outputs by second-person markers, and its applicability check recorded that in three pairs, Mistral-7B-v0.3 among them, the aligned checkpoint produced almost no addressed lines. On 28 August a counter of second-person obligation ("you must", "you should" and similar) fell from base to aligned checkpoint in 20 of 20 models (the sixteen and the four of Appendix H), but that counter was never validated and none of its outcomes was scored. At 16:33:53Z on 31 August, 47 minutes before the commit, the second-person rate and four other counters were read between the SFT and the final checkpoint of four AI2 sequences (Tülu-3-8B, OLMo-2-13B, OLMo-2-32B and an OLMo-2-7B RL stage), where the second-person rate fell in all four. The headline statistic, the base-to-aligned second-person rate on this panel, was first computed at 17:36:48Z, and we found no earlier computation of it.

Failed prediction 1, that the fall would appear only in the Tülu and OLMo lineages, is part of the same commit. What it was based on is not recorded. When it was written, the only second-person rate on disk was the AI2 reading above, the unvalidated obligation counter had fallen in every lineage, and the line classifier had found almost no addressed lines in one aligned checkpoint from another lineage.

Pre-registration record for the second protocol. The headline of this paper was chosen by a rule written before the chat-template numbers existed, and that rule turned out to be incomplete. It asked only *how many* models remain interval-disjoint and *downward* under the chat template, and offered three headlines indexed by that count. The measurement broke the question: seven models are disjoint and *upward*. Read as written, the rule selects its most conservative branch, recorded as branch B with a line added to the Limitations: it keeps branch B's title, *Under a Fixed Prompt Format, Alignment Removes the Persons*, and narrows the register result to the raw-continuation protocol in the abstract (rule commit `53ca500c`). We record this rather than quietly rewriting it, because the rule fired correctly and it was the rule that was wrong. The branch actually taken was added afterwards, and its threshold — at least three positive and disjoint models under the shared degeneracy filter — was fixed before the filtered numbers were computed but *after* an unfiltered draft had been seen, in which four of eleven models were already positive. The decision statistic was therefore not contaminated, but the choice of threshold was not blind, and we do not describe this branch as pre-registered. The filtering tool is in the released repository, and the measurement (commit `3c351b73`) is listed in its registration index with its timestamp and digest; the branch record follows with the original documents at camera-ready. The filter itself was specified before it was run and only one filtering rule was executed: a cell enters if it is non-degenerate in all four quadrants.

Random effects across the sixteen models. The same non-independence applies here: fifteen base checkpoints carry sixteen post-trained endpoints, Tülu-3-8B and Llama-3.1-8B share a base, and six Qwen sizes share a pipeline, so the random-effects variance is estimated across sixteen contrasts that are not sixteen independent draws. The pre-registered analysis treats each model as its own contrast with its own prompt-clustered interval, and that analysis remains primary. To see

what happens when the sixteen models are instead modelled as draws from a population, we fit one random-effects model over the sixteen pre-registered ΔM_1 estimates and their intervals. The pooled shift is -4.44 per thousand with a 95% interval of $[-5.39, -3.50]$, still excluding zero, against -4.47 under a fixed-effect pooling. Between-model heterogeneity is large and we report it rather than smoothing it: $\tau = 1.79$, $I^2 = 89\%$, $Q = 142.4$ on 15 degrees of freedom. The practical consequence is the interval, not the point: treating the models as draws from a population widens it by $3.2\times$. The direction is what replicates across models. The magnitude is not, and the paper claims only the first.

The in-context assistant prompt, in full. A third prompt format, the *in-context assistant* prompt, answers a different question: what the base checkpoint does when it is given the assistant task itself rather than a text to continue. It re-runs the 34 debate prompts, 192 generations per prompt, inside the URIAL frame (Lin et al., 2024), three worked examples and a fixed answer header, at the same temperature and top-p. The frame is the released `inst_1k` prompt at $K = 3$, commit `127f94e`, SHA-256 `3a3989a9...`, used unedited, and its own second-person density was counted before any generation was run, so the copying account could be stated in advance rather than after the numbers (§4).

Models, snapshots and decoding. Every generation reported in this paper is traceable to a checkpoint snapshot hash recorded in the run manifest. The alignment checkpoint sequences used for the weight-space and data size experiments are listed below. The sixteen-model panel manifests are released with the code.

checkpoint sequence	stage	model	snapshot
Tülu-3-8B	base	`meta-llama/Llama-3.1-8B`	`d04e592bb4f6`
Tülu-3-8B	SFT	`allenai/Llama-3.1-Tulu-3-8B-SFT`	`f2a0b46b0cfd`
Tülu-3-8B	DPO	`allenai/Llama-3.1-Tulu-3-8B-DPO`	`a7beb67e33ff`
Tülu-3-8B	RL	`allenai/Llama-3.1-Tulu-3-8B`	`666943798adb`
OLMo-2-13B	base	`allenai/OLMo-2-1124-13B`	`3fefddc1bf18`
OLMo-2-13B	SFT	`allenai/OLMo-2-1124-13B-SFT`	`b7ec47ed94f3`
OLMo-2-13B	DPO	`allenai/OLMo-2-1124-13B-DPO`	`97a7f33fc8b4`
OLMo-2-13B	Instruct	`allenai/OLMo-2-1124-13B-Instruct`	`3a5c85baefbb`
OLMo-2-7B RLVR	DPO	`allenai/OLMo-2-1124-7B-DPO`	`e34ea60adff2`
OLMo-2-7B RLVR	step 60	`@step_60`	`701db8a04495`
OLMo-2-7B RLVR	step 120	`@step_120`	`1449245a2b3c`
OLMo-2-7B RLVR	step 180	`@step_180`	`d2cc98e0a938`
OLMo-2-7B RLVR	step 240	`@step_240`	`0a106d76da43`
OLMo-2-7B RLVR	step 300	`@step_300`	`2bb37f4c942b`
OLMo-2-7B RLVR	step 360	`@step_360`	`f942bc3aae37`

The six RLVR stages are not six repositories. They are branches of `allenai/OLMo-2-1124-7B-Instruct`, and the `@step_`n entries above give the branch; the snapshot column is the commit that branch resolved to when we downloaded it. A loader that takes the newest snapshot of that repository collapses all six stages onto one checkpoint without saying so, so each stage was pinned to its own revision and its identity checked against the weights rather than against the branch name.

Human labelling package. The forty sentences of Table A3 were drawn from 664 counter-flagged correction sentences across the three substrates, through a fragment gate applied before any labelling: a sentence must end in terminal punctuation and run between five and sixty words. The gate fired $6 + 40 + 13$ times in production (too short, too long for a labelling form, no terminal punctuation) and its two checks were measured separately against a previous package's known-broken items before the threshold was set. The form is a single question with three options and no counter output. The delivered labelling form carries `sha256` prefix `6ecfb5210f46e461`, and its key is held in a sibling directory that the labelling never saw. One item was returned unlabelled and one was set to NO-CONTEXT in the clarification round (Appendix C); both are excluded by name rather than imputed.

Minimal pairs. The mood-swap stimuli, used by an instrument reported in the repository, are minimal pairs. The method has a long precedent in acceptability work (Suijkerbuijk et al., 2025).

Human reference corpora. Two strips, both measured with the counters of Section 2 and the same forty-word floor. *Spoken/interactive*: the Supreme Court oral-argument corpus (speaker roles `J` and `A`) and the ChangeMyView subreddit corpus, both from ConvoKit. The cluster unit is the case and the conversation respectively. *Written expository*: best-voted StackExchange answers across ten non-code communities (`flax-sentence-embeddings/stackexchange_title_best_voted_answer_jsonl`, site-balanced, 20,000 answers, underlying content CC BY-SA), wikiHow article prose (`soyasis/wikihow_small`, 1,871 articles, wikiHow content is CC BY-NC-SA and is used for measurement only and not redistributed), and OpenStax textbook paragraphs (`crumb/openstax-text`, 76 books, CC BY 4.0, with the book as the cluster unit). Intervals are cluster bootstraps over those units with 500 draws. No corpus text is redistributed. Only the counted statistics are released.

Intervals, null, placebo and the noise floor. CIs are bootstrapped over prompts, never over sentence pairs, and the null permutes stage labels within a matched cell and rebuilds the statistic over 200 draws. The placebo splits each prompt's *own* generations in half, so that between-prompt variance does not leak into the calibration. **This is not the placebo the pre-registration named.** The registered placebo split prompts within a checkpoint; on the debate prompts only 4 of 16 models clear it (its $p95$ band runs 4.1–6.7 per thousand, wide enough to swallow most of the effect), and that count was read before the paired placebo was written. The paired within-prompt placebo ($p95$ 0.9–1.2) is the better instrument and all 16 models clear it, but it replaced a registered one after the first count had been seen, and we list it as a deviation (Appendix B) rather than as the bar we set.

Human reference: the assistant-role anchor. OASST1 (Köpf et al., 2023) contributes the one role-matched human reference: English `assistant` messages clearing the same 40-word floor, 18,828 of them, from which a uniform reservoir draws 6,528 — the panel's own n, so the two are read at the same sample size — over 3,028 conversation trees, the tree being the cluster unit. The corpus is released under Apache-2.0; no message text is redistributed here, only the counted statistics. The `prompter` side does not reach the 2,000-row floor we set for a band (1,790 eligible) and is reported as UNMEASURABLE rather than shown. **The sampler was repaired after these numbers were first read**: the original reservoir retained each accepted record with a fixed probability rather than n/i, which concentrates the draw in the tail of the file. Over a 100,000-record stream the old rule put 93% of a 1,000-record sample in the last decile against a uniform 10%; the courtroom rows move most under the repair (advocates $10.3 \to 6.6$ per thousand, justices $24.1 \to 23.3$, Reddit $19.6 \to 18.8$), while this anchor barely moves ($15.0 \to 14.7$).

Decoding and seeds. ELICIT-99 generations use temperature 0.9, top-p 0.95, four sampling seeds per stage, and a per-stage token budget that was raised once, before any outcome was read, because truncation was found to depend on both stage and stimulus class. The debate-prompt outputs were generated with a single seed per checkpoint. Its manifests do not record which seed, and the multi-seed audit reported in the limitations is the measurement of that gap. Seed values are integers taken from the run date and carry no other meaning.

Counters and judges. The register and form counters are rule-based and deterministic given the released code, over spaCy (Honnibal et al., 2020) part-of-speech and dependency parses (spaCy 3.8.14, `en_core_web_sm` 3.8.0). The open-weight cross-check judge was run at temperature 0. The primary judge is reached through an agent interface that does not expose decoding parameters, so no temperature is reported for it. The second label judge of Appendix C is reached the same way, so each of its labels is the majority of three runs. Each judge uses one model version throughout—primary `Claude Opus 5`, whose act-level labels (Appendix C) were produced on 31 August 2026, open-weight cross-check `Qwen2.5-32B-Instruct`, and second label judge `Claude Sonnet 5`, run on 21 September 2026—and no result compares judge versions, so no version table is reported. Prompts are released with their SHA-256 digests.

Seed coverage of the register battery. The sixteen-model panel was generated with a single sampling seed per checkpoint, and the generation manifests of that panel do not record which seed. The single-seed status is an inference from the pipeline default rather than a logged fact. A separate multi-seed audit regenerates each model with four sampling seeds and re-reads it with the same estimator. It has completed **all 16** (the manifest lists them), and they include the model that carries the smallest effect in the panel, which is the one a seed artifact would break first. Re-read with four generation seeds, all of them hold: the shift stays negative, the prompt-clustered interval still excludes zero, and the effect still exceeds the paired placebo. Four-seed pooled estimates differ from the single-seed ones by 0.05–0.61 per thousand, against seed-to-seed spreads of 0.08–0.52. What survives the re-read is the sign, the interval and the placebo, in all 16; the *magnitude* does not travel as well. 8 of the 16 single-seed readings fall outside their own four-seed range (up to 5.3 seed-sd, Gemma-3-4B, and 4.7, OLMo-2-32B), so we read magnitudes at the four-seed pooled value wherever it exists and mark the rest.

LIMITATIONS IN DETAIL

What the second protocol leaves open. The chat-template split of Figure 2(b) is a measurement without a mechanism. We can say that the same weight contrast changes sign between two prompt protocols in seven of sixteen models, and we cannot say why those seven. Scale and the presence of a chat template on the base checkpoint fail to sort them; the lineage label sorts them and is confounded with the pipeline, which is why it is a sorting variable and not a mechanism. Of the two candidates we named as untested, one now is: 9 of the 16 templates prepend a default system line and the split does not follow that either (§3). Whether the *base* checkpoint ships one is **confounded with lineage and underpowered, not excluded**. The other, pretraining exposure to chat-formatted text, we still cannot measure with open checkpoints alone. Two further gaps are ours rather than the data's. The stage attribution is protocol-dependent and we report it that way rather than choosing a protocol; it reaches one positive model, OLMo-3-7B, and it is missing for that model under the template because its checkpoints do not all ship the same chat template. The copying account for the in-context assistant prompt is **narrowed further and still not closed**: the base checkpoints exceed the density of the three examples they are shown but not that of the prefix taken whole, and the control that decides it — the same prefix rewritten without a single second person, measured at 0.0 — was run on 16 of 16 models. It removes about 27% of the base band and leaves the rest (11.6–21.2 per thousand), so the density the bases produce is mostly not copied from the frame. Both sides are read on the six single-strength prefix conditions, the ones on which the person-stripped frame fits the $1{,}024$-token window its corpus was generated at, each checkpoint in its own format. Mistral-7B-v0.3 is excluded: its person-stripped prompt exceeds that $1{,}024$-token window on its own tokenizer even at single strength, and the corpus was not regenerated. On the chat-template protocol the base checkpoint is also asked to answer in a format it was not trained on, and that asymmetry is intrinsic to the comparison rather than filterable. The in-context assistant prompt removes the genre half of the objection and does not remove the split: five models still go the other way under it, and what separates them is the pipeline, confounded with the lineage label (§3). The first five checkpoint sequences' sides, the two caveats on the stage reading and the degeneracy filter are in Appendix D.

What the targeted preference runs do not separate. Selection on person marking co-selects other properties: in the person runs the chosen side also carries fewer refusals, fewer code blocks and fewer tokens than the rejected side, and the unselected run carries none of these. The runs co-select a register bundle alongside the person axis, and we separated the two by training the minimal pair: a run matched on refusal rate and length but carrying no net person signal shifts the count by -0.24 per thousand $[-0.51, -0.01]$ against -6.89 for the person-selected run, clearing neither the bar nor its own placebo band. The bundle is therefore small rather than absent, and the dimensionality of what remains is open. Whether a prompt alone withdraws the persons is only half-open: the in-context assistant prompt carries an instruction header of exactly that kind and the base answers at 21.3–32.4 per thousand rather than at the aligned band, but the factorial condition that would separate the system line from the worked examples was not run. **We did run all sixteen models with a one-line neutral system prompt, and it does not settle this either**: the reading requires a model to clear a degeneracy shelf (distinct-4 ≥ 0.60), 14 of 16 fall below it under that prompt, and 2 of 16 is beneath the pre-registered floor of 12, so the outcome is UNMEASURABLE by its own rule and **we claim nothing from it**. The two models that can be read *disagree in sign*, each with an interval clear of zero ($+2.64$ $[+1.77, +3.50]$ and -2.59 $[-3.65, -1.55]$); with the denominator under the floor

that is a description, not a result. Our own preference runs are all DPO; from one supervised base, the released DPO and SimPO (Meng et al., 2024) checkpoints both withdraw the second person on the raw-continuation protocol, the reference-free one further (-6.53 $[-8.10, -5.21]$ against -2.77 $[-3.40, -2.09]$ per thousand), so the withdrawal is not specific to DPO; each checkpoint was trained on its publisher's own pipeline, so the two objectives are not compared at an equal data size.

What we do not explain. On the raw-continuation prompt we localize the effect to how the training stages share it — the preference stage carries more on five checkpoint sequences, the supervised stage on two, and two are tied — and the preference stage's part to its data size. Whether it continues into reinforcement fine-tuning we cannot say on this evidence: of the three checkpoint sequences that move again after the preference stage, the one whose placebo we paired, Tülu-3-8B, moves -0.46 there, inside the paired placebo band of its own DPO stage ($p95$ 0.56), so we report the stages and not a continuation. Two independence caveats belong with the stage result: the three checkpoint sequences we opened first are one pipeline at three scales rather than three independent replications, and the three preference corpora are not independent either, since the Tülu-3 mixture contains UltraFeedback (Appendix D). We do not explain *why* either stage withdraws the persons, nor why the interpolated delta returns toward base on one of the four checkpoint sequences and not the others, which is reported in the repository. Head- or circuit-level analysis is the natural next step and is outside this paper, as is testing whether the retreat can be removed post hoc — an activation-space ablation of the kind Anand et al. (2026) demonstrate. Every count in this paper is also taken on a single turn. Ibrahim et al. (2025) find that most anthropomorphic behaviours first appear only after several turns, so a single-turn density is a lower bound on what a conversation would show, and the direction we report is not guaranteed to be the direction a multi-turn measurement would find. Our second protocol sharpens that limit rather than removing it: on the same first turn, addressing the model in its assistant role reverses the sign of the contrast in seven of sixteen models (Figure A2), so what the model does as an assistant is not read off what it does as a continuation engine.

Coverage of ELICIT-99. ELICIT-99 now covers fifteen of sixteen models; the one it does not is Qwen2.5-1.5B, whose aligned outputs fall to the pre-registered degeneracy shelf. **The bar is met at 12 of 16, exactly the bar.** The registration asks for twelve of sixteen models whose intervals exclude zero, and it names no placebo in that count (Appendix A quotes it). Twelve models clear it. The paired within-prompt placebo, a required control outside the bar, is cleared by nine, so the bar, read as interval plus placebo, would not be met (Appendix B). **The in-context assistant prompt registered no bar of its own, and read against this one it does not clear it either**: ten of sixteen models clear the interval on the identical string and eleven with each checkpoint in its own format, short of twelve, and the paired placebo removes none of them. Four further caveats belong here and are named rather than argued. In Gemma-3-4B, -12B and -27B the aligned ELICIT-99 outputs reach the $1{,}024$-token cap in 37, 68 and 74% of draws (the base outputs in 21, 43 and 56%), so part of their density there is read on truncated text; OLMo-3-7B's (24%) is the only other aligned checkpoint above a tenth, and every other aligned checkpoint stays under 1% (Appendix Table A8). Act-level judgements are threshold-dependent and we report presence rather than change; on the reading↔production map, "uncorrelated" at $n = 6$ means *not distinguishable from chance* rather than *shown to be zero*. We measure a text-level gradient and a model's conditional probabilities over it, and nothing here distinguishes a system that *is* positioned from one that reliably *writes as if* positioned. And all sixteen models are English-instruction models: person marking is overt in English and optional in a pro-drop language, so a density counted on pronouns is not portable as it stands (Milička et al., 2025). ELICIT-99's growth, the placebo bands, the human strips' status as anchors rather than paired contrasts, and the spread of the written strip are in Appendix D.

B Deviations from pre-registration

This appendix exists because the universal claim it replaces — that every reported finding was pre-registered — is not true of this paper, and the pre-registration record is public enough that a reader would find the gap. Table A1 states, per result, what was registered, what was done, and which of the two the reported number follows. Where a deviation improved the instrument we say so and still call it a deviation; the standard we hold to is that the reader learns it from us.

Table A1: **Deviations from pre-registration.** *Registered* = an acceptance bound existed in a committed document before the statistic was computed. Rows marked *descriptive* carry no outcome name anywhere in the paper. *Exploratory* = a threshold was chosen after a draft of the result had been seen.

Result	Status	What deviated	Reported as
Depersonalization, M_1 (§3)	registered	The registered bound is $\geq$12/16 negative with CIs excluding zero. The registration's own placebo split prompts within a checkpoint and only 4/16 models clear it; the paired within-prompt placebo, which all 16 clear, was written after the first count was read.	Registered bound in §1; paired placebo named as the later, better instrument
First registered outcome	registered	The first outcome file is stamped 17:36:48Z; the second run, which adds the paired placebo and two secondary counters, is stamped 17:52:46Z.	The earlier stamp is the first outcome
Readings before the bar	registered	The second-person rate was read on four AI2 training stages at 16:33:53Z; the bar was committed at 17:20:46Z.	Headline first computed 17:36:48Z (Appendix A)
ELICIT-99, M_1	registered	Read as interval plus placebo, the bar is not met; the registration names the interval alone.	Met at 12 of 16; paired placebo cleared by 9
Targeted preference runs (raw continuation)	registered	None.	Outcome as registered
Targeted preference runs (chat template)	registered	Ran, and the registration's own degeneracy gate fired: one run falls below the distinct-4 shelf (0.59 against a floor of 0.60). The slot collision that had produced that run's generations without the template was repaired and the run regenerated in an empty slot, four seeds, all twelve generation sets complete.	`UNMEASURABLE-DEGENERATE`; no outcome name, levels in Appendix D.7
Ordinary instructions (Appendix D.4)	descriptive	Instrument fixed and committed before the run; no acceptance bound and no outcome name. The aligned outputs are each model's own chat template and the base outputs the same model in the URIAL frame, so the two checkpoints differ in prompt as well as in weights, and the base text is truncated at its first code fence (median 36% of rows). The first-person column counts first-person forms directly, excluding the capitalised *US*; the derived measure the first reading printed (first person minus sentence-opening vocatives) is not reported.	Descriptive; no outcome name
Genre-fixed continuation (Appendix D.2)	descriptive	The four genre headers were written by the assistant that ran the experiment, and the design, the single reading path and the genre-conformity rule were committed before any of these generations existed. No bar was registered.	Descriptive; no outcome name
Bundle control, paired run	registered	The registration's person-neutrality criterion was written in pair units, and its own gate measured it degenerate: 93.4% of the pool fell inside the band, so it eliminated nothing. It was replaced — before training, and before any outcome was read — by a two-part criterion on the run's net person signal. The outcome bar, the pair count and the reading order were not touched.	Outcome under the repaired criterion; the correction is at the head of the registration
URIAL prompt set (§4)	descriptive	The registration was written after generation had started and says so.	Levels and counts from all sixteen prefix conditions

Table A1 (continued).

Result	Status	What deviated	Reported as
URIAL prompt length	descriptive	The first generation used a 1,024-token window; the longest prompt in the panel is 1,235 tokens, no cell is truncated at 2,048, and the paper reads 2,048.	2,048 throughout; the person-stripped corpus, not regenerated, stays at 1,024
The IFEval cost of the restoring sentence (Appendix D.4)	Descriptive	The registration named a cluster-bootstrap body that is bound to the pronoun measure and cannot take a difference of accuracies; its procedure was kept (prompts resampled, same seed, 2,000 draws) and added as a separate primitive with its own test. The official evaluator's random seed was also fixed, because it fills unspecified constraint parameters at random.	A measured cost, with the share due to the sentence asking for a continuation left unseparated
The code-block-matched control (Appendix D.7)	Attempted, not run	Matching refusals, length and code blocks at once, with no pronoun difference, reaches a ceiling of 3,348 pairs against the 7,537 each run uses; three ways of coding the code-block axis give 1,674 to 3,348.	A ceiling, not a null
The chat-template readings of OLMo-2-7B and Tülu-3-70B (§4, Table 2)	Descriptive	Registered before generation as descriptions, with no bar and no outcome name, and read with the registered stage tool and its paired placebo; each checkpoint sequence's stages share one template, and Tülu-3-70B's post-preference stage was not run.	Fills the matching dashes of Table 2; the stage rule of §4 is read as written
Qwen2.5 bases read through their own chat template (Appendix D.3)	Descriptive	Registered before it was computed, with no bar and no outcome name; the generations already existed, so no model was run for it; four of the six sizes were read.	Reported as description
Person-stripped control, same-string comparison	descriptive	Only the base outputs of the person-stripped corpus were generated, by design, so that the person-free prompt set would first be read against each model's own template. The same-string comparison has no aligned outputs on that corpus and therefore cannot be reached by any model — a named outcome that no model could produce.	`UNMEASURABLE-BY-DESIGN`; the control is reported on the task-equal comparison only
Chat-template 2×2 headline	exploratory	The reported branch was chosen after an unfiltered draft had been seen.	Stated in Appendix A
Band without the system line; template length (Appendix D.3)	descriptive	Read after the draft, no bound; outputs `08d17c66f2f27c27`, `aa4f0b134b0ae42a`.	Values only
Zephyr, OLMo-3, Tülu-3-70B checkpoint sequences	descriptive	No registration; instruments fixed, no acceptance bound.	Checkpoint sequence cards, no outcome name
Origin classes	descriptive	Redefined three times during the program.	Final definition only, with the count
Human anchors	descriptive	No bar. The sampler was repaired after the numbers were first read (§D.2); advocates move from 10.3 to 6.6.	Both values, with the repair
Scorer specificity margin (§5)	registered	1.5 was chosen after the first fit returned 1.86, then registered before the fits it judges.	Margin and its provenance

Table A1 (continued).

Result	Status	What deviated	Reported as
Stage-vector angle (§6)	registered	Three of the six predictions committed before the vectors were built — that no checkpoint sequence reaches a large cosine, that at least one is near zero, and that most stay below the band — **are structural consequences of seven billion coordinates, and the registration says so before the measurement**. They carry no evidential weight. The informative predictions are the two on local alignment, whose null is exact.	Local alignment only; the cosine reported as a magnitude, not as a result
Angle × pipeline group	*unmeasurable, declared*	Two checkpoint sequences against four gives a sharpest attainable permutation probability of 0.13, so the comparison cannot reach a conventional bound. Registered as unmeasurable **before** the run.	Named as unmeasurable; no outcome

Table A1 (continued).

Result	Status	What deviated	Reported as
Instruction reversal (§3)	registered	Three readings, in order. The first pass scored 12 of 14 models against a bar of 12, because the tool looked up two checkpoint sequences' base checkpoint under a name it assumed rather than the one on disk. Resolving the name from disk gave 14 of 14 and moved the register reading's median from 1.73 to 1.46; the outcome name did not change. That reading was interim: the registered analysis covers sixteen models, and two were still being generated. With all sixteen read the medians are 1.72 (register) and 3.14 (named forms). A pronoun-free register sentence was run afterwards, when an outside reading showed that the first carries a *you*; it was registered before it ran but lies outside the original plan, and it moves the register median from 1.72 to 1.24 without changing the outcome name.	The 16-model reading, with both earlier readings stated
Dolci corpus read (App. D.3)	descriptive	Reported first on a 20,000-pair sample, then read whole. Both readings agree in sign and direction; the full read narrows the intervals about threefold.	Full read, with the sample's agreement stated
Pairs without person content (§5)	registered	The step count was first carried over from the earlier runs (2,302) and the run's own gate fired before training; it was re-derived from this run's length losses by the registration's own rule (2,297).	Outcome as registered, at the re-derived step count
Tripled pronoun gap, second seed (App. D.7)	registered	None.	Outcome as registered
Reference-free objective (Appendix A)	registered	Registered before either checkpoint's outputs existed, for a raw-continuation and a chat-template protocol. Only the raw-continuation protocol was generated: the SimPO checkpoint ships no tokenizer or chat template of its own and borrows its base's, whose embedding rows it matches on every row we sampled.	Raw-continuation protocol; the chat-template protocol not generated
Leaving the continuation on the raw-continuation prompt (§3)	descriptive	Flagging rule written before the count and committed with it; ceiling rule committed before its count; no bound. Outputs `cd1fc53b4577595a`, `f08ab0314f9504dc`, `bb35e8ae4557349b`.	16 of 16 with flagged generations removed, and under the ceiling
Lineage sign, OLMo-3 merged into AI2 (§3)	descriptive	Computed after the six-level test was read, with its tool, seed and permutations; output `b879b82a4cf14b21`.	12 of 14, $\mathrm{frac} = 0.092$
Person-selected runs, training length (§5)	registered	Registered one epoch at batch 16; 458 steps were set in advance from a measured pair-drop rate, and the trainer records 0.987–0.989 epochs.	Outcome read as registered
Panel sampling seed (§3)	descriptive	One seed, not recorded in the manifests; the four-seed re-read covers 16 of 16 models.	Magnitudes descriptive

Table A1 (continued).

Result	Status	What deviated	Reported as
Addressed or generic "you" (App. C)	registered	Agreement bar (0.80) set after labelling, before comparison.	Not met; its labels are not used
Reply or continue (App. D.3)	registered	Same bar; six cells settled by one of the authors, blind to the judge.	Not met; stays in the appendix
Addressed or generic, second label judge (App. C)	registered	Same bar and sentences; judge tried after the first missed it.	Met; its labels split the rate in §2
Reply or continue, second label judge (App. D.3)	registered	Same bar and outputs; judge tried after the first missed it.	Not met; stays in the appendix
Five later readings (App. D.2–D.7)	descriptive	Paths committed before counting; no bar.	Values only
Chat-template first person, unfiltered cells (App. D.3)	descriptive	Path committed before counting; no bar; all sixteen models read, and the unfiltered second-person column reproduced as a gate.	Values only
Identical-string ELICIT-99 (§3, App. D.2)	descriptive	Path committed before generation; no bar; all sixteen models read.	Values only
Pairs without person content, second seed (§5, App. D.7)	registered	Same configuration, seed changed; the driver's calls were repaired before any generation existed.	Holds: sign and interval as registered

Table A1 (continued).

Result	Status	What deviated	Reported as
Stage attribution by preference-signal lineage (§4)	descriptive	The account that the preference stage installs the withdrawal where a judge lineage is shared and the supervised stage where it is not was written before the lineage of the two checkpoint sequences outside the AI2 and Zephyr pipelines was read. SmolLM2-1.7B's card names UltraFeedback for its preference stage, so a checkpoint sequence on which the supervised stage carries the withdrawal shares that lineage, and the account fails on measurement.	Withdrawn; the stage is reported per checkpoint sequence as an observation
Four-seed re-read, last three models (Table 1)	registered	Registered before reading, bar the single seed's sign; all three hold; output `e25d5680403ea684`.	Sign; magnitudes as ranges
Direct first-person count, capital *US* excluded (§3)	descriptive	Rule committed before the count; outputs `4e08e3c1f7da294b`, `11f62ffa1d0c565f`.	Counts only
Prose lines alone, both persons (§3)	descriptive	Rules committed before each count; outputs `863f0e5a996374bb`, `3872f2202fecf6ad`.	Counts only
Chat-template first person, turn-cut and echo-free (§3)	descriptive	Rule committed before the count; output `3ba0f71194e743f6`.	Counts only
Model-by-protocol signs; band on three prompt formats (§3)	descriptive	Read from published readings, rule committed first; outputs `fe457b203027de88`, `a8135d1c89711528`.	Signs and ratios only
Sign test at dependence units (Limitations)	descriptive	Unit definitions committed before the test; output `682ffccdfb1b0ce8`.	n and p per unit
Person-selected runs, training seeds 2 and 3 (§5)	registered	Registered before training, bar the sign in each new seed; the first reading of two generation sets failed, and they were re-read with the same tool and arguments; output `e4f607f2aad004e6`.	Sign; magnitudes as ranges
Pronoun gap-flattened run, training seeds 2 and 3 (§5)	registered	Each seed registered before training, bar the displacement from the first seed's flattened run (seed 2) and the sign against the supervised checkpoint (seed 3); the second seed does not clear its own paired placebo and its share is printed, marked; outputs `7628d436acd20051`, `5e8d12d06389f713`.	Sign; share on all three seeds, the second marked
Natural-pronoun gap run, training seeds 2 and 3 (§5)	registered	Registered before training, bar the sign against the supervised checkpoint in each new seed; the sign holds in both; the flattened-minus-natural interval was declared as description and computed after the registered reading; outputs `e6ca02a13bb5e012`, `d2e3fcc5901386f3`.	Sign; magnitudes as a range
Degeneracy, echo and turn-cut applied at once, chat-template (§3)	descriptive	Read after the template split was seen; the bar — at least three of the seven rising models surviving — was committed in the rule before the count.	Six of seven survive; output `0ebf497dc148cbf9`
System-line echoes, chat-template outputs	descriptive	The echo flag reads the prompt's last 200 characters, where no template's system line falls; read after the split was seen.	One model changes class, Qwen2.5-14B down to null; output `cc9594069ac76ddb`

Table A1 (continued).

Result	Status	What deviated	Reported as
ΔM_1 by line type, both protocols	descriptive	Rule committed before the count. The equivalence gate — the three line classes recombining to the outputs' own M_1 — was set at 10^{-6} and measured 8.2×10^{-2}, so the reading is **not reported here**.	Tool record; not reported. Output `1d324f6813bf4dca`
Panel table read at four sampling seeds	descriptive	Levels and shifts pooled over four generation seeds; the equivalence check against the two pre-registered seed readings measured 0.	Counts unchanged, 16 of 16; output `89c04ecac9f460a5`
Geometric dispersion beside the band ratio	descriptive	Read after the ratios were seen; a second dispersion statistic on the same cards, no bound.	Output `a747e79858db6beb`
Third placebo: prefix conditions split within each prompt (§1)	descriptive	Specified after the registered placebo cleared 4 of 16 and before this count; the rule fixed 200 draws, the 95th percentile and a reporting threshold of 12 of 16.	16 of 16 clear; output `2aa3782d1ee6bb69`
Prose sentences alone, the six list-dependent models	descriptive	Rule committed before the count; whole sentences are classified rather than lines, so a sentence reaching a list line is dropped whole. No bound.	Values only; output `724acde1d2ee4f42`
Prose sentences alone, the other ten models	descriptive	Same rule, extended before this count to the models that hold on prose lines, so that a by-sentence count covers all sixteen. No bound.	10 of 10 hold; 13 of 16 with the six; output `0f2aab1434403222`
Model cards and objective texts, read for person and register	descriptive	Search strings and reading labels committed before reading; 5 of 32 card slots are not on disk and are reported as missing.	Values only; output `3155766934e42bf2`
ΔM_1 by line type, tolerance at the printed digit	descriptive	Tolerance 0.005 per checkpoint, half the last printed digit, set after the 10^{-6} tolerance above had failed and its largest deviation had been seen; per-model deviations were not read before the rule.	11 of 16 models reported; output `14033520ae768dc1`

What the registrations cost us. For the registered results, each acceptance bar was fixed before its statistic was computed, except where the rows above record otherwise, and each registration is a timestamped commit in the released repository. Not every registered prediction held: of the 26 registrations that carry a bar or an outcome name fixed in advance and whose results are reported here, 13 carry at least one prediction that failed. They are listed below, one line per registration; every row points to its evidence line (file, line and digest) in the released repository's record of failed predictions except the last two, which point to that record alone. The eight rows marked *repository*, these two among them, are those whose results are not reported here.

#	registration	failed	reported	what the failed prediction said	what was measured
1	the registered bare-prompt direction of the second person (Depersonalization, M_1)	1	here	The second-person decline would appear only in the Tülu/OLMo lineage (a local pattern).	The prediction failed on run A and could not be measured on run B.
2	the task-matched ground: the neutral panel's prompts inside the URIAL frame	3	here	On the identical-string comparison, 6–8 families would fall and 2–4 rise with intervals excluding zero, the fallers mostly AI2 and Qwen and every riser Gemma or Llama. The base models' median in the URIAL frame would be 12–18 per thousand.	10 models fell and 6 rose, both outside their bands; 9/10 fallers were AI2 or Qwen, but one riser, OLMo-3-7B (AI2), was neither Gemma nor Llama. The base median was 20.13. *(scored on the 1,024-token card)*
3	the default system line, present or absent inside the same string	1	here	Among the interval-disjoint families, the signs of the shift would be indistinguishable from random (sign agreement u below 0.543).	Sign agreement measured u = 0.5565, above the 0.543 bar.
4	two ladders outside the AI2 and Zephyr recipes (SmolLM2-1.7B, NeuralHermes-2.5)	1	here	On both the SmolLM2 and Hermes ladders the outcome would be the one named "same step".	Both checkpoint sequences read "different step": the supervised stage moved –3.3356 (SmolLM2-1.7B) and –5.4761 (NeuralHermes-2.5) with intervals excluding zero, the preference stage +0.0740 and +0.1497 with intervals covering zero.
5	re-scoring person-swapped responses with two public reward models	1	here	Both reward-model scorers would read as penalising.	Both scorers read as rewarding (Δ = –0.0004 and –0.5971, person-free minus person-marked score).
6	the person dial (four arms of 7,537 preference pairs selected on person marking)	4	here	The unselected arm's shift against the supervised checkpoint would be –1.0 to –2.5. On the person arms force would separate from the persons, the force-selected arm would not train force at this dose, and the embedding estimator would move less than 1 null-sd in every arm.	The layer-separation, force-below-data-size and embedding-band predictions all failed on the same measurement, so what the runs steer is register rather than person marking; the unselected run's actual shift was +0.225.
7	a third targeted run on minimal pairs built by exchanging a single determiner	4	here	On the minimal-pair arms the mean dial magnitude would be 3.0–5.5 per thousand, force would move at z 2.0–6.0 on the person arms, and the downward shift would exceed the upward one. The embedding estimator would move less than 1 null-sd in every arm.	The shift size, the force z, the down-over-up asymmetry and the embedding-estimator band were each scored as not holding. *(runs below the degeneracy shelf; magnitudes not read)*
8	pairs without person content	3	here	The arm trained on pairs without person difference would not withdraw (the first-ranked of 7 outcomes, at 0.34), and it would withdraw less than the neutral arm, with a shift above –1.1224.	The run withdrew (the outcome given 0.26), with a shift of –1.2041; the ranking's Brier score was 0.774.
9	the reference-free objective: released DPO and SimPO checkpoints from one supervised base	1	here	From the same supervised base, the SimPO step would withdraw less than the DPO step, with non-overlapping intervals.	The SimPO stage was larger than the DPO stage; the DPO reference stage read "same sign" and no output fell below the degeneracy shelf, so the prediction was scorable.
10	the scorer fit: a joint fit with first-person density and two surface controls	3	here	In the joint fit the second-person coefficient would stay between 0.10 and 0.19 in absolute value, its margin over the other controls would fall below 1.5 (a general style penalty), and it would be largest on the helpfulness rating.	The coefficient measured 0.1925, 0.0025 above the upper bound; the margin was 3.13; and the largest coefficient fell on instruction following.
11	the mechanical form counters (bare imperative, second-person obligation, hedge, interrogative directive, hortative)	8	here	Aligned models would use fewer bare imperatives ($\geq$4/6 families), more hedges ($\geq$5/6) and more interrogative directives ($\geq$4/6) than their bases, with the bare-imperative drop at the SFT$\rightarrow$DPO rung on $\geq$2/3 ladders. The hedge counter would give the largest, most stable difference, sentences per generation would rise ($\geq$4/6), the intact-only filter would change no cell's outcome, and hortatives would rise and clear the bar in $\geq$4/6.	Against thresholds of 4, 5, 2 and 4, the first set measured 2/6, 1/6, 0/3 and 1/6; the second set measured 1, 3, 29 and 2 against thresholds of 4, 4, 30 and 4.
12	the reading cascade: identical text scored under a real and a matched pseudo-neutral context	1	repository	On the CMV substrate, the share of families whose interpersonal-command channel separates positively would lie between 0.375 and 0.625.	The CMV share measured 0.333.
13	injecting the pre-specified weight-space dominance direction	2	repository	For Tülu-8B and Zephyr-7B, the dominance-direction arm would register movement along that direction ("on-axis").	Both read "off-axis": the movement registered outside the dominance direction rather than along it.
14	the grammatical mood swap: rule-based imperative–request minimal pairs under a short and a long frame	1	repository	A dominant-phrased request would register on the command side (the dominant-request trap cell firing) in at least 3 of 6 families.	The dominant-request cell fired in 2 of 6 models (0.333).

#	registration	failed	reported	what the failed prediction said	what was measured
15	undoing the preference delta by interpolation	2	repository	The erasure would sit in the single DPO weight delta, so undoing that delta would work on OLMo-2-32B (a monotone curve approaching the base at α = –0.5) and on Zephyr (dSFT→dDPO).	On OLMo-2-32B the curve was monotone but did not approach the base, and Zephyr read as scattered.
16	the behavioural anchor: the supervised weights stepped either way along the dominance direction at a fixed KL budget	2	repository	The ruler would show the dominance direction on Zephyr, with both perturbation branches inside the bar, and on Tülu as well.	Neither showed it: on Zephyr the plus branch cleared the bar (0.0000) and the minus branch did not (0.2000); on Tülu the branches read 0.3500 and 0.2833.
17	rung-by-rung reading with the embedding estimator	4	repository	No pair's embedding-estimator interval would lie wholly below 0, the largest shift would fall at an interior rung in every family with at least 3 dense rungs, and Gemma-3-27B's dense pair would show no net erasure. On every OLMo-2 ladder the largest step would be SFT→DPO.	OLMo-2-32B's base→SFT interval lay wholly below zero ([–0.0347, –0.0048]), the Qwen2.5 maximum fell at the largest model (14B), and Gemma-3-27B moved +0.0314 with an interval excluding zero. The largest OLMo-2 stage was SFT→DPO at 13B but DPO→Inst at 32B.
18	the first scorer fit, whose specificity margin set the later bar	3	here	The second-person coefficient would stay under the floor, with $\lvert\beta\rvert$ between 0.005 and 0.050, and the first-person coefficient would be larger in absolute value. The modal-density placebo would not clear the floor.	The second-person coefficient measured 0.2116, 4.2 times the predicted ceiling; the first-person coefficient was smaller (0.179 against 0.212); and the placebo prediction did not hold.
19	the low-dose targeted preference arms (four DPO arms on form-selected pairs)	1	here	The dominance-axis shift would follow each arm's direction, negative for imperative-down and positive for imperative-up, with $\lvert\Delta\rvert$ above the MDE.	Imperative-up came out negative and already below the bar, and no run exceeded the bar.
20	the injected-needle control	1	repository	—	*in the repository's record of failed predictions*
21	a six-counter reading of three training ladders	1	repository	—	*in the repository's record of failed predictions*

Not counted: two operational predictions (whether runs would finish) and two base-rate statements. Eight rows fall outside the 26, because their results are not reported here: they were scored and they failed, and their wording and score are in the repository's record. Six of them are passages that moved to the repository when the appendix was cut. The *what the failed prediction said* column quotes each registration verbatim, in its own vocabulary: *family* is its word for what this paper calls a model, and *rung*, *ladder*, *arm*, *leg* and *dose* for training stage, checkpoint sequence, run, output and data size; *dial*, *recipe* and *step* for the targeted preference runs, a post-training pipeline and a training stage.

C Annotation and validation

What was labelled, and by whom. This box covers the act-level labels, human and judge. The mood-swap instrument has a separate mechanical validation on minimal pairs, and that one is described in the repository. Three annotation layers stand behind the counters here, and two more check judges; we state each with its size and its limit. *Two-judge act labels*: 580 responses to ELICIT-99 and to neutral control prompts, labelled independently by two judges under a fixed rubric, giving the agreement figures below and Table A2. *Human anchor*: one of the authors labelled fifty of those items once. *Human correction labels*: forty counter-flagged sentences put to one labeller under a single three-way question, of which thirty-eight were labellable (Table A3). *Addressed or generic "you"*: one hundred second-person sentences put to one labeller, one of the authors, under a three-way question, against two LLM judges, and thirty fresh ones put to the same labeller after the second label judge was chosen (below). *Reply or continue*: one hundred chat-template outputs put to the same labeller, against the same two judges' opening-mode labels (Appendix D.3). In total 318 items carry a human label, and each of the 318 items was labelled once. No annotation layer here is crowd-sourced, and none is large enough to license a per-model claim.

A design iteration we discarded. A first version of the person-restoration comparison paired each impersonal correction with a variant in which only the subject determiner was exchanged for another impersonal determiner, on the assumption that this would hold "something changed" constant while adding no person. Forty items went to one labeller. Nine were answered outside the offered response set and three were withdrawn because their content had been shown to the labeller in advance, leaving twenty-eight. The labeller chose the altered variant in twenty-one of twenty-one experimental items and in seven of seven control items, a difference of exactly zero. The answer letter tracked the presentation order item for item, so the pattern is not a position bias. All seven usable control items happened to exchange *the* for *this*, and the labeller reported reading *this* as the more marked of the two. We read this as a failure of the control rather than as a result: a one-token change was enough to move the choice whether or not the token added a person. The second version replaces that control with one that adds a third-person possessive, so that both variants add a person and the contrast is which person is added.

The blind protocol. Every labelling package in this paper was built and delivered blind, by one procedure. Items are drawn stratified across the strata of interest and then shuffled. The package identifier is a hash of its own content, so a package cannot be silently re-cut, and the answer key is written to a directory outside the delivered package, so the labeller sees items and a question and nothing else. The counter that produced the flags is imported by the package builder rather than reimplemented, so the sentences shown are the sentences the counter actually flagged.

How each judge was kept blind, and by what. The two judges are blind for different reasons and we state the difference rather than call both "blind" and stop. The open-weight cross-check judge is blind *by mechanism*: it is shown one item at a time, allowed a single output token, and read from the token distribution, so there is no context in which an identifier or a judge label could reach it. The second label judge is blind *by discipline*: each item is passed through an orchestrating context in a fresh window that carries the two sentences and the question and nothing else, and the totals are not computed or shown while the items are being answered. Discipline is the weaker of the two guarantees and we do not present it as the stronger one. The identity of the second model is recorded by the authors rather than read from the interface, which is a further declaration rather than a measurement. The orchestrating context is not itself blind: it built the packages and holds the key, so what the arrangement guarantees is that no judge label reaches the scoring window, not that the pipeline is blind end to end. One package manifest from an earlier round of this comparison recorded the design alongside the items; the judge contexts did not carry it, and the delivery tool now enforces that separation with a fixed field whitelist.

Clarification rules for the correction labels. Five rules were fixed before the labels were read, and they are the only post-hoc decisions of record. (k1) An item with no resolvable context receives NO-CONTEXT and leaves the denominator by name rather than being guessed. (k2) "Directed at a person" means the reader or addressee. Criticism of a third party or of a group counts as claim-directed. (k3) One label was converted under k2, from person-directed to claim-directed, and it is the only converted label. (k4) A correction the writer partly agrees with is still claim-directed. A weak mention that declares nothing is not a correction. (k5) The impersonality reading is tied to the counter's own second-person equivalence note rather than read independently. The rules were applied in one clarification round with the assistant, which changed four of the forty labels: two marked hesitations were resolved, one label was converted (k3), and one item the labeller had labelled was set to NO-CONTEXT under k1. Applying or withholding the flattening these rules imply moves the impersonal share between 0.808 and 0.852, and we report that band rather than a single figure.

Act-level judgements are threshold-dependent. Over 580 jointly labelled items, two independent judges agree on the *presence* of status acts (acts by which a reply takes standing over its reader, such as correcting, evaluating, setting the agenda or offering conditional help) at $\kappa = 0.06$–0.58 across the four act types for which presence is defined (0.06 for evaluation, 0.58 for conditional help). On the remaining two neither judge ever recorded the act, so κ is undefined rather than low. The more generous judge labels between 1.6 and 27.5 times more often depending on the act, and the top of that range is one sparse act (110 positives against 4), so the spread is a statement about rare acts, not a uniform bias. We therefore report status-act *presence*, zero on neutral prompts and non-zero on elicited ones, and decline to report a status-act *change* under alignment: it is below our minimum detectable effect and judge-dependent, and we say so rather than reporting the number we have.

Table A2: **Register counters against act-level labels.** The correction counter and the impersonality share, checked against the 580 act-level labels on responses to ELICIT-99 and to neutral control prompts. DET.: does a correction occur (counter ≥ 1 correction sentence vs. the rubric's yes/no). IMP.: among responses where both detect a correction, is it impersonal. Degenerate margins are flagged: where one cell of the table is empty, κ has no base and the raw agreement is the readable number.

contrast	labeller	both yes	rubric only	counter only	both no	agree	κ
det.	Primary judge	56	200	4	320	0.65	+0.22
det.	Cross-check judge	36	30	24	490	0.91	+0.52
det.	Authors (12 items)	2	5	0	5	0.58	+0.25
imp.	Primary judge	45	5	6	0	0.80	-0.11†
imp.	Cross-check judge	29	3	4	0	0.81	-0.11†
imp.	Authors (12 items)	NOT MEASURABLE: the impersonality question was not put to this labeller					

†Degenerate margin: one cell is empty, so κ is not interpretable and the agreement column is the readable number. The authors' row covers the 12 of the authors' 50 items on which the correction question was put (10 where the two judges disagreed, 1 where both said yes and 1 where both said no). The 50 are disagreement-enriched (40 were sampled where the two judges disagreed), so that row is agreement *within the disagreement stratum*, not field accuracy. The impersonality question was never put to the authors.

The counters of Section 2 are rule-based, and the question of what they miss is answerable from labels we already hold: the 580 responses to ELICIT-99 and to neutral control prompts carry act-level judgements from two independent judges under a fixed rubric, whose first question (did the assistant declare a premise or claim of the user's to be wrong) is the correction our counter tries to detect, and whose second asks how that correction was opened: bluntly at the user, hedged, or aimed at the claim rather than the person. Two results, and one of them is a limitation.

First, the counter is a high-precision, low-recall detector of corrections. All three labellers show the same shape: it misses corrections a judge sees and almost never invents one: four false positives in 580 against the primary judge, none in the authors' items. The impersonal-correction share of Table A22 is therefore computed over lexically marked corrections, not over every correction a reader would recognise. The direction of the finding is unaffected, but the denominator should be read with that scope.

Second, within the corrections both instruments detect, the labels agree on four in five, and κ is not the statistic to read there: for both judges the "neither impersonal" cell is empty, so κ has no base and returns a value near zero for a table that is in fact almost entirely on the diagonal: the subset is nearly homogeneous, which is itself what Table A22 reports. The residual disagreement is nameable rather than random. We predicted before measuring that a blunt refusal (*No. That's incorrect.*) carries no second person and would be scored impersonal by the counter against the rubric's DIRECT. That cell fired but was small (3 and 0). The larger cell is its mirror: the rubric calls a correction claim-directed while the counter finds a second-person form elsewhere in the same sentence and scores it person-directed (5 and 3). Our counter is the stricter of the two instruments, and it is strict in a direction that works against the effect we report.

Table A3: **The correction counter against a human labeller.** Forty counter-flagged sentences, drawn balanced across three substrates and across the counter's own two decisions, labelled by one person under a single three-way question: does this sentence correct a claim, and is the correction aimed at the reader or at the claim. PRECISION: the share the labeller agreed were corrections at all. IMPERSONAL: among those, the share aimed at the claim. Two items were set to NO-CONTEXT because the package showed one sentence and the deictic reached back into the previous one (one returned unlabelled, one labelled with that note). They leave the denominator by name. Intervals are Wilson 95%. Recall is *not* measurable here: every item was already flagged by the counter.

substrate	n	precision			impersonal share		
		count	rate	95% CI	count	rate	95% CI
aligned assistant	12	11/12	0.92	[0.65, 0.99]	10/11	0.91	[0.62, 0.98]
online debate	14	10/14	0.71	[0.45, 0.88]	8/10	0.80	[0.49, 0.94]
courtroom argument	14	5/12	0.42	[0.19, 0.68]	4/5	0.80	[0.38, 0.96]
pooled	40	26/38	0.68	[0.53, 0.81]	22/26	0.85	[0.66, 0.94]

The judge comparison leaves one question open, what the counter's flags look like to a person, and forty of them were put to a human labeller under a single three-way question (the judge comparison itself is Table A2), drawn balanced across three substrates and across the counter's own two decisions (Table A3). Three things came back.

Precision, and its scope. Pooled, 68% of the counter's correction flags ($26/38$, CI $[0.53, 0.81]$) are corrections to a human reader. The remaining third are weak mentions rather than declarations. Recall is not measurable in this package and we do not report it: every item was already flagged by the counter, so the sentences it misses are structurally absent. The pooled figure is also substrate-mixed, and the spread is the finding: $11/12$ on aligned assistant text against $5/12$ on courtroom argument. At twelve items per substrate those intervals overlap narrowly ($[0.65, 0.99]$ against $[0.19, 0.68]$), so we report the difference as *suggestive and untested*, not as an established fact. It matters because the human anchor of Figure A3 is courtroom and debate speech: if the counter over-flags corrections there, the human end of the M_3 comparison is measured over a noisier denominator than the model end.

The impersonal share survives. Among the corrections the labeller accepted, 85% were aimed at the claim rather than the reader ($22/26$, CI $[0.66, 0.94]$), close to the 0.91–0.98 we report for aligned assistants, and reached on a sample deliberately balanced towards the counter's person-directed decisions, which should have pushed it down. Four of the forty labels were returned with an explicit hesitation or a reading the labelling round resolved differently. Taking each the other way moves this share between 0.81 and 0.85 and the precision between 0.66 and 0.71, so no statement here rests on how those four were settled. In the courtroom subset the same share is $4/5$. That is four items and we draw nothing from it beyond naming the direction it points, which is *against* the size of the gap we report, not for it.

Where the counter and the person part. Four items were labelled reader-directed. Two are a boundary the counter cannot see: an argument that moves from the claim to the person making it, and a criticism that becomes personal through the situation rather than through a pronoun. One label was revised during the labelling round under a rule written at the same time: a criticism aimed at a *group* is not aimed at the reader, which is the same mechanism our own pilot found in the counter, from the other side. Two further items could not be resolved: the package showed one sentence and the deictic reached back into the sentence before it. One was returned unlabelled; the labeller labelled the other while noting that dependence, and it was set to NO-CONTEXT under k1. Both leave the denominator by name rather than being guessed.

Taken together: the direction of the M_3 finding is unaffected and its errors still undercount impersonality, so 0.91–0.98 remains a lower bound. What this package adds is a precision figure to report alongside it, and a named risk on the human side of the comparison that a larger labelled set should settle.

Addressed or generic "you": two judges against one labeller. The counters cannot tell a "you" addressed to the reader from a generic one (§2). A judge, Qwen2.5-32B-Instruct at temperature 0, labelled $2{,}000$ second-person sentences drawn across the sixteen models and both checkpoints as addressed to a particular reader, generic, or neither (a quotation, invented dialogue, or undeterminable). One labeller, one of the authors, answered the same three-way question for 100 of those sentences, drawn by model; the package named neither the model, the checkpoint nor the judge's label. The two agree on 76 of 100 (model-clustered 95% interval $[0.70, 0.83]$), Cohen's $\kappa = 0.55$ $[0.43, 0.67]$, against a chance agreement of 0.47. The disagreements sit on the boundary the counters cannot see: 11 sentences the labeller read as addressed the judge called generic and 6 went the other way, and of the 7 the labeller left undetermined the judge agreed on 3. Agreement is 43 of 55 on aligned outputs and 33 of 45 on base outputs. We had fixed, before comparing the two, that the judge's split would enter the main text only at an agreement of at least 0.80; it falls short. The judge is a Qwen2.5 model, like six of the sixteen. A second label judge (Claude Sonnet 5, same model family as the assistant) was tried after the first missed the bar; its agreement on the same 100 items is $84/100$ (model-clustered 95% interval $[0.79, 0.89]$, $\kappa = 0.69$ $[0.62, 0.78]$, chance agreement 0.48; on the same sentences it matches the labeller 46 times among aligned and 38 among base checkpoints; the two judges agree with each other on 83); its labels are used only where that bar is met. Both judges were compared with the labeller on these same 100 sentences, so the second was chosen on them. It meets the bar, and it then labelled all $2{,}000$ sentences under the same prompt, three times with a majority label. With its labels, the addressed rate (the second-person rate times the addressed share of its cell) falls from base to aligned on raw continuation in 8 of 16 models and rises in 8 (§2); with the open-weight cross-check judge's labels it falls in 11. The generic rate, the second-person rate times the generic share of its cell, falls in 16 of 16 and rises in none; 12 of the falls have a prompt-clustered 95% interval below zero and 8 exceed their paired placebo, a split of the base checkpoint's outputs into two halves within each prompt (200 splits). Labels were assigned under a three-question rule: (1) a quoted, fictional, or template- or prompt-artefact sentence is UNDETERMINED; (2) a "you" that refers to something this interlocutor said, asked, did or has is ADDRESSED; (3) otherwise, a "you" that can be replaced by "one", "anyone" or "people" without changing the meaning is GENERIC. The rule was fixed after cell 40 and cells 1–40 were re-read under it. After the second label judge had been chosen, the same labeller answered the same question for 30 further sentences drawn from the same $2{,}000$, none of them among the 100: the two agree on 28 of 30, the open-weight cross-check judge and the labeller on 23, and the two judges on 25. At this size we read the totals only: the

Table A4: **Sixteen models on the debate prompts, ELICIT-99 and the in-context assistant prompt, in full.** Table 1 with the elicited and in-context assistant prompts restored. M_1 is second-person forms per thousand tokens and Δ1st the first-person shift on the debate prompts, counted directly (the debate-prompt columns pool four further sampling seeds). DISJ. and bold mark a negative shift whose prompt-clustered 95% CI excludes zero. Models with no elicited entry carry their measured reason: D-SHELF = degeneracy shelf (distinct-4 < 0.60). The prompt sets, the counts and the task-matched reading are in Appendix D.

		debate prompts						ELICIT-99†			in-context assistant prompt*		
model	lin.	base	inst.	ΔM_1	%	1st base	Δ1st	base	inst.	ΔM_1	base	inst.	ΔM_1
Mistral-7B-v0.3	Mistral	13.1	3.6	**-9.51**	-73	22.2	-15.3	32.4	16.3	**-16.03**	30.1	28.1	**-2.02**
Tülu-3-8B	AI2	12.0	2.9	**-9.04**	-76	22.8	-17.1	39.0	15.5	**-23.50**	30.0	13.5	**-16.49**
OLMo-2-13B	AI2	12.5	5.8	**-6.76**	-54	23.1	-12.5	34.0	16.7	**-17.30**	26.1	16.3	**-9.85**
OLMo-2-32B	AI2	13.8	8.6	**-5.17**	-37	26.8	-10.1	30.5	17.0	**-13.53**	32.4	18.0	**-14.41**
Llama-3.1-8B	Meta	12.0	7.3	**-4.69**	-39	22.8	-6.4	39.0	14.6	**-24.46**	30.0	30.4	+0.45
OLMo-3-7B	AI2	14.0	9.5	**-4.54**	-32	24.8	-8.2	17.6	18.1	+0.54§	21.3	23.6	+2.37
Qwen2.5-72B	Alibaba	10.9	6.7	**-4.14**	-38	20.2	-14.0	23.4	18.1	**-5.26**	22.1	13.1	**-9.04**
Llama-3.1-70B	Meta	11.7	7.8	**-3.86**	-33	27.1	-8.8	34.5	15.0	**-19.51**	21.5	28.5	+7.04
Qwen2.5-7B	Alibaba	10.1	6.5	**-3.62**	-36	17.6	-9.4	19.0	17.4	-1.63	23.8	11.1	**-12.70**
Qwen2.5-14B	Alibaba	10.7	7.2	**-3.44**	-32	17.9	-13.5	21.4	18.7	-2.61	26.7	15.0	**-11.63**
Qwen2.5-3B	Alibaba	10.3	7.3	**-3.06**	-30	17.6	-8.5	26.0	18.2	**-7.85**	21.4	14.3	**-7.09**
Gemma-3-4B	Google	11.3	8.3	**-3.02**	-27	22.9	-9.8	29.8	19.5	**-10.32**§	23.2	45.6	+22.41
Qwen2.5-32B	Alibaba	10.8	8.0	**-2.77**	-26	21.1	-15.4	23.0	17.7	**-5.25**	27.4	16.1	**-11.29**
Gemma-3-27B	Google	11.5	8.9	**-2.59**	-23	28.5	-16.2	24.4	17.1	**-7.33**§	24.5	51.3	+26.79
Gemma-3-12B	Google	12.2	9.9	**-2.33**	-19	25.6	-16.8	25.6	16.4	**-9.18**§	24.3	46.4	+22.17
Qwen2.5-1.5B	Alibaba	10.1	8.3	**-1.77**	-17	20.4	-5.3	D-SHELF			23.3	16.5	**-6.78**

package's own frame lists agreement per class and per checkpoint as unmeasurable here, and a 95% interval on a rate of this kind spans about ± 0.14. The cue classifier of Appendix D.1 labels 17 of the same 100 sentences. Its generic label matches the labeller in 5 of 8 and its addressed label in 6 of 9 (model-clustered 95% intervals $[0.25, 1.00]$ and $[0.31, 1.00]$), and the labeller read 78 of the 83 it leaves undetermined as addressed or generic.

D MEASUREMENTS AND CONTROLS BEHIND THE MAIN TEXT

Table A4 below gives Table 1 of the main text with every column restored, the ELICIT-99 and in-context assistant prompts among them, and Table A5 gives the contrast drawn in Figure 2(b) with its prompt-clustered intervals, the two template levels, the retention columns and the task-matched column. Nothing in them is new; the main-text table and figure are the same measurements with the extremes kept in view. Several tables in this appendix print the registered reading, taken at the single sampling seed the panel was generated at, rather than the four further sampling seeds that Table 1 pools: Tables A9, A10 and A15, and the unnumbered tables under *Prose lines alone, for both persons* and *The second person by model and protocol*. That reading is given in full in *The panel at one sampling seed* (Appendix D.2); in all sixteen models its sign is the same as in Table 1.

Notes on Table A4. The paired within-prompt placebo replaced the registered one after the first count was read. It is a required control outside the bar, which names only the interval, and on top of the interval it is cleared by 16/16 on the debate prompts, 9/15 on ELICIT-99 (the placebo band is much wider there) and 10/16 on the in-context assistant prompt (identical string). The same contrast is read on three sets of cells, so it carries three values: this table pools four further sampling seeds over all cells, Table A5 reads one seed on the cells non-degenerate in all four quadrants, and Table A6 reads one seed on all cells, unfiltered. PROTOCOL: on the debate prompts every checkpoint, base and instruct alike, is prompted with the same raw-continuation text scaffold; no chat template is applied on either checkpoint (0 of the 133 generation sets stored on this protocol, which also cover the checkpoint sequences, the training runs and the four models of Appendix H, carries a template marker in its prompt, although 39 of them come from checkpoints that ship one), so that contrast is not confounded with a change of prompt format. †On ELICIT-99 the protocol follows each checkpoint: a checkpoint with a chat template is prompted through it and one without through the raw-continuation scaffold, so in 10 of the 15 measured models the two checkpoints differ in prompt format as well as in training, and **a format change is not separated from alignment on this prompt set**. The 5 models whose checkpoints share a format are Qwen2.5-3B, Qwen2.5-7B, Qwen2.5-14B, Qwen2.5-32B, Qwen2.5-72B. *On the in-context assistant prompt both checkpoints are read on the identical string, so the only difference between them is the weights, and the columns are read on all sixteen prefix conditions (0 cells truncate). *__The in-context assistant prompt, in full.__ Both checkpoints are read on the identical string, so the only difference between them is the weights. The task lifts every base into 21.3–32.4 per thousand across the 16 scored models, above the 16.3 of the three example answers; the lowest base sits only 0.02 per thousand above the 21.25 of the prompt taken whole (its instruction header alone runs at 60.2) — a margin below this design's own resolution — so the copying account is excluded at the answers and left open at the whole prefix. The contrast there is 10 down, 5 up and 1 FORMAT-ONLY (a

Table A5: **The same contrast under three prompt protocols.** Every model with its prompt-clustered 95% CI, plus the two template levels and the retention columns; Figure A1 draws the raw-continuation and chat-template contrasts as levels, in the row order of Table 1. ΔM_1 is instruct − base in second-person forms per thousand tokens. A cell (prefix condition × prompt × draw) enters the first two columns only if it is non-degenerate in all four quadrants, so those two are read on the *same* cells; *kept* is the surviving share and *dgn.* the base outputs' degeneracy share under the template. The daggered rows, the task prompt set and the unfiltered reading are below.

	M_1 under template		ΔM_1 (instruct − base)				
model	base	inst.	raw-continuation	chat-template	task*	dgn.	kept
Gemma-3-27B	8.9	41.2	−2.81 [−5.25,−0.61]	+32.27 [+30.42,+33.90]	+16.67 [+15.19,+18.26]	0.35	44%
Gemma-3-4B	13.8	40.1	−3.17 [−5.04,−1.44]	+26.35 [+24.61,+27.89]	+15.75 [+14.17,+17.28]	0.42	41%
Gemma-3-12B	11.5	36.2	−3.28 [−5.05,−1.67]	+24.61 [+22.39,+26.61]	+11.34 [+9.70,+12.94]	0.28	47%
Llama-3.1-70B	8.3	23.1	−4.31 [−5.12,−3.51]	+14.76 [+13.51,+16.08]	+1.84 [+0.97,+2.75]	0.41	37%
OLMo-3-7B	17.5	27.4	−4.52 [−5.43,−3.53]	+9.92 [+8.63,+11.20]	+5.76 [+4.59,+7.00]	0.17	67%
Llama-3.1-8B§	12.5	20.0	−5.44 [−6.76,−4.22]	+7.51 [+6.26,+8.62]	−10.05 [−11.08,−9.11]	0.39	41%
Qwen2.5-1.5B†	15.0	16.4	−2.31 [−4.02,−0.97]	+1.43 [+0.05,+2.92]	−8.65 [−9.55,−7.83]	0.54	28%
Mistral-7B-v0.3	13.2	13.1	−9.99 [−11.50,−8.70]	−0.11 [−2.62,+1.85]	−16.73 [−17.93,−15.54]	0.17	62%
Qwen2.5-72B	10.5	10.3	−3.87 [−6.13,−2.03]	−0.17 [−1.46,+1.10]	−12.22 [−13.25,−11.25]	0.13	58%
Qwen2.5-3B	14.0	11.1	−3.27 [−4.80,−1.84]	−2.90 [−3.99,−1.89]	−10.73 [−11.45,−10.05]	0.20	53%
Qwen2.5-14B	16.3	11.2	−3.90 [−5.86,−2.24]	−5.10 [−6.75,−3.69]	−15.81 [−16.68,−14.99]	0.19	48%
Qwen2.5-7B	15.9	8.7	−4.28 [−6.13,−2.52]	−7.23 [−8.53,−5.97]	−15.33 [−16.06,−14.62]	0.31	47%
Qwen2.5-32B	19.4	12.1	−2.46 [−3.88,−1.26]	−7.30 [−9.08,−5.64]	−15.56 [−16.58,−14.72]	0.15	54%
OLMo-2-32B	17.5	4.8	−4.75 [−5.57,−3.94]	−12.65 [−14.67,−10.97]	−27.58 [−28.70,−26.27]	0.30	51%
OLMo-2-13B	18.6	4.0	−6.56 [−7.68,−5.55]	−14.55 [−16.32,−12.95]	−22.00 [−22.67,−21.41]	0.36	46%
Tülu-3-8B§	20.4	2.0	−9.29 [−10.83,−7.85]	−18.41 [−19.88,−16.96]	−28.01 [−28.62,−27.43]	0.34	46%

raw-continuation: 16/16 disjoint and down. chat template: 7 up, 7 down, 2 null. task: 11 down, 5 up, 0 format-only.

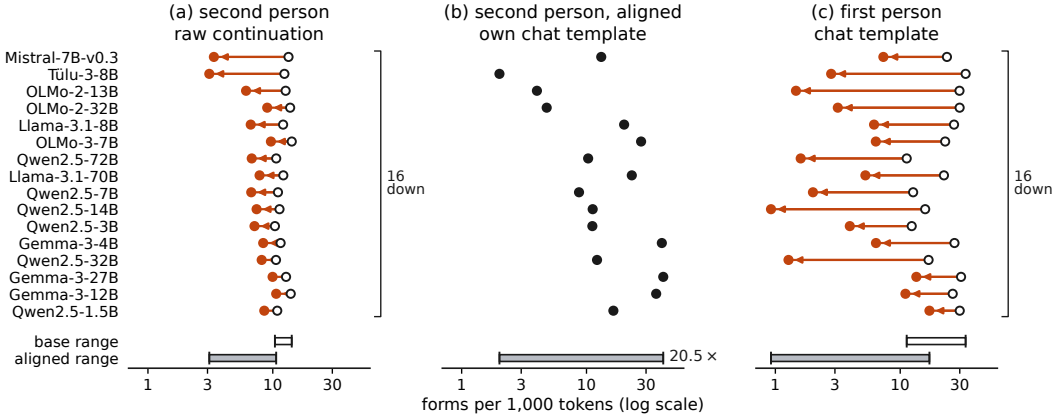

Figure A1: **Second- and first-person levels, base to aligned.** The levels behind Figure 2. (a) Second-person rate on raw continuation, base (open) to aligned (filled). (b) Aligned second-person level under each model's own chat template, with the shift in Figure A2. (c) Derived first-person rate under the chat template, base to aligned. All three panels read one sampling seed on the cells non-degenerate in all four quadrants (Appendix D.1), so panel (a) is not Table 1, which pools four further seeds over all cells. Forms per 1,000 tokens. Dotted lines mark intervals covering zero, and bars mark each group's range. Rows follow Table 1.

shift below the model's minimum detectable effect) across 16 of 16 models with 0 missing generation sets. The reading against each aligned checkpoint's own template is in Table A5. §These rows reach the 1,024-token cap on ELICIT-99 in at least a tenth of aligned draws: Gemma-3-4B (37% aligned, 21% base), OLMo-3-7B (24% aligned, 79% base), Gemma-3-12B (68% aligned, 43% base) and Gemma-3-27B (74% aligned, 56% base). Part of their density there is read on truncated text (Appendix Table A8). We print the stamp rather than drop the rows, and we do not claim a direction for the bias it leaves.

Table A5: symbols, the in-context assistant prompt and the unfiltered reading. Tülu-3-8B and Llama-3.1-8B descend from the *same* base checkpoint, and **the two pipelines answer at** 2.01 **and** 20.02 **second persons per thousand in their own formats** — the sign here is set by the pipeline, not by the base weights, and that statement needs no base outputs. † marks the one positive whose CI lower bound is +0.05 at the lowest retention. **task* is a separate measurement on a separate prompt set and is *not* read on the filtered cells of the first two columns: the base outputs are generated with the assistant task as three in-context examples (Lin

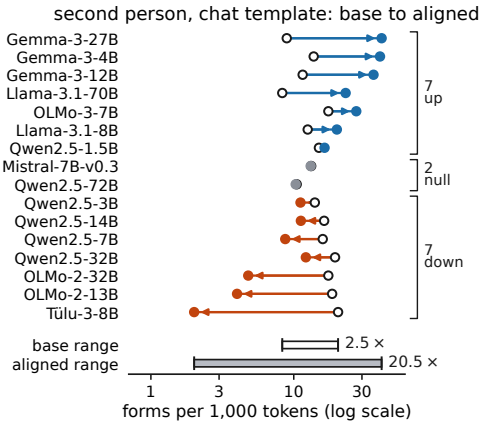

Figure A2: **Second-person shift through each model's own template, base to aligned.** The shift behind Figure A1(b). Second-person forms per thousand tokens through each aligned model's own template, base (open) to aligned (filled), on the filtered cells of Figure A1, rows sorted by the shift; dotted lines mark intervals spanning zero, bars each band's range.

et al., 2024) and the aligned outputs are read through the model's own template, so each checkpoint is prompted in the format it was trained for while the task is held fixed. Read against the base outputs instead, the same weights differ by 7.91 points between the two chat formats, but that checkpoint is prompted through a template it never saw and invents turn markers of its own, so we do not rest on it. *dgn.* is the base outputs' degeneracy share under the template. Read *without* the four-quadrant filter the chat-template column is 6 up and 10 down *by sign*, or 6 up, 8 down and 2 null once each interval is read; the only model whose sign changes is Qwen2.5-1.5B — the row already marked † — so the filter does not create the split. **Sign counts and interval counts are different numbers** and both are given here, because the filtered column above (7/7/2) is classified by interval. It carries the same distinct-4 floor per checkpoint and no checkpoint was dropped by it. Read on all sixteen prefix conditions at the 2,048-token window, where no prompt is truncated, its outcome is 11 down, 5 up and 0 FORMAT-ONLY across 16 of 16 models.

Notes on Table A6. **The blocks marked (tmpl.) read the same checkpoints through the chat template, and on four of the five the protocol moves the localization the same way**: **the supervised stage's share grows and the preference stage's share shrinks**, from 102.1%, 76.0%, 72.9%, 66.1%, 85.9% on raw continuation to −0.4%, 42.4%, 104.9%, 8.1%, 16.8% under the chat template (Zephyr-7B, Tülu-3-8B, Tülu-3-70B, OLMo-2-7B, OLMo-2-13B). On Tülu-3-70B the preference stage's share rises instead, from 72.9% to 104.9%: under the template the preference stage carries the drop while the supervised stage does not clear its interval. On Zephyr-7B the exchange is complete and the two stages swap which one is null; on the other four the preference stage still clears its interval, so the pre-registered outcome name does not change there. **The base→SFT row of a (tmpl.) block is the one asymmetric comparison in this table**: these five bases ship no chat template of their own (measured), so that row reads the base checkpoint through a template it never saw, while every preference row compares two checkpoints that both carry one. OLMo-2-32B and OLMo-3-7B have no chat-template block: their checkpoints do not all ship the same chat template, so their stages cannot be read on one string. **The chat-template rows here and in Table A5 are not the same reading**: this table is computed on all cells, that one only on cells that are non-degenerate in all four quadrants, so the same model reads −17.17 here and −18.41 there. The same contrast thus appears on three sets of cells: Table 1 pools four further sampling seeds over all cells, Table A5 reads one seed on the filtered cells, and this table reads one seed on all cells, unfiltered. The absolute levels behind these shares are in Appendix D. Which preference objective is used is itself a measured choice (Ivison et al., 2024; 2023; Dong et al., 2023a).

Scale sequences behind the supervised stage's share. **The supervised stage's share does not follow scale.** Within OLMo-2 — one pipeline read at three scales — the supervised stage carries 31%, 11%, 45% of the end-to-end withdrawal at 7B, 13B, 32B, which is not monotone in scale, while the withdrawal itself does fall monotonically (10.5, 6.3, 4.8 per thousand). Tülu-3 gives two points in the same direction of scale and the share rises rather than falls: 19% at 8B against 27% at 70B. **Size moves the amount withdrawn; *within a pipeline* it never makes the supervised stage the larger one.**

Table A6: **How the training stages share the withdrawal, in full.** The training stages of Table 2 with prompt-clustered 95% intervals, the change in first-person density and in the force total, and the end-to-end sum for each checkpoint sequence; the two sequences outside the AI2 and Zephyr pipelines follow in their own block. Blocks marked (tmpl.) read the same checkpoints through the chat template. Bold marks an interval excluding zero.

sequence	stage	ΔM_1	95% CI	Δ1st	Δforce
Tülu-3-8B	base→SFT	**-1.63**	$[-2.77, -0.39]$	-6.9	-0.0027
	SFT→DPO	**-6.62**	$[-8.10, -5.21]$	-9.9	-0.0061
	DPO→RL	**-0.46**	$[-0.77, -0.19]$	-0.3	+0.0005
	end to end	-8.71	$[-9.99, -7.48]$		
OLMo-2-13B	base→SFT	**-0.67**	$[-1.27, -0.07]$	-2.6	-0.0031
	SFT→DPO	**-5.39**	$[-6.24, -4.66]$	-9.5	+0.0012
	DPO→Instruct	-0.22	$[-0.56, +0.09]$	-0.5	-0.0016
	end to end	-6.28	$[-7.36, -5.38]$		
OLMo-2-32B	base→SFT	**-2.16**	$[-2.68, -1.65]$	-2.7	-0.0050
	SFT→DPO	**-2.68**	$[-3.49, -1.89]$	-7.7	-0.0000
	DPO→Instruct	+0.07	$[-0.31, +0.45]$	-0.4	+0.0001
	end to end	-4.77	$[-5.49, -3.99]$		
Zephyr-7B	base→SFT	+0.08	$[-1.01, +1.47]$	+0.6	-0.0000
	SFT→DPO	**-4.06**	$[-5.11, -3.06]$	-4.9	-0.0033
	end to end	-3.98	$[-5.16, -2.64]$		
OLMo-3-7B	base→SFT	**-2.62**	$[-3.46, -1.77]$	-4.0	-0.0013
	SFT→DPO	**-1.52**	$[-2.08, -0.96]$	-4.9	+0.0019
	DPO→RLVR	**-0.56**	$[-1.07, -0.07]$	+0.9	+0.0002
	end to end	-4.70	$[-5.68, -3.65]$		
Tülu-3-70B	base→SFT	**-1.81**	$[-2.71, -0.73]$	-7.6	-0.0024
	SFT→DPO	**-4.87**	$[-6.03, -3.86]$	-11.1	-0.0025
	end to end	-6.67	$[-7.52, -5.89]$		
OLMo-2-7B	base→SFT	**-3.24**	$[-3.86, -2.56]$	-6.3	-0.0088
	SFT→DPO	**-6.95**	$[-8.20, -5.80]$	-11.6	-0.0006
	DPO→Instruct	**-0.33**	$[-0.56, -0.11]$	-0.4	-0.0020
	end to end	-10.51	$[-11.48, -9.58]$		
Outside the AI2 and Zephyr pipelines (raw continuation)					
SmolLM2-1.7B	base→SFT	**-3.34**	$[-4.37, -2.33]$	—	—
	SFT→DPO	+0.07	$[-0.44, +0.60]$	—	—
	end to end	-3.26	—		
SmolLM2-1.7B: second-person density $14.29 \rightarrow 10.96 \rightarrow 11.03$ per thousand (base, SFT, DPO); paired-placebo 95th percentile 1.13 and 1.10 for the two stages (200 draws); one generation seed, 12 generations per prompt in each prefix condition; intervals from 1000 prompt-clustered resamples; first-person and force changes not computed.					
NeuralHermes-2.5	base→SFT	**-5.48**	$[-6.53, -4.37]$	—	—
	SFT→DPO	+0.15	$[-0.34, +0.57]$	—	—
	end to end	-5.33	—		
NeuralHermes-2.5: second-person density $13.44 \rightarrow 7.96 \rightarrow 8.11$ per thousand (base, SFT, DPO); paired-placebo 95th percentile 1.10 and 0.89 for the two stages (200 draws); one generation seed, 12 generations per prompt in each prefix condition; intervals from 1000 prompt-clustered resamples; first-person and force changes not computed.					

Table A6 (continued).

sequence	stage	ΔM_1	95% CI	Δ1st	Δforce
Zephyr-7B (tmpl.)	base→SFT	**-20.68**	$[-21.60, -19.79]$	-27.5	-0.0221
	SFT→DPO	+0.08	$[-0.82, +0.83]$	-2.2	+0.0003
	end to end	-20.60	$[-21.78, -19.47]$		
Tülu-3-8B (tmpl.)	base→SFT	**-9.55**	$[-11.39, -7.77]$	-20.1	-0.0133
	SFT→DPO	**-7.28**	$[-8.93, -5.80]$	-9.2	-0.0041
	DPO→RL	**-0.34**	$[-0.48, -0.21]$	-0.3	+0.0004
	end to end	-17.17	$[-18.33, -15.96]$		
OLMo-2-13B (tmpl.)	base→SFT	**-10.98**	$[-12.17, -9.81]$	-23.7	-0.0154
	SFT→DPO	**-2.27**	$[-3.07, -1.49]$	-3.4	-0.0001
	DPO→Instruct	**-0.32**	$[-0.43, -0.20]$	-0.3	-0.0003
	end to end	-13.57	$[-14.99, -12.22]$		
OLMo-2-7B (tmpl.)	base→SFT	**-19.25**	$[-20.75, -17.86]$	-22.0	-0.0243
	SFT→DPO	**-1.70**	$[-2.26, -1.13]$	-2.3	-0.0001
	DPO→Instruct	**-0.15**	$[-0.23, -0.08]$	-0.1	-0.0008
	end to end	-21.10	$[-22.83, -19.58]$		
Tülu-3-70B (tmpl.)	base→SFT	+0.25	$[-1.10, +1.40]$	-5.0	-0.0062
	SFT→DPO	**-5.45**	$[-6.18, -4.76]$	-5.6	-0.0018
	end to end	-5.20	$[-6.63, -4.01]$		

Bold: interval excludes zero. On the raw-continuation prompt the preference stage clears its interval on all 7 AI2 and Zephyr checkpoint sequences of this table and on neither of the two outside them. Under the template we can read 5 of them as a single string; the preference stage still clears its interval on 4 of those 5, and on Zephyr-7B the two stages swap which one is null.

D.1 INSTRUMENTS AND SETUP

Table A7 lists the terms the main text uses.

Table A7: **Terms used in the main text.**

model	a base and a released post-trained checkpoint, compared as a pair (16 in the panel)
family	a model series (six: Llama-3.1, Qwen2.5, Gemma-3, Mistral, OLMo, Tülu)
base / aligned checkpoint	a model's pretrained and released post-trained checkpoint
post-training pipeline	the data, stages and objectives that turn a base into an aligned checkpoint
training stage	supervised fine-tuning (SFT), preference optimisation or a later RL stage
checkpoint sequence	the checkpoints one pipeline releases after each training stage
lineage	the lab that trained a model (five, OLMo and Tülu both AI2's; six if OLMo-3 counts apart)
prompt set	the debate prompts (34), ELICIT-99 (99) or the ordinary instructions (300)
prompt format	raw continuation, the aligned chat template given to both checkpoints, or the in-context assistant prompt given to the base checkpoint
prefix condition	one of 16 fixed prefixes placed before each debate prompt
first- / second-person rate	forms of "I" and "we", or of "you", per 1,000 tokens (M_1: second person)
withdrawal	a drop in a default rate that one sentence in the prompt can reverse
interpersonal register	whether a text has a writer and a reader in it
paired placebo	each prompt's outputs split in half 200 times, with an effect counted only above the 95th percentile of those differences
filtered set	the prompt, prefix and sample combinations whose four generations (base and aligned, raw continuation and chat template) are all non-degenerate, that is, none is flagged as empty, too short, repetitive, dominated by function words or not in Latin script
echo	an output that repeats the end of its own prompt
fabricated turn marker	a turn label, such as "User:", that the model writes itself
pronoun gap	the difference in second-person rate between the chosen and the rejected response of a preference pair
appendix names, kept there	M_1 second-person rate, M_2 addressee-directed acts, M_3 impersonal-correction share, M_4 first-person volitionals, M_5 distancing index; FORMAT-ONLY a shift the prompt format alone produces; UNMEASURABLE an outcome the design cannot reach

Generation caps, and how often a generation reaches one. The debate prompts, in all three prompt formats, and the ordinary instructions are generated with a cap of 160 new tokens, and ELICIT-99 with 1,024. Table A8 gives, for each model and checkpoint, the share of generations that reach the cap, counted by re-tokenizing each output with its own checkpoint's tokenizer and counting a generation that ends within three tokens of the cap. On ELICIT-99, where the generator also recorded its own token count, the two agree in at least 97% of generations on 31 of the 32 generation sets; the exception is one base checkpoint whose recorded count is itself unusable. **Most debate-prompt generations run to the cap on both checkpoints**: on raw continuation a median 76% of base and 95% of aligned generations do (62–100% and 62–100%), so rates on this prompt set are rates over the first 160 tokens of a continuation that would have run on. On the ordinary instructions the aligned answers reach the cap in 55–71% of cases, and the base continuations in 74–100% before they are cut at their first code fence or new query. In Gemma-3-4B, -12B and -27B the aligned ELICIT-99 outputs reach the 1,024-token cap in 37, 68 and 74% of draws (the base outputs in 21, 43 and 56%), so part of their density there is read on truncated text; OLMo-3-7B's (24%) is the only other aligned checkpoint above a tenth, and every other aligned checkpoint stays under 1%.

(i) Mechanical form counters. Five sentence-level counters over spaCy (Honnibal et al., 2020) POS and dependency parses: bare imperative, second-person obligation, hedge, interrogative directive, hortative (categories close to those used in speech-act annotation, International Organization for Standardization, 2020; Conrad et al., 2026). Validity was measured before use on a 70-item set labelled by the assistant (an LLM agent), stratified by surface cue: balanced accuracy 0.958 (bare imperative) and 1.000 for hedge, interrogative directive and hortative.

The obligation counter carries no outcome. It reached only two gold positives from two conditions, because second-person obligation is essentially absent from this corpus (0 of 68 natural items). **This is a measurement, not a defect of the counter.**

Table A8: **Share of generations that reach the token cap** (%), base and aligned outputs, counted by re-tokenizing each output with its own checkpoint's tokenizer. Cap 160 new tokens except ELICIT-99 (1,024). Raw, template and in-context are the three formats of the debate prompts; the in-context aligned checkpoint reads the identical string. Instruction base outputs are counted before they are cut.

	raw		template		in-context		instructions		ELICIT-99	
model	base	al.	base	al.	base	al.	base	al.	base	al.
Gemma-3-12B	70	93	89	93	97	99	98	65	43	68
Gemma-3-27B	70	91	95	98	97	100	97	68	56	74
Gemma-3-4B	75	88	79	96	96	99	98	67	21	37
Llama-3.1-70B	72	97	86	92	99	100	100	62	22	0
Llama-3.1-8B	79	100	84	94	99	100	100	62	20	0
Mistral-7B-v0.3	76	71	79	98	100	100	100	61	56	0
OLMo-3-7B	100	87	100	100	100	100	100	68	79	24
OLMo-2-13B	77	67	72	100	95	100	75	71	2	0
OLMo-2-32B	75	62	54	100	77	100	74	66	6	0
Qwen2.5-1.5B	84	97	81	63	96	94	96	55	76	0
Qwen2.5-14B	77	100	64	100	96	100	96	60	1	0
Qwen2.5-32B	74	100	66	100	88	100	92	61	13	0
Qwen2.5-3B	85	100	64	100	95	100	94	64	10	0
Qwen2.5-72B	62	99	86	99	98	100	97	62	12	1
Qwen2.5-7B	78	99	75	100	98	100	97	62	4	0
Tülu-3-8B	79	65	82	99	99	100	100	64	20	0
median	76	95	80	99	97	100	97	63	20	0

A bar can sit below its own floor. A plain-accuracy bar of 0.90 is passed by a counter that never fires whenever the positive rate is under 10%. We measured that floor (0.971 and 0.900 for two of the five counters) and therefore report balanced accuracy. The bar value did not move, but its form was declared before measurement.

(ii) Two prompt protocols, raw continuation and chat template. The chat template is not a nuisance parameter. Reading the same stage under the raw-continuation prompt and under each model's own chat template changes the *name* of the outcome for four of twelve counter×checkpoint sequence cells, reported in the repository. Prompt format is known to move evaluation outcomes on its own (Sclar et al., 2023; Voronov et al., 2024), which is why we read every stage under both protocols rather than picking one.

Setup of the depersonalization contrast (§3). For each of sixteen models (Table 1, five lineages, 1.5B–72B) we compare the base checkpoint with its instruct/aligned checkpoint on the two prompt sets of this section; the in-context assistant prompt is read in §4 and in full below. One is the debate prompts, which request no task. The other is ELICIT-99, built to afford status acts (its construction is below, in *How the two prompt sets and the in-context frame are built*). We count, per thousand tokens, second-person forms (M_1), first-person forms (directly or derived, as §2 states), addressee-directed acts (M_2), the share of corrections that name no person (M_3), first-person volitionals (M_4) and a distancing index (M_5). We also count a *force total* (imperatives, deontic modals and evaluative predicates, personal and impersonal). CIs are prompt-clustered. A paired placebo splits each prompt's own generations and must be exceeded. The counters were checked against act-level and human labels rather than assumed: the correction counter under-detects relative to a judge, and 68% (26/38) of its flags are corrections to a human reader, so M_3 is reported over counter-flagged corrections with its precision (Appendix C).

The five counters, their definitions and what each one is blind to. The register battery is five deterministic counters over parsed sentences, reported per thousand tokens or as shares. M_1 counts second-person forms and M_4 first-person volitionals. M_1 **in full**: its numerator is every case-insensitive, word-bounded match of *you*, *your*, *yours*, *yourself* and *yourselves*, plus one per sentence that opens with an address form (*hey*, *hi*, *hello*, *look*, *listen*, *friend*, *dear reader*) followed by a comma or an exclamation mark. Contractions count through their pronoun, since the apostrophe is a word boundary: *you're*, *you've*, *you'll* and *you'd* each contribute one. **What is outside it**: a second person carried by a bare imperative, which has no subject and is counted by M_2 and the force total instead; archaic and non-standard forms (*thou*, *thee*, *u*); and vocatives outside the seven address forms above. The counter does not resolve reference, so a generic *you* and an addressed *you* count alike. **Its denominator** is spaCy tokens with whitespace removed, so punctuation, list bullets and markdown markers are all counted as tokens; a sentence containing nothing but punctuation is dropped before either side is counted. M_2 counts addressee-directed acts: a deontic or evaluative predicate with a second-person subject, or a prose-standing imperative. M_3 is the share of corrections that name no person, and M_5 a distancing index of third person, articles and prepositions against first and second. Alongside them we count a *force total* over imperatives, deontic modals and evaluative predicates, personal and impersonal alike. It moves only if the illocutionary layer moves rather than the interpersonal one. Because a bare imperative has no subject

by definition, the imperative part of M_2 is built from prose-standing imperatives rather than from second-person ones, and the distinction is measured rather than assumed.

How the two prompt sets and the in-context frame are built, and at what cost. ELICIT-99 affords status acts through false premises, bad plans, requests to ratify reasoning, borderline requests and contested opinions, with the target vocabulary excluded from the prompts by a leak gate derived from the stimulus file's own rule line. The debate prompts answer what a model does when nothing *asks* it for a status act — the prompts request no task; the prefix conditions are themselves status-loaded and are held fixed across checkpoints, so they block rather than confound the contrast. ELICIT-99 answers what it does when something does. Both are generated at temperature 0.9 and top-p 0.95: the debate prompt set is 34 debate-continuation prompts $\times$ 16 prefix conditions $\times$ 12 draws, 192 generations per prompt ($6{,}528$ responses per checkpoint) at one seed, and ELICIT-99 is 99 prompts at four sampling seeds per stage, under a per-stage token budget raised once before any outcome was read (Appendix A).

Our own hedge counter measured the wrong thing. Our hedge counter was first read as a "hedge axis". Split by its own word lists, 82–83% of what it counted was epistemic hedging (*maybe*, *perhaps*) and about 2% was addressee-directed request framing, and the split is stable across all five training runs we decomposed. We therefore read that axis as an *epistemic* hedge axis throughout, and we separate the two vocabularies in the released code. The correction changes no outcome. It changes what the outcome is about.

Why the act-level threshold instability is the expected one. The agreement figures themselves are in Appendix C; we report status-act *presence* and decline to report a status-act *change* under alignment, which is below our minimum detectable effect and judge-dependent. That instability is the expected one rather than a surprise: model judges are known to carry position and verbosity biases, to prefer their own generations (Panickssery et al., 2024), and to shift with the order in which candidates are shown (Zheng et al., 2023; Wang et al., 2023), which is why every paired judgement in this paper is shown in both orders and why the judge is a secondary instrument throughout. The agreement figures, and every other annotation layer in this paper, are collected in Appendix C.

The lexicon's failure, and where the caution is shared. A word-level dominance lexicon fails on this contrast, and the failure is one of category, not of noise: a dominance norm records what a reader feels *reading* a word, while writer-dominance is carried by structure. **Independent work reaches the same caution from the annotation side**: scalar dominance labels are unreliable where paired comparisons are not (Buechel & Hahn, 2017; Chrzan et al., 2026), and politeness strategy is a poor stand-in for force (Danescu-Niculescu-Mizil et al., 2013; Zhao & Hawkins, 2025).

The denominator swap, in full. Read per sentence instead of per thousand tokens, the register and force counts swap places and force separates in more models than register (Figure A4); this is why the body reports no magnitude ratio between the two layers. **The sign of the register fall survives the swap** — it is downward wherever it separates — and that invariance, not the ratio, is what the body carries.

Sentence embeddings are blind to the contrast, and to mood. Two exploratory panels agree. On a panel of 596 judged passages, a variance decomposition of a blind judge's dominance scores gives sentence embeddings $R^2 = 0.003$, against 0.1473 for deterministic form counters and 0.005 for the lexicon: form explains the same decisions $29.5\times$ better than the lexicon does. On a second panel of 598 passages, the embedding direction *correlates* with the judge's scores at $\rho = +0.024$, inside an anisotropy-matched placebo band drawn from the cloud's own distribution ($K = 200$; $|\rho|$ 95th percentile 0.1468). These panels are exploratory and carry no pre-registered bar; we report them because they decide which instrument the rest of the paper may use, not because they establish a magnitude. Mood is a syntactic contrast that survives almost no pooling, and an embedding trained for semantic similarity is built to discard it.

Generic *you* and addressed *you*: what the split can and cannot say. M_1 counts second-person forms and does not ask whom they name. We split the sentences carrying one into ADDRESSEE, GENERIC and UNDETERMINED with a cue list fixed before the measurement, with ADDRESSEE taking precedence where both fire. Across the 16 models the generic count falls in 16 and the addressed count in 14. A preliminary check by the assistant (an LLM agent, not a human labeller) on 40 sentences found the generic label right $12/13$ of the time and the addressed label $7/14$. In that check the deontic cue (*you should*, *you must*) fired on generic advice as readily as on advice to a reader. Of the 13 the cues left UNDETERMINED, 13 were resolvable on reading (overall agreement $19/40$). On the authors' labels for 100 second-person sentences (Appendix C) the classifier labels only 17. Its generic label is right $5/8$ of the time and its addressed label $6/9$. Of the 83 it leaves UNDETERMINED, the authors read 78 as addressed or generic. **The asymmetry of the preliminary check is not reproduced on human labels.** UNDETERMINED is 76% of all second-person sentences in the panel. Both halves are read from the judge-labelled check of Appendix C (§2). **We do not claim the withdrawal is**

confined to either half: on human labels neither of the classifier's two labels is right often enough to carry that claim.

D.2 Robustness on the raw-continuation prompt

Readings the main text compresses. Section 3 states that the fall survives several further readings; the counts are here. On prose lines alone (list, numbered and heading lines dropped, *US* not counted) the first person still falls in 16 of 16, every interval excluding zero. On the four prefix conditions that carry no first person the fall holds in 15 of 16 (Qwen2.5-1.5B's interval covers zero). Counted directly, with outputs that echo the prompt removed and each output cut at its first fabricated turn marker, the count is still 16 of 16. On ELICIT-99 the first person falls in 7 of 15 models on borderline requests, where two rise and six show no clear change. The degeneracy rule written before counting flags 2.7% of aligned outputs (0.9 to 5.6% per model, against 2.2% of base outputs); with those removed the largest change in any model is 0.18, and a stricter rule that removes 52% of the aligned output also leaves 16 of 16. In the six models whose prose-line interval covers zero, 24 to 55% of aligned lines are list, numbered or heading lines. Of the three restoring sentences, the one that contains one "you" restores 1.7 times the drop (15 of 16), and naming the forms restores 3.1 times (16 of 16).

Reading the human-band figure: the rows, the shaded band and the readings. The seven human references of Figure A3 are, in its row order, Supreme Court justices, Reddit debate (ChangeMyView) and Supreme Court advocates (spoken and interactive, filled circles); wikiHow how-to, StackExchange answers and OpenStax textbook prose (written expository, open squares); and OASST1 volunteers writing assistant replies (an open square), the one human reference written for the same task the models are doing. The shaded band is the OpenStax textbook row. The readings, the counts behind each row and the sensitivity of the middle panel are given elsewhere in this appendix.

Force: a different stage, and ruler-dependent. The force total clears its own null in 10 of 16 models and in both directions, so it is not the one-directional object that register is, and the stage at which it moves is not the stage at which the persons leave. **The comparison between the two layers does not survive a change of denominator** and we report no ratio between them. **What does not move with the ruler is the *sign* of the register fall**, which is downward wherever it separates. The ruler comparison, the model-level numbers and the dilution decomposition are in this appendix.

Bare imperatives: the rate does not move, the composition does. Across 3.8 million parsed sentences from the panel and its extensions, bare imperatives are 1.04% of sentences and two thirds (66.8%) stand in prose rather than in a numbered procedure. The rate shifts by at most 0.94 points in any model; the procedural share shifts by much more (Appendix Table A23).

The prefix conditions, and the headline read without them. A prefix condition is a procedure, not a fixed string. Twelve of them stay on one axis: six are a single utterance (either pole of arousal, dominance and tone) and six are four utterances from one pole. The other four mix two axes, two dominance and two arousal utterances. Every pole has a pool of five utterances; in draw c the condition's j-th slot takes pool item $(j + c) \bmod 5$ and the slots are shuffled with a seeded permutation, so across the twelve draws a condition cycles through its pools. The prefix is followed by the neutral frame "{thesis} is {stance}. Here is why:". Table A12 gives what each of the sixteen prefix conditions brings of its own once the shared thesis scaffold is removed, and Table A9 reads the debate-prompt contrast on the nine that bring no second person. Table A10 counts the second persons inside quoted speech model by model, and Table A11 crosses the three robustness checks of §3 model by model.

Second persons inside quoted speech. In the raw-continuation outputs the second persons are not in someone else's mouth. Running the chat-template outputs' check on the raw-continuation ones—quoted spans, speaker labels and dash-dialogue, and the complement clauses of speech verbs—puts 2% to 6% of the base outputs' second persons inside such a span (median 4%). **The chat-template outputs are not one figure but two regimes**: thirteen of its sixteen models put essentially none of their second persons after a fabricated turn marker, and the three whose templates write that boundary in plain text put 47–61% there — so the raw-continuation outputs' few per cent is read against a distribution with a floor at zero, not against a single chat-template number. **Read outside those spans the withdrawal is unchanged:** 16 **of** 16 **down and interval-disjoint.** A deliberately over-cutting ceiling reading, which also discards everything after the first blank line and so throws away ordinary paragraph breaks, leaves 15 of 16 (the exception is Qwen2.5-32B, whose interval then covers zero). **Two limits of this reading, named**: the detector under-counts — straight-quote dialogue and any speech verb outside our list are missed — so the in-span share is a *lower* bound on exactly the quantity we want small; and because M_1 is a density, excising a span removes tokens as well as second persons, so the direction of the change is not conservative by construction. **So we measured it**: the excision removes 3–7%

Table A9: **The headline read on the nine prefix conditions that carry no second person.** The debate prompts' 16 prefix conditions are a regulator we did not design out: seven of them contain second persons of their own (from 102 to 1,360 occurrences across the 408 prompt-draw cells), and a base checkpoint can copy them. Splitting each condition's own prefix from the shared thesis scaffold by longest common suffix leaves nine conditions with *none* (Table A12). Read on those nine alone, the withdrawal is still negative and interval-disjoint in 14 of the 16 models — **the direction is not an artifact of the prefix conditions, but part of the magnitude is**. The pre-registered $16/16$ count on the full panel remains the primary reading; this is a second, descriptive one.

model	full panel ΔM_1	nine you-free prefixes	95% CI	prefixes disj.
Mistral-7B-v0.3	-9.86	**-8.42**	[-10.42, -7.02]	16/16
Tülu-3-8B	-8.71	**-7.37**	[-8.83, -6.17]	16/16
OLMo-2-13B	-6.28	**-5.62**	[-6.83, -4.62]	16/16
OLMo-2-32B	-4.77	**-4.76**	[-5.51, -4.17]	13/16
Llama-3.1-8B	-4.77	**-4.66**	[-5.76, -3.71]	16/16
OLMo-3-7B	-4.70	**-4.08**	[-5.20, -3.10]	15/16
Qwen2.5-7B	-3.90	**-3.07**	[-4.87, -1.79]	10/16
Llama-3.1-70B	-3.78	**-3.22**	[-4.02, -2.52]	13/16
Qwen2.5-72B	-3.69	**-2.48**	[-4.62, -0.94]	8/16
Gemma-3-4B	-3.63	**-2.65**	[-4.19, -1.54]	10/16
Qwen2.5-14B	-3.49	**-1.85**	[-3.33, -0.80]	9/16
Qwen2.5-3B	-3.46	**-2.66**	[-4.09, -1.49]	11/16
Gemma-3-27B	-3.07	-1.53	[-4.18, +0.51]	14/16
Gemma-3-12B	-2.80	**-1.88**	[-3.72, -0.43]	13/16
Qwen2.5-32B	-2.18	**-1.47**	[-2.94, -0.42]	10/16
Qwen2.5-1.5B	-1.61	+1.08	[-0.04, +2.03]	7/16

Table A10: **The raw-continuation outputs' second persons, read outside quoted, attributed and reported speech.** Share of the base outputs' second-person tokens falling inside a detected span, and ΔM_1 recomputed with those spans removed from both checkpoints. *ceiling* additionally treats everything after the first blank line as another speaker's turn, which over-cuts. Intervals are prompt-clustered (500 resamples). Over-detection works against the result, not for it: it removes text and leaves less to count.

model	in-span %	tokens cut %	ΔM_1	outside spans	ceiling
Gemma-3-12B	4.0	4.8	-2.80	-2.57	-5.93
Gemma-3-27B	4.6	5.0	-3.07	-2.79	-4.78
Gemma-3-4B	3.1	4.1	-3.63	-3.39	-5.83
Llama-3.1-70B	4.8	5.2	-3.78	-3.75	-3.77
Llama-3.1-8B	4.0	4.8	-4.77	-4.80	-4.80
Mistral-7B-v0.3	3.1	3.6	-9.86	-9.90	-9.28
OLMo-3-7B	3.2	4.1	-4.70	-4.61	-4.52
OLMo-2-13B	4.0	5.3	-6.28	-6.34	-6.32
OLMo-2-32B	4.7	5.4	-4.77	-4.83	-3.80
Qwen2.5-1.5B	2.2	3.1	-1.61	-1.44	-1.58
Qwen2.5-14B	4.2	5.3	-3.49	-3.30	-2.30
Qwen2.5-32B	5.9	6.7	-2.18	-2.03	-0.32°
Qwen2.5-3B	2.2	3.2	-3.46	-3.35	-3.07
Qwen2.5-72B	4.3	4.9	-3.69	-3.60	-2.98
Qwen2.5-7B	2.0	2.9	-3.90	-3.83	-2.82
Tülu-3-8B	4.0	4.8	-8.71	-8.78	-8.55
down, interval-disjoint			**16/16**	**16/16**	**15/16**

$^{\circ}$ interval covers zero. Limits, named: straight-apostrophe quotation is not counted, the speech-verb list is our own, and unmarked turn transitions appear only in the ceiling column.

of the tokens and **raises** M_1 **in all** 16 **models**, which means the spans it removes are *less* second-person-dense than the prose around them — the second persons are not concentrated in the dialogue.

The first-person fall on the prefix conditions that carry no first person. Twelve of the 16 prefix conditions carry first-person forms of their own, so a base checkpoint could copy them. Four carry none (the arousal conditions, one- and four-utterance, in both directions), and two of these carry no second person either. We re-read the first-person change on those four alone (Table A13), on the original sampling seed, with the

Table A11: **The three robustness re-reads, crossed.** Each column is the model as it was read in the card that produced it: a per-sentence count whose interval excludes zero (Table A21), the nine prefix conditions that carry no second person (Table A9), and the four-seed re-read with its placebo band (Appendix A). The three are cleared by 12, 14 and 16 models and by the same 11 together.

model	per sentence	you-free prefixes	four seeds
Gemma-3-12B	✓	✓	✓
Gemma-3-27B	✓	—	✓
Gemma-3-4B	✓	✓	✓
Llama-3.1-70B	✓	✓	✓
Llama-3.1-8B	✓	✓	✓
Mistral-7B-v0.3	✓	✓	✓
OLMo-3-7B	✓	✓	✓
OLMo-2-13B	✓	✓	✓
OLMo-2-32B	✓	✓	✓
Qwen2.5-1.5B	—	—	✓
Qwen2.5-14B	—	✓	✓
Qwen2.5-32B	—	✓	✓
Qwen2.5-3B	✓	✓	✓
Qwen2.5-72B	—	✓	✓
Qwen2.5-7B	✓	✓	✓
Tülu-3-8B	✓	✓	✓
cleared / not cleared / not read	**12** / 4 / 0	**14** / 2 / 0	**16** / 0 / 0

Table A12: **What each prefix condition brings of its own.** Second- and first-person counts of the condition's prefix alone, after the shared thesis scaffold is removed, summed over the 408 prompt×draw cells it appears in. The nine conditions with no second person are the pool read in Table A9. **The first person is the larger stimulus and we did not control for it**: the dominance-low conditions carry it heavily.

condition	tokens	2nd person	M_1 of prefix	you-free
`ARO+x4`	60,772	1,360	22.38	–
`DOM-x2_ARO+x2`	67,878	884	13.02	–
`DOM+x2_ARO+x2`	64,410	680	10.56	–
`ARO+x1`	15,314	340	22.20	–
`DOM-x4`	75,358	340	4.51	–
`DOM-x2_ARO-x2`	66,620	204	3.06	–
`DOM-x1`	18,816	102	5.42	–
`ARO-x1`	14,498	0	0.00	✓
`DOM+x1`	17,014	0	0.00	✓
`ARO-x4`	58,086	0	0.00	✓
`DOM+x4`	68,558	0	0.00	✓
`DOM+x2_ARO-x2`	63,152	0	0.00	✓
`TON+x1`	16,844	0	0.00	✓
`TON-x1`	16,096	0	0.00	✓
`TON+x4`	67,776	0	0.00	✓
`TON-x4`	64,036	0	0.00	✓

prompt-clustered interval of the full panel (1,000 draws). The same loop reproduces the full panel's first-person change in all 16 models, to within 10^{-9}, before any subset is read. **On the four conditions the first person falls in** 15 **of** 16 **models**, counted directly or as derived, and Qwen2.5-1.5B is the exception (interval covering zero). On the two conditions that carry neither person the first person falls in 15 and rises in 1, and the second person falls in 10, rises in 1 and is unclear in 5. The four conditions hold a quarter of the data and the two an eighth, so intervals are wider and an unclear reading is a loss of power, not a finding of no change. The reading is descriptive: its rule was committed before the count, with no bar.

Raw continuation with the genre fixed by a header. The debate stubs leave the genre of the continuation open, so a model that answers in a less personal genre and a model that uses fewer persons within a genre are not separated. To separate them we re-ran raw continuation with a one-line bracketed header in front of the same 34 stubs: `[Personal blog post]`, `[Reply in a discussion forum]`, `[How-to guide]` and `[News report]`. None of the four contains a first- or second-person form. Both checkpoints read the identical plain string, with no chat template, at the panel's own decoding settings and 12 samples per

Table A13: **The first-person fall without first-person prefixes.** Aligned minus base per 1,000 tokens, raw continuation, original seed. *All 16*: every prefix condition. *Four*: the conditions with no first person. *Two*: the conditions with neither person. Δ1st is counted directly, with a capitalised *US* not counted. **Bold**: prompt-clustered 95% interval excludes zero.

	all 16 prefixes	four		two	
model	Δ1st	Δ1st	95% CI	ΔM_1	Δ1st
Mistral-7B-v0.3	**−15.34**	**−11.65**	[−13.92, −9.42]	**−6.35**	**−9.97**
Gemma-3-12B	**−16.62**	**−10.67**	[−12.55, −8.61]	−1.18	**−6.38**
Tülu-3-8B	**−16.82**	**−8.62**	[−10.33, −6.88]	**−4.47**	**−6.96**
Gemma-3-27B	**−15.53**	**−8.11**	[−9.83, −6.38]	+0.31	**−5.84**
Qwen2.5-32B	**−15.12**	**−7.90**	[−9.00, −6.75]	**−2.52**	**−2.35**
OLMo-2-13B	**−12.56**	**−7.85**	[−9.23, −6.57]	**−3.86**	**−5.10**
Qwen2.5-7B	**−9.45**	**−7.34**	[−8.50, −6.16]	**−1.94**	**−2.98**
OLMo-2-32B	**−10.65**	**−7.18**	[−9.02, −5.42]	**−1.91**	**−5.83**
Qwen2.5-72B	**−13.41**	**−6.66**	[−8.17, −5.29]	**−1.95**	**−2.42**
Gemma-3-4B	**−9.34**	**−6.43**	[−8.05, −4.69]	−1.02	**−3.67**
Qwen2.5-14B	**−13.00**	**−6.24**	[−7.48, −5.14]	−1.29	**−2.97**
Qwen2.5-3B	**−8.05**	**−3.32**	[−4.77, −1.80]	−0.78	**−3.84**
Llama-3.1-70B	**−8.69**	**−3.14**	[−5.09, −1.29]	**−2.23**	**−2.61**
OLMo-3-7B	**−7.70**	**−3.03**	[−4.61, −1.31]	**−3.23**	**−5.58**
Llama-3.1-8B	**−6.21**	**−2.79**	[−4.36, −1.31]	**−3.53**	**−4.39**
Qwen2.5-1.5B	**−5.22**	+0.72	[−0.84, +2.39]	**+1.78**	**+1.71**
down / up / unclear	16/0/0	15/0/1		10/1/5	15/1/0

prompt, so each checkpoint has 1,632 generations. The design, the single reading path and the genre rule below were committed before any of these generations existed. The reading is descriptive: there is no bar and no outcome name. **With the genre fixed, both forms still fall**: "you" in 15 of 16 models across the four genres together and "I" in 15 of 16, with 14 and 14 of those above their paired placebo (Table A14). Genre by genre the counts are 15/0/1 for "you" and 16/0/0 for "I" on blog, 12/0/4 for "you" and 13/0/3 for "I" on forum, 12/0/4 for "you" and 15/0/1 for "I" on how-to, 15/0/1 for "you" and 14/0/2 for "I" on news (down / up / unclear); each genre carries a quarter of the data, so an interval covering zero there is a loss of power and not a result. On prose lines alone, across the four genres together, "you" falls in 14 models per token and 13 per sentence, and "I" in 13 and 13, with no model rising in either. **The header fixes the request, not the realised genre.** Each generation is classified by a rule fixed before counting: forum cues, then how-to cues, then news cues, none of which contains a person, with blog as the residual class. A cue class is entered only when its cue fires, so those three shares are lower bounds on the genre a reader would name, and blog's is an upper bound. Asking for a genre moves the class a generation lands in by at most 9 percentage points (aligned how-to: 52% of how-to requests against 44% of the others), and by three points or fewer in the other three, where the cue rule fires rarely. The two checkpoints also still differ from each other: pooled over the four requests the base generations are classified blog in 65% and how-to in 27%, the aligned ones in 49% and 46%. The header narrows the genre difference between the checkpoints without closing it, so what this prompt set adds is that the fall survives when the same genre is requested of both checkpoints, not that genre is held constant.

Where ELICIT-99 reaches, and the one model it does not. Qwen2.5-1.5B's reason is printed by name in Table A4 rather than left blank; the 70B and 72B pairs have since been generated and are scored. The placebo band on this prompt set is far wider than on the debate prompts — 95th percentiles of 5.1 to 9.1 per thousand against 0.9 to 1.2 — and the three models that clear the interval but not the placebo are Qwen2.5-3B, Qwen2.5-32B and Qwen2.5-72B. ELICIT-99's growth is a matter of record: it was read at eleven models, then thirteen, then fifteen, and every pair that closed a gap was generated after the earlier reading and measured with the same tool against the same bar, which has not moved. The reading↔production map is pairwise-complete and its n ranges from six to sixteen; against an n-matched shuffle reference no pair survives, but at $n = 6$ that reference itself admits $|\rho|$ up to 0.829 at the 95th percentile. The human strips are anchors, not paired contrasts: the conversational corpora are interactive and the written ones edited, while model outputs are single-turn, and **the two text populations answer different prompts**. The written strip is also not one register: across the ten question-answering communities we sampled, second-person density spans 2.7 to 37.4 per thousand, so it should be read as a range of expository registers rather than as a single human baseline. In Czech or Spanish the same retreat would have to be counted on agreement morphology instead of on pronouns.

The second person model by model, with the fragile positive named. On the continuation prompt it is not a lineage artifact, and the evidence for that is the ranking rather than the extremum: the four AI2 models occupy ranks 1, 3, 6 and 9–10 of sixteen by proportional drop, interleaved with the others and not clustered. We scope the claim to this prompt because under the chat template the lineage label *does* sort the split, and the sign itself moves (Figure A2). **The lineage sign reading depends on where OLMo-3 is put.** Read with OLMo-3 as its own level, 13 of the 14 models that carry a lineage fall on its majority side (frac = 0.030), five of six lineages being sign-pure and the exception a single Qwen2.5 model; merged into AI2, the five lineages

Table A14: **Raw continuation with the genre fixed by a one-line header.** Aligned minus base, per 1,000 tokens, first person counted directly. *all* pools the four genres. **Bold**: prompt-clustered 95% interval excludes zero. The middle block reads the pooled panel on prose lines only, per token and per sentence. The lower block is the genre a generation was classified into, summed over the 16 models, against the genre its header requested; blog is the residual class, so its diagonal is an upper bound.

	all		blog		forum		how-to		news	
model	ΔM_1	Δ1st	ΔM_1	Δ1st	ΔM_1	Δ1st	ΔM_1	Δ1st	ΔM_1	Δ1st
Qwen2.5-1.5B	**−2.88**	−0.68	**−3.05**	**−4.18**	**−2.09**	+0.83	−3.63	+0.85	−2.71	−0.51
Qwen2.5-3B	**−1.99**	**−2.32**	**−5.02**	**−4.45**	+0.01	+0.26	−0.91	**−2.34**	**−2.26**	−2.89
Mistral-7B-v0.3	**−10.87**	**−7.77**	**−12.19**	**−11.59**	**−9.35**	**−8.82**	**−10.54**	**−5.64**	**−11.45**	**−4.69**
Qwen2.5-7B	**−4.99**	**−5.16**	**−3.26**	**−6.88**	**−2.19**	**−3.16**	**−10.33**	**−7.06**	**−4.20**	**−3.58**
OLMo-3-7B	**−10.10**	**−9.22**	**−10.03**	**−8.19**	**−9.10**	**−8.46**	**−9.79**	**−9.84**	**−11.10**	**−9.99**
Llama-3.1-8B	**−7.04**	**−6.71**	**−7.57**	**−8.54**	**−6.66**	**−5.95**	**−8.07**	**−7.29**	**−5.91**	**−5.12**
Tülu-3-8B	**−8.77**	**−10.39**	**−8.43**	**−13.36**	**−8.47**	**−10.14**	**−10.88**	**−8.15**	**−7.41**	**−9.84**
Gemma-3-4B	**−7.57**	**−7.17**	**−8.44**	**−3.55**	**−6.67**	**−7.26**	**−7.29**	**−10.22**	**−7.89**	**−7.61**
OLMo-2-13B	**−7.57**	**−8.17**	**−8.80**	**−10.39**	**−6.82**	**−6.72**	**−7.33**	**−8.26**	**−7.44**	**−7.49**
Gemma-3-12B	**−8.00**	**−9.21**	**−7.61**	**−10.00**	**−6.80**	**−7.54**	**−9.16**	**−9.60**	**−8.48**	**−9.69**
Qwen2.5-14B	**−2.36**	**−4.63**	**−3.18**	**−4.47**	+1.20	**−2.42**	**−3.10**	**−5.95**	**−4.50**	**−5.87**
Gemma-3-27B	**−9.19**	**−6.88**	**−8.54**	**−6.42**	**−9.57**	**−6.79**	**−10.35**	**−6.68**	**−8.37**	**−7.60**
OLMo-2-32B	**−7.92**	**−9.58**	**−7.44**	**−9.83**	**−9.31**	**−10.67**	**−8.48**	**−10.03**	**−6.27**	**−7.69**
Qwen2.5-32B	−2.54	**−4.60**	−1.68	**−5.92**	−0.76	**−3.87**	−3.27	**−3.86**	**−4.46**	**−4.72**
Qwen2.5-72B	**−2.06**	**−4.40**	**−3.36**	**−6.28**	+1.24	−1.56	−1.69	**−4.76**	**−4.60**	**−5.17**
Llama-3.1-70B	**−4.36**	**−4.10**	**−5.61**	**−4.91**	**−5.64**	**−3.64**	**−2.63**	**−4.48**	**−3.19**	**−3.02**
down / up / unclear	15/0/1	15/0/1	15/0/1	16/0/0	12/0/4	13/0/3	12/0/4	15/0/1	15/0/1	14/0/2
down, above placebo	14	14	11	13	10	10	10	14	13	13

	prose, per token		prose, per sentence	
model	ΔM_1	Δ1st	ΔM_1	Δ1st
Qwen2.5-1.5B	**−2.30**	−1.18	−0.023	+0.018
Qwen2.5-3B	**−2.05**	**−2.54**	**−0.032**	**−0.035**
Mistral-7B-v0.3	**−12.10**	−0.11	**−0.215**	−0.025
Qwen2.5-7B	**−5.21**	**−4.97**	**−0.090**	**−0.081**
OLMo-3-7B	**−9.78**	**−9.41**	**−0.142**	**−0.137**
Llama-3.1-8B	**−7.96**	**−6.50**	**−0.145**	**−0.113**
Tülu-3-8B	**−10.28**	**−11.72**	**−0.191**	**−0.214**
Gemma-3-4B	**−5.66**	−2.76	**−0.116**	**−0.070**
OLMo-2-13B	**−7.70**	**−8.31**	**−0.126**	**−0.133**
Gemma-3-12B	**−7.41**	**−6.91**	**−0.145**	**−0.145**
Qwen2.5-14B	**−2.73**	**−5.36**	**−0.043**	**−0.093**
Gemma-3-27B	**−8.85**	**−3.84**	**−0.162**	**−0.079**
OLMo-2-32B	**−8.55**	**−9.72**	**−0.143**	**−0.161**
Qwen2.5-32B	−1.15	**−4.40**	+0.000	**−0.070**
Qwen2.5-72B	−1.10	**−4.58**	−0.008	**−0.075**
Llama-3.1-70B	**−5.99**	**−3.59**	**−0.108**	−0.057
down / up / unclear	14/0/2	13/0/3	13/0/3	13/0/3

	base outputs, classified as				aligned outputs, classified as			
header asks for	blog	forum	how-to	news	blog	forum	how-to	news
blog	4,444	40	1,607	437	3,309	6	2,922	291
forum	4,069	43	1,936	480	3,286	10	2,937	295
how-to	4,176	38	1,906	408	2,780	8	3,409	331
news	4,294	26	1,646	562	3,352	5	2,688	483
matches the request	blog 68%, forum 1%, how-to 29%, news 9%				blog 51%, forum 0%, how-to 52%, news 7%			

§1 counts, it is 12 of 14 (frac $= 0.092$) and the sign no longer clears the bar. The magnitude result holds under both labellings. **Lineage and pipeline are confounded here**, so neither labelling names a mechanism. Prompting for status acts scales the measurement and holds its sign in fourteen of the fifteen models scored on ELICIT-99, the exception being OLMo-3-7B; base second-person density is higher on ELICIT-99 than on the debate prompts in all fifteen, by a median of 2.27×. The per-model detail, the length-matched first window and the prompt-format caveat that keeps that prompt set from carrying the contrast are in Appendix D and Table A4. The dilution decomposition names its own denominator: aligned sentences are longer in 14 of 16 models, and the length term carries about a fifth of the median model's fall, most or all of it in four Qwen2.5 models (Appendix E).

Whether the aligned outputs leave the continuation on the raw-continuation prompt. A raw continuation prompt can still be answered as an assistant, and we counted how often it is. A rule written

before the count reads only the opening of each generation — its first 160 characters, after leading whitespace, quotation marks and markdown are stripped — and assigns the first class that matches: a refusal (*I'm sorry*, *I cannot*, *As an AI*), an assistant opening (*Sure*, *Certainly*, *Here is*, *Let me*) or a meta-comment on the text (*This argument*, *It seems*, *In summary*, *You've provided*); everything else is read as a continuation, and the same rule is applied to the base outputs as its structural reference. Of 104,448 aligned generations, 2,815 (2.7%) are flagged — 136 refusals, 291 assistant openings and 2,388 meta-comments — against 2,269 (2.2%) in the base outputs; refusals and assistant openings alone are 0.4% (427 of 104,448), and the meta-comments are the rest. Per model the aligned share runs from 0.9% to 5.6% and the base share from 0.8% to 3.5%. **With the flagged generations removed from both checkpoints, ΔM_1 stays negative and interval-disjoint in 16 of 16 models** (Table A15); the largest change is 0.18 per thousand (Gemma-3-27B, $-3.07 \rightarrow -3.24$), and the full-panel ΔM_1 recomputed by the same counter reproduces the registered single-seed reading in all sixteen. Between 0.3% and 4.3% of an aligned checkpoint's second persons sit in its flagged generations. **The reading has three limits.** The rule reads only the opening, so a generation that changes mode mid-text is not seen; a deliberately over-cutting reading, its rule committed before its count, that flags a refusal or assistant marker at the start of any line or sentence and any turn marker or markdown structure anywhere in the generation removes 51.9% of aligned and 31.7% of base generations and leaves 16 of 16 negative and interval-disjoint, the fall moving by at most 2.44 per thousand (Qwen2.5-14B, $-3.49 \rightarrow -5.93$). The 31.7% it removes from the base outputs, which have no assistant mode to leave, is that rule's false-positive floor; the fall widens in all but two models and narrows slightly in Llama-3.1-70B and Mistral-7B-v0.3. Its class precision was not measured. And on a seeded sample of flagged generations from three models, the meta-comment class also collects legitimate debate continuations (*The argument is based on…*) and the refusal class a continuation such as *I can't prove it*, so among openings the removed set errs on the large side, and 16 of 16 holds with that larger set removed. The flagged share is therefore neither a lower nor an upper bound on how often the aligned outputs leave the continuation. The reading is descriptive: the rule was written before the count, but no acceptance bound was (Appendix B).

Table A15: **The raw-continuation-prompt contrast with the generations that leave the continuation removed.** Flagged share is the percentage of a checkpoint's 6,528 generations whose opening a rule written before the count reads as a refusal, an assistant opening or a meta-comment. ΔM_1 is recomputed on both checkpoints without the flagged generations, with its prompt-clustered 95% CI; the last column is the share of the aligned outputs' second persons that sit in its flagged generations. Rows follow Table 1.

	flagged share		ΔM_1		aligned 2nd persons
model	aligned	base	all	without flagged [95% CI]	in flagged
Mistral-7B-v0.3	0.98%	1.06%	-9.86	**-9.93** [-11.57, -8.40]	1.3%
Tülu-3-8B	1.33%	2.31%	-8.71	**-8.73** [-10.10, -7.48]	0.3%
OLMo-2-13B	3.74%	3.48%	-6.28	**-6.34** [-7.46, -5.44]	1.7%
OLMo-2-32B	2.85%	2.65%	-4.77	**-4.77** [-5.49, -4.07]	1.7%
Llama-3.1-8B	1.65%	2.31%	-4.77	**-4.77** [-5.78, -3.81]	1.0%
OLMo-3-7B	0.89%	0.75%	-4.70	**-4.73** [-5.75, -3.77]	0.8%
Qwen2.5-7B	2.60%	1.04%	-3.90	**-3.82** [-5.49, -2.28]	0.9%
Llama-3.1-70B	1.52%	2.28%	-3.78	**-3.74** [-4.41, -3.09]	1.0%
Qwen2.5-72B	3.75%	2.37%	-3.69	**-3.66** [-6.03, -1.76]	2.8%
Gemma-3-4B	3.65%	2.39%	-3.63	**-3.65** [-5.31, -2.08]	2.9%
Qwen2.5-14B	3.72%	2.77%	-3.49	**-3.50** [-5.27, -2.04]	2.9%
Qwen2.5-3B	1.61%	1.75%	-3.46	**-3.48** [-5.17, -1.89]	0.9%
Gemma-3-27B	2.99%	2.44%	-3.07	**-3.24** [-5.52, -1.25]	3.9%
Gemma-3-12B	5.59%	2.33%	-2.80	**-2.80** [-4.51, -1.36]	4.3%
Qwen2.5-32B	4.32%	1.81%	-2.18	**-2.16** [-3.44, -1.02]	3.5%
Qwen2.5-1.5B	1.93%	3.02%	-1.61	**-1.68** [-3.15, -0.41]	1.6%

Prose lines alone, for both persons (§3). A rule committed before each count drops every line that opens as a heading, a bulleted or numbered item, or consists of bold text alone, and reads what remains with the canonical counters; the first person is counted directly, with a capitalised *US* not counted. Bold marks a prompt-clustered 95% interval that excludes zero. The drop is line-level, so a sentence that runs from a prose line into a list line is not seen by this rule. The sentence column classifies whole sentences instead, each generation parsed once: a sentence touching any list line is dropped whole. By sentence the fall holds in 13 of 16 models; it holds by neither unit in Gemma-3-12B, Gemma-3-27B and Qwen2.5-1.5B.

model	lines dropped, base	aligned	ΔM_1, all lines	ΔM_1, prose lines	prose sentences	Δ1st, prose
Gemma-3-12B	0.128	0.445	−2.80	+0.69	+0.36	**−14.02**
Gemma-3-27B	0.137	0.455	−3.07	−0.35	−0.59	**−12.53**
Gemma-3-4B	0.160	0.432	−3.63	−1.16	**−1.36**	**−6.11**
Llama-3.1-70B	0.059	0.085	−3.78	**−3.92**	**−3.92**	**−9.27**
Llama-3.1-8B	0.051	0.075	−4.77	**−4.92**	**−4.92**	**−6.40**
Mistral-7B-v0.3	0.307	0.635	−9.86	**−8.68**	**−8.62**	**−13.91**
OLMo-3-7B	0.055	0.078	−4.70	**−4.71**	**−4.70**	**−7.87**
OLMo-2-13B	0.090	0.120	−6.28	**−6.27**	**−6.26**	**−12.55**
OLMo-2-32B	0.072	0.061	−4.77	**−4.77**	**−4.79**	**−10.66**
Qwen2.5-1.5B	0.113	0.240	−1.61	−1.13	−1.15	**−4.61**
Qwen2.5-14B	0.127	0.480	−3.49	**−1.67**	**−2.26**	**−12.57**
Qwen2.5-32B	0.103	0.455	−2.18	−0.34	**−0.99**	**−14.07**
Qwen2.5-3B	0.099	0.388	−3.46	**−2.69**	**−3.07**	**−7.09**
Qwen2.5-72B	0.123	0.549	−3.69	−1.28	**−2.05**	**−11.66**
Qwen2.5-7B	0.121	0.438	−3.90	**−2.80**	**−3.10**	**−8.12**
Tülu-3-8B	0.051	0.100	−8.71	**−9.04**	**−9.03**	**−17.48**
negative, interval excluding zero			16	10	13	16

Line types, recombined at the counter's printed digit. The three line classes recombine to each checkpoint's own M_1 within 0.005, half the last printed digit, in 11 of 16 models; the rest are marked. Among those 11, on the raw-continuation prompt ΔM_1 is negative and interval-disjoint on prose lines in 8 and on list lines in 10, and under the template it is positive and disjoint on prose lines in 5 and on list lines in 1. **Assignment is by line**: a sentence that runs from a prose line into a bullet is not seen. Rows are sorted by the raw-continuation contrast; bold marks a prompt-clustered 95% interval that excludes zero, and the recombination tolerance was set after an earlier, stricter tolerance had failed (Appendix B).

	raw-continuation			chat-template		
model	prose	list	heading	prose	list	heading
Mistral-7B-v0.3	−8.7	−7.6	+14.4	+10.1	−5.4	−12.1
Tülu-3-8B	−9.0	−3.9	+0.0	−17.0	−7.3	+0.0
OLMo-2-13B	−6.3	−6.3	+1.8	−12.1	−5.5	+0.3
OLMo-2-32B	−4.8	−4.8	+5.3	−10.6	−5.8	+0.9
Llama-3.1-8B	*tool record; not reported*					
OLMo-3-7B	*tool record; not reported*					
Qwen2.5-7B	−2.8	−3.4	−3.8	−5.6	−1.2	−0.4
Llama-3.1-70B	−3.9	−2.5	−1.5	+23.1	+5.6	−2.5
Qwen2.5-72B	−1.3	−3.8	+1.0	+10.1	−2.6	−0.1
Gemma-3-4B	*tool record; not reported*					
Qwen2.5-14B	−1.7	−3.9	−1.1	+0.6	−1.7	−3.6
Qwen2.5-3B	−2.7	−5.3	+1.3	+1.6	−4.3	+0.1
Gemma-3-27B	*tool record; not reported*					
Gemma-3-12B	*tool record; not reported*					
Qwen2.5-32B	−0.3	−1.9	−4.6	−1.0	−3.1	−7.4
Qwen2.5-1.5B	−1.1	−2.7	+1.6	+3.2	−4.3	+2.0

Where the controlled readings fail: inside the placebo band, or a wider interval? For each model whose fall does not separate on prose lines or per sentence, we compare the fall with the 95th percentile of its paired within-prompt placebo, built on the same pairs and the same readings. Of the six models that do not separate on prose lines, Gemma-3-12B, Gemma-3-27B, Gemma-3-4B and Qwen2.5-32B sit inside the band; Qwen2.5-1.5B and Qwen2.5-72B lie beyond it, with an interval that covers zero. Of the four models that do not separate per sentence, Qwen2.5-1.5B, Qwen2.5-14B and Qwen2.5-32B sit inside the band; Qwen2.5-72B lies beyond it, with an interval that covers zero. Of the three models that do not separate on prose selected by sentence, Gemma-3-12B and Gemma-3-27B sit inside the band; Qwen2.5-1.5B lies beyond it, with an interval excluding zero in this paired reading. Inside the band means the reading cannot tell the fall from a split of the model's own outputs; beyond it with an interval covering zero means the interval, not the effect, is what the reading loses.

The main contrast per 1,000 words. Table 1 reads the second-person rate per 1,000 tokens, a denominator that counts punctuation, list markers and markdown. Here the same outputs, the same pooled seeds, the same prompt-clustered intervals and the same paired placebo are read per 1,000 words, where a word is a token carrying at least one letter, so that punctuation, list markers, markdown and numerals leave the denominator; only the denominator changes, and the numerator is unchanged. Read with tokens, the code reproduces Table 1 exactly (0 differences over sixteen models). With words the fall holds in 16 of 16 models and clears the placebo in 16; a rule fixed before the count sent the result to the main text at 12 or more and to the Limitations below 12. The base outputs carry 0.87 to 0.88 words per token and the aligned outputs 0.80 to 0.88. Bold marks an interval excluding zero.

	per 1,000 tokens		per 1,000 words			words per token
model	Δ	$\pm$	Δ	$\pm$	placebo p95	base / aligned
Mistral-7B-v0.3	$\mathbf{-9.51}$	1.39	$\mathbf{-10.93}$	1.59	0.72	0.87 / 0.86
Tülu-3-8B	$\mathbf{-9.04}$	1.31	$\mathbf{-10.25}$	1.47	0.58	0.88 / 0.88
OLMo-2-13B	$\mathbf{-6.76}$	0.97	$\mathbf{-7.69}$	1.10	0.65	0.87 / 0.87
OLMo-2-32B	$\mathbf{-5.17}$	0.66	$\mathbf{-5.95}$	0.75	0.59	0.87 / 0.87
Llama-3.1-8B	$\mathbf{-4.69}$	1.10	$\mathbf{-5.32}$	1.24	0.58	0.88 / 0.88
OLMo-3-7B	$\mathbf{-4.54}$	0.85	$\mathbf{-5.16}$	0.96	0.59	0.87 / 0.87
Qwen2.5-72B	$\mathbf{-4.14}$	2.37	$\mathbf{-4.16}$	2.66	0.56	0.87 / 0.81
Llama-3.1-70B	$\mathbf{-3.86}$	0.48	$\mathbf{-4.51}$	0.55	0.60	0.87 / 0.88
Qwen2.5-7B	$\mathbf{-3.62}$	1.36	$\mathbf{-3.79}$	1.50	0.56	0.88 / 0.84
Qwen2.5-14B	$\mathbf{-3.44}$	1.49	$\mathbf{-3.51}$	1.66	0.63	0.87 / 0.83
Qwen2.5-3B	$\mathbf{-3.06}$	1.64	$\mathbf{-3.23}$	1.82	0.54	0.88 / 0.85
Gemma-3-4B	$\mathbf{-3.02}$	1.35	$\mathbf{-2.77}$	1.49	0.62	0.87 / 0.81
Qwen2.5-32B	$\mathbf{-2.77}$	1.50	$\mathbf{-2.70}$	1.67	0.55	0.87 / 0.83
Gemma-3-27B	$\mathbf{-2.59}$	1.67	$\mathbf{-2.16}$	1.89	0.66	0.87 / 0.80
Gemma-3-12B	$\mathbf{-2.33}$	1.44	$\mathbf{-1.85}$	1.61	0.59	0.87 / 0.81
Qwen2.5-1.5B	$\mathbf{-1.77}$	1.20	$\mathbf{-1.82}$	1.36	0.46	0.88 / 0.86

The other readings under both denominators, each on its own set as in the main text: "you" under the chat template (filtered set, one seed): 7 down, 7 up, 2 covering zero per 1,000 tokens and 7 down, 7 up, 2 covering zero per 1,000 words; "you" on ordinary instructions (in-context assistant prompt for the base, own template for the aligned model): 16 down, 0 up, 0 covering zero, 16 falling above the placebo per 1,000 tokens and 16 down, 0 up, 0 covering zero, 16 falling above the placebo per 1,000 words; "I" counted directly on raw continuation (original seed): 16 down, 0 up, 0 covering zero per 1,000 tokens and 16 down, 0 up, 0 covering zero per 1,000 words; derived first person under the chat template: 16 down, 0 up, 0 covering zero per 1,000 tokens and 16 down, 0 up, 0 covering zero per 1,000 words; "I" counted directly on ordinary instructions: 12 down, 2 up, 2 covering zero, 12 falling above the placebo per 1,000 tokens and 12 down, 2 up, 2 covering zero, 11 falling above the placebo per 1,000 words; "you" on prose lines alone: 10 down, 0 up, 6 covering zero per 1,000 tokens and 10 down, 0 up, 6 covering zero per 1,000 words; "you" on prose selected by sentence, with this paired reading's own intervals, which separate one model more than the 13 of §3: 14 down, 0 up, 2 covering zero, 13 falling above the placebo per 1,000 tokens and 12 down, 0 up, 4 covering zero, 11 falling above the placebo per 1,000 words. Measures per sentence divide by the number of sentences and do not change with the denominator; the code checks that they are identical under both. The rule and the equivalence gate were committed before any count per word was read; the re-read is descriptive.

The panel at one sampling seed. Table 1 is read on four sampling seeds pooled, with the first person counted directly. The registered reading is the single seed the panel was generated at, and it is given here unchanged, with the first person derived as it was registered; the count is the same under both, 16 of 16 negative with the interval excluding zero and above the paired placebo.

model	base	inst.	ΔM_1	$\pm$	Δ1st	model	base	inst.	ΔM_1	$\pm$	Δ1st
Mistral-7B-v0.3	13.6	3.8	-9.86	±1.71	-15.2	Qwen2.5-7B	10.4	6.5	-3.90	±1.70	-9.6
Tülu-3-8B	11.6	2.9	-8.71	±1.25	-17.0	Qwen2.5-14B	10.7	7.2	-3.49	±1.62	-13.3
OLMo-2-13B	11.9	5.6	-6.28	±1.03	-12.6	Qwen2.5-3B	10.5	7.0	-3.46	±1.70	-8.2
OLMo-2-32B	13.7	9.0	-4.77	±0.77	-10.7	Gemma-3-4B	11.5	7.9	-3.63	±1.74	-9.8
Llama-3.1-8B	11.6	6.8	-4.77	±0.99	-6.2	Qwen2.5-32B	10.2	8.0	-2.18	±1.21	-15.4
OLMo-3-7B	14.2	9.5	-4.70	±1.02	-8.0	Gemma-3-27B	12.0	8.9	-3.07	±2.23	-15.7
Qwen2.5-72B	10.3	6.6	-3.69	±2.10	-13.7	Gemma-3-12B	12.6	9.8	-2.80	±1.62	-16.4
Llama-3.1-70B	11.5	7.7	-3.78	±0.63	-8.7	Qwen2.5-1.5B	10.1	8.5	-1.61	±1.31	-5.5

Intervals clustered on the thesis. The 34 prompts are 17 theses, each in two stances, so prompt-clustered intervals treat the two stances of a thesis as independent. We resampled theses instead (1000 draws on four pooled sampling seeds; with prompts as the unit the loop reproduces the second-person intervals of Table 1 exactly; the first-person column here is the derived measure, which that table replaces with the direct count on the same seeds). Resampled by thesis, ΔM_1 stays negative with an interval excluding zero in 16 of 16 models (16 of 16 by prompt) and Δ1st in 16 of 16 (16 by prompt); the intervals widen by a median factor of 1.39 and 1.28. With 17 clusters a percentile interval is coarse, and the thesis absorbs the dependence between two stances, not the scaffold and prefix conditions that every prompt shares.

	ΔM_1 (four seeds pooled)			Δ1st, derived (four seeds pooled)		
model	shift	by prompt	by thesis	shift	by prompt	by thesis
Mistral-7B-v0.3	-9.51	$[-11.00, -8.22]$	$[-11.85, -7.78]$	-15.16	$[-16.08, -14.13]$	$[-16.34, -13.85]$
Tülu-3-8B	-9.04	$[-10.36, -7.73]$	$[-11.04, -7.34]$	-17.28	$[-18.64, -15.94]$	$[-18.88, -15.67]$
OLMo-2-13B	-6.76	$[-7.80, -5.86]$	$[-8.35, -5.62]$	-12.69	$[-13.41, -11.91]$	$[-13.56, -11.84]$
OLMo-2-32B	-5.17	$[-5.89, -4.57]$	$[-6.23, -4.36]$	-10.14	$[-10.68, -9.56]$	$[-10.82, -9.40]$
Llama-3.1-8B	-4.69	$[-5.83, -3.62]$	$[-6.32, -3.34]$	-6.36	$[-7.25, -5.45]$	$[-7.51, -5.16]$
OLMo-3-7B	-4.54	$[-5.41, -3.70]$	$[-5.62, -3.45]$	-8.46	$[-9.45, -7.41]$	$[-9.72, -7.20]$
Qwen2.5-72B	-4.14	$[-6.71, -1.96]$	$[-8.12, -1.51]$	-14.28	$[-15.45, -13.17]$	$[-15.78, -12.75]$
Llama-3.1-70B	-3.86	$[-4.34, -3.38]$	$[-4.45, -3.24]$	-8.84	$[-9.90, -7.67]$	$[-10.28, -7.17]$
Qwen2.5-7B	-3.62	$[-5.03, -2.31]$	$[-5.77, -1.91]$	-9.64	$[-10.65, -8.64]$	$[-10.92, -8.46]$
Qwen2.5-14B	-3.44	$[-5.02, -2.05]$	$[-5.84, -1.76]$	-13.79	$[-14.86, -12.85]$	$[-15.18, -12.55]$
Qwen2.5-3B	-3.06	$[-4.76, -1.47]$	$[-5.64, -1.11]$	-8.56	$[-9.49, -7.69]$	$[-9.88, -7.48]$
Gemma-3-4B	-3.02	$[-4.47, -1.77]$	$[-4.93, -1.26]$	-10.10	$[-10.89, -9.28]$	$[-11.04, -9.19]$
Qwen2.5-32B	-2.77	$[-4.37, -1.38]$	$[-5.28, -0.97]$	-15.71	$[-16.83, -14.69]$	$[-17.09, -14.36]$
Gemma-3-27B	-2.59	$[-4.33, -1.00]$	$[-5.14, -0.56]$	-16.29	$[-17.38, -15.08]$	$[-17.59, -14.82]$
Gemma-3-12B	-2.33	$[-3.82, -0.93]$	$[-4.42, -0.62]$	-16.64	$[-17.75, -15.32]$	$[-18.07, -14.89]$
Qwen2.5-1.5B	-1.77	$[-3.07, -0.66]$	$[-3.82, -0.34]$	-5.54	$[-6.47, -4.74]$	$[-6.78, -4.47]$

Which person falls further, in absolute density and in rate. In absolute density the first person falls further than the second in all sixteen models (Table 1). That ordering is a statement about absolute density and not about rate: in proportional terms the first person falls further in only 10 of 16, so we report the absolute difference and make no claim that one person recedes before the other. The ratio of second to first person has no consistent direction on the debate prompts (six fall, ten rise). On ELICIT-99 the proportional ordering does hold in all fifteen scored models, which we give as description rather than as replication: that prompt set was built to afford status acts, and its base densities are higher by a median of $2.27\times$ (1.2–$3.4\times$ across the fifteen; Appendix E).

How far a model withdraws is not shared, and not a consequence of scale. On the continuation prompt base models cluster on second-person density, and aligned models spread. Alignment does not push every model to a common floor. It separates them. The random-effects synthesis of Appendix A is this observation in statistical form: the models share the sign, in 16 of 16, and not the size, at $I^2 \approx 89\%$, and the paper claims only the sign.

Written expository prose, and where the aligned band sits against it. Three written expository corpora, namely StackExchange answers, wikiHow instructions and OpenStax textbook prose, widen the human range from 10.3–24.1 per thousand in conversation to 4.0–50.8 in writing. The aligned range (2.9–9.8) falls below the two highest conversational groups but overlaps the advocate row (6.6), and overlaps the written band only at its impersonal end and reaches below it (2.9 against the written floor of 4.0): nearest is the textbook at 4.0, and on impersonal correction share the aligned band (0.91–0.98) brackets the textbook's 0.93 while every other reference sits below 0.84. The match is partial, because aligned force, 0.007–0.030 acts per sentence, is below the textbook's 0.073. In Biber's (1988) terms, the aligned assistant sits at the informational pole of Dimension 1, where textbooks sit, and base models and human conversation sit toward the involved pole.

Bare imperatives: the rate holds, the procedural share moves. We report the procedural/prose split because an undifferentiated imperative count conflates a procedural step with an order. The rate moves by at most 0.94 percentage points in any model, while the procedural share moves by much more, so what alignment changes here is composition and not volume.

ELICIT-99: coverage, token caps, and the model it does not reach. On ELICIT-99 the withdrawal is larger than on the debate prompts in twelve models and smaller in three: Qwen2.5-7B ($-3.90 \to -1.63$), Qwen2.5-14B ($-3.49 \to -2.61$) and OLMo-3-7B, whose elicited shift is $+0.54$ with an interval covering zero. Twelve of the fifteen separate on the full text, and twelve on the pre-registered length-matched first window, which now covers the same fifteen. They are not quite the same twelve: OLMo-3-7B and Qwen2.5-7B separate only in the window, Qwen2.5-3B and Qwen2.5-72B only on the full text — and OLMo-3-7B is the sharper case, since its full-text shift is $+0.54$ with an interval covering zero while its windowed shift is -4.51 $[-6.98, -1.88]$, which is what a length-matched reading is for. ELICIT-99 holds the sign everywhere except in that same model. That prompt set reads each checkpoint through its own template, however, so for ten of its fifteen models a change of prompt format is not separated from alignment (Table A4); the debate prompts, where the scaffold is identical on both checkpoints, are what carry the contrast. On an identical string, with the aligned model's own chat template around each prompt given to both checkpoints (4 samples of each of the 99 prompts, up to 1,024 new tokens; the strings match in all 16 models), "you" falls in 12 models, rises in 1 (Gemma-3-4B) and does not separate in 3, and 9 of those 12 clear the paired placebo; "I" falls in 15 and does not separate in 1, 11 clearing the placebo. On prose lines alone, with list, numbered and heading lines removed,

the second person splits (5 fall, 5 rise, 6 do not separate) while the first person falls in 10 and rises in none. This reading is descriptive, and its path was committed before the generations existed.

Force moves less, and not one way. **Where the two layers separate is the stage.** Read stage by stage with prompt-clustered intervals, the force total moves on only 3 of the 9 stages of the three checkpoint sequences we opened first, the only ones with force intervals, and not at the stage the persons move at: it clears zero at supervised fine-tuning on both OLMo-2 checkpoint sequences among them (-0.0031 and -0.0050), where the persons barely move, and at the preference stage on Tülu-3-8B (-0.0061), the one checkpoint sequence of the three where the two layers share a stage (Table 2). Across the wider base→instruct contrast force clears its null in 10 of 16 models, with 11 of its sixteen shifts falling and 5 rising on the pre-registered ruler — a split a two-sided sign test cannot separate from chance ($p = 0.21$); the two layers are plotted side by side in Appendix Figure A4. We report no magnitude ratio between the layers, because the ratio does not survive the ruler swap. What does survive is a count: on either *matched* ruler register separates at least as often as force and only downward. Per thousand tokens register separates in 16 of 16 models and force in 14 of 16, all downward; per sentence register separates in 12 of 16, all downward, and force in 10 of 16, of which two point up (Appendix E). Model by model the two do move together — the rank correlation between the register shift and the force shift is $\rho = +0.67$ across the sixteen ($+0.82$ read per sentence), against a label-permutation null of 0.00 ± 0.26 — so the layers are coupled across models even though they are not carried by the same stage. What does not happen is that corrections disappear, and what happens to them is smaller than we first read it. Base and aligned text alike correct impersonally: the impersonal share of corrections is 0.90–0.96 in the base checkpoints and 0.91–0.98 in the aligned ones, and both bands sit far above every conversational human group we measured, whose shares run 0.53–0.83 (Figure A3). Alignment moves that share only a little. On the debate prompts it rises in 8 of 16 models (9 intervals exclude zero, one of them downward); on ELICIT-99 3 of 15 separate, two upward and one a reversal (OLMo-3-7B, -0.20), so the direction is not one-way on that prompt set either (Appendix Table A22). **Impersonal correction is not something alignment introduces; it is a property of the substrate that alignment nudges.** What alignment withdraws is not the act but its grammatical subject and object. Whether that substitution is what a reader registers as less dominant we tested directly and did not establish: in a pre-registered blind comparison the personal form was read as the more dominant one, but by less than the design's minimum detectable effect, because the third-person control was itself read as dominant far above chance. The substitution is measured; its uptake by a reader is not (Appendix E).

The register it lands in (§3). **The nearest human comparison is human volunteers writing assistant replies** (OASST1, 14.7 [14.0, 15.5] per thousand), above every aligned checkpoint and above the base band (Appendix Figure A3).

Where the aligned assistant lands: the rest of the reading. "Below the humans" is the wrong description once written registers are measured with the same counters: three written expository corpora — StackExchange answers, wikiHow instructions and OpenStax textbook prose — widen the human range from 10.3–24.1 per thousand in conversation to 4.0–50.8 in writing, and the nearest human comparison is neither of those groups but a third, the one the aligned band is furthest from (§3). Unfiltered, six aligned checkpoints read above the OASST1 interval, one inside and nine below, so the stronger claim that none lands inside holds only under the degeneracy filter and we do not rest on it. **Read through the templates it fails by splitting rather than by shifting**: seven aligned checkpoints read above its interval and nine below — unfiltered, six above, one inside and nine below — and the upper seven are exactly the models whose contrast reverses under the template. On impersonal correction the base band sits above the human rows too (0.90–0.96 against 0.53–0.83), and the aligned band lands on the textbook row on both counters that measure person — but it does *not* sit below every conversational group: twelve of the sixteen aligned models sit above the lowest one. Against the written references the aligned range (2.9–9.8 per thousand) reaches below the written floor, with the textbook the nearest reference on both counters that measure person. The OASST1 row is a reservoir sample of 6,528 messages, drawn uniformly from the 18,828 English assistant messages that clear the length floor, over 3,028 conversation trees, read with the same counters and with the 40-word floor applied to the human rows only; on the continuation prompt its 14.7 [14.0, 15.5] sits above every aligned checkpoint and above the base band (10.1–14.2). The one model that lands inside the interval unfiltered is Qwen2.5-1.5B at 14.6. The seven models above it have an impersonal-correction share of 0.84, *below* the aligned band's 0.91–0.98: where they do correct, the volunteers name someone more often than the aligned models do. **The aligned range is itself protocol-bound**: read through each model's own chat template the aligned band widens to 2.0–41.2 per thousand, its upper end Gemma-3-27B under that protocol alone (Table A5), which crosses the conversational human range rather than sitting below it.

Repairing the courtroom rows, and the middle panel's floor. The rows of Figure A3 are, in order of second-person density: Supreme Court justices, Reddit debate and Supreme Court advocates (spoken and interactive); wikiHow how-to, StackExchange answers and OpenStax textbook prose (written expository); and OASST1 volunteers writing assistant replies. The shaded row is the OpenStax textbook. Impersonal correction separates models from human speakers on both checkpoints and alignment moves it only a little. Once the

courtroom corpus is sampled uniformly across its 64 years rather than from its tail, advocates read 6.6 [6.4, 6.9] per thousand. The justice/advocate ratio is 3.5× on the uniformly sampled corpus; a sampler concentrated in recent terms read advocates at 10.3 instead, and the repair is in Appendix A. **Sensitivity of the middle panel**: the correction counter's precision was measured against a human labelling (68%, 26/38) and it under-detects impersonality, so every value in that panel—human rows and model bands alike—is a lower bound, and the rows lowest on it are the ones a better counter could move most (Appendix C).

The five stimulus classes of ELICIT-99. Each class of ELICIT-99 was read on its own with ELICIT-99's pairing, prompt clustering and paired placebo (19–20 prompts per class, so intervals are wide and a null means not resolved at that size). ΔM_1 excludes zero downward in 11 of 15 models on false premises, 9 on bad plans, 10 on requests to ratify reasoning, 5 on borderline requests and 15 on contested opinions; Δ1st does so in 15, 15, 15, 7 and 15 in the same order. The borderline request class is the weakest on both counters.

ΔM_1	false premise	bad plan	reasoning to ratify	borderline request	contested opinion
Gemma-3-12B	−9.52	−15.95	−4.91	−4.49	−11.45
Gemma-3-27B	−8.84	−14.46	−7.37	−4.69	−4.61
Gemma-3-4B	−8.54	−20.93	−11.08	−2.39	−13.13
Llama-3.1-70B	−18.68	−13.01	−19.97	−13.33	−24.60
Llama-3.1-8B	−19.48	−20.41	−13.83	−25.94	−31.90
Mistral-7B-v0.3	−15.86	−6.30	−13.06	−20.81	−25.72
OLMo-3-7B	−2.45	−4.17	+8.96	+13.20	−4.16
OLMo-2-13B	−11.05	−20.07	−16.93	−17.10	−23.38
OLMo-2-32B	−13.33	−10.69	−13.53	−10.14	−14.82
Qwen2.5-14B	−1.66	−4.13	−2.56	+4.54	−3.73
Qwen2.5-32B	−3.90	−0.45	−7.71	−6.87	−8.20
Qwen2.5-3B	−7.68	−2.89	−11.38	+3.36	−13.68
Qwen2.5-72B	−2.53	−7.52	−5.90	+1.23	−4.36
Qwen2.5-7B	+0.24	−0.31	−0.74	+3.73	−7.14
Tülu-3-8B	−18.82	−21.79	−15.34	−26.05	−31.66
down / up / null, ΔM_1	11/0/4	9/0/6	10/1/4	5/1/9	15/0/0
down / up / null, Δ1st	15/0/0	15/0/0	15/0/0	7/2/6	15/0/0

Bold: the prompt-clustered 95% interval excludes zero. Qwen2.5-1.5B is on the degeneracy shelf of ELICIT-99 and not read.

Past the first 160 tokens. On the debate prompts most aligned outputs reach the 160-token cap, so their rates describe openings. The instruction prompts of Appendix D.4 were generated in a 2,048-token window, and there we read the first 160 tokens of each output and the rest as separate texts (aligned checkpoint in its own template, base checkpoint in the in-context assistant prompt, the B1 cell of Table A16). In the first 160 tokens "you" falls in 16 of 16 models, all 16 clearing the paired placebo, and "I" in 12 (2 rise). After them, Mistral-7B-v0.3 and the three Gemma-3 models write too little to read (0 to 22 pairs). Of the other 12, "you" falls with an interval excluding zero in 6 and rises in none, at a median shift of −8.85 per 1,000 against −8.89 in the first 160 tokens, but on a median of 230 pairs rather than 1,195, and only one model clears its placebo. "I" falls in 1 of 12 after the first 160 tokens, at a median of −0.35 against −2.53. The fall in "you" keeps its size where later text exists, then, but the later text is too sparse to separate it model by model, and the first-person fall is concentrated in the openings on these prompts. The reading is descriptive; its path was committed before the count.

D.3 The chat template

Figure A2 shows the base-to-aligned shift behind Figure A1(b).

The one rising model that names its corpus. The seventh model names its preference corpus, so the audit the other six block can be run once. **It is exploratory**: it was built after the split was seen, on one model and one corpus, and it settles nothing. Read whole — 149,994 of Dolci-Think-DPO-7B's 150,000 pairs, the six dropped because one side exceeds the parser's length limit — **the two persons go opposite ways**: the chosen side carries slightly *more* second person (+0.17 per thousand, 95% CI [0.15, 0.19]) and much less first person (−2.68, [−2.72, −2.63]). **A 20,000-pair sample read first agrees**: both signs and both readings are the same there (+0.16 and −2.68), and the full read narrows the intervals about threefold without moving either estimate. **We do not read either gap as a preference signal**: every pair is labelled by a delta-learning heuristic, the chosen side always a 32B model's answer and the rejected always a 0.6B model's, so the person difference is also a difference between two model sizes, and nothing here separates them. **The test this section wants — whether the data behind a rising disclosed model rewards the persons — therefore does not close**, and it fails to close structurally rather than statistically: the one corpus we can open is not built in a way that could carry it. **What it does show is smaller and stranger**: there is nobody in this corpus. No designer named the axis, no rater scored a response, no human ranked a pair; the pronoun gap is a by-product of which

two models were paired as strong and weak. A preference corpus can carry an unnamed register pronoun gap *with no evaluator in it at all* — one step further from any stated objective than an unaudited judge. Direction again outruns magnitude: the second-person gap sits an order of magnitude below the 1.6 to 9.9 withdrawal on the same ruler.

The chat-template base outputs open a turn of their own. The Tülu-3 and OLMo-2 templates mark the turn boundary in plain text (<|user|>), and a base checkpoint that never saw the template can emit that string inside its own output, opening a second speaker whose words are then counted as the model's. Searching each model's base outputs for the boundary strings *that model's own template* uses, three models do it at scale: OLMo-2-13B on 5,212 of 6,528 rows (79.8%), Tülu-3-8B on 4,429 of 6,528 rows (67.8%), OLMo-2-32B on 3,551 of 6,528 rows (54.4%), and in those three 61%, 55%, 47% of the base outputs' second persons fall after the first marker. Cutting each row at its first marker lowers the base level and shrinks the contrast: OLMo-2-13B 17.72 → 16.70 per thousand and ΔM_1 $-13.57 \rightarrow -12.56$; Tülu-3-8B 19.12 → 17.81 per thousand and ΔM_1 $-17.17 \rightarrow -15.87$; OLMo-2-32B 16.66 → 14.03 per thousand and ΔM_1 $-11.88 \rightarrow -9.24$. **No sign changes and the $7/7/2$ split is unaffected**; what moves is the size of three chat-template base cells. The other thirteen models are at or near zero, so this is a property of the plain-text templates rather than of the prompt set. **One limit is named rather than hidden**: a transition the model marks only with a blank line is invisible to a string search, so every count here is a lower bound.

The chat-template outputs echo the prompt, and the split survives it. Under a chat template a checkpoint can rewrite the user turn instead of answering it, and the second persons counted on such a row are the prompt reflected back rather than the model's own. Flagging a row as an *echo* when any 40-character window of its own prompt's last 200 characters reappears verbatim in the generation, the echo share runs 0.10–0.42 on the base outputs and 0.05–0.46 on the aligned ones — large enough that the degeneracy filter, which looks for repetition rather than for copying, does not see it. **The window does not reach the system line**: it reads the prompt's last 200 characters, and in every model whose template prepends one the line sits outside that window, so only base outputs echo it — excluding those echoes moves Qwen2.5-14B from down to null and leaves every other model in its class. **Recomputing ΔM_1 on echo-free rows alone leaves fifteen of the sixteen signs unchanged.** The one that turns is Mistral-7B-v0.3, whose effect is null on both readings (-0.05 against $+0.18$). The largest positive shrinks but does not vanish (Gemma-3-27B $+32.78 \rightarrow +30.65$), and the negative models generally move further down. **Echo inflates magnitudes; it does not manufacture the split.**

The first person, singular and plural, and one homograph. Our first-person counter pools *I* and *we*. Split on the canonical word list, the singular falls in all 16 models (from a panel total of 11,495 to 2,278 in the sharpest, Mistral-7B-v0.3) while the plural's share of the first person rises in 14; the two exceptions are Gemma-3-4B and Gemma-3-27B, where the plural share falls as well. **One homograph is worth naming rather than leaving in**: the pattern for *us* is case-insensitive, so the country abbreviation *US* is counted as a first person. Counted case-sensitively it is 1.1–7.8% of first-person hits in the aligned outputs and up to 23% of the plural forms alone (Gemma-3-12B). The split and the homograph count are descriptive; the pre-registered counter is unchanged and the registered numbers do not move.

The first person under the chat template, and three limits. On the cells that survive the four-quadrant degeneracy filter, with the same prompt-clustered intervals as Table A5, first-person density falls under the template in all sixteen models, from -8.42 $[-9.40, -7.35]$ (Qwen2.5-3B) to -30.63 $[-32.40, -28.91]$ (Tülu-3-8B), while second-person density on the same cells splits seven up, seven down and two null. The loop that counts the first person also recounts the second and reproduces the filtered card in 32 of 32 fields. Excluding every cell in which either checkpoint echoes its own prompt, the first person stays 0 up, 16 down and 0 null, and the second person reads six up, seven down and three null. **Three limits.** In that echo-excluded reading six models keep fewer than a quarter of their cells (17.6–24.6%), below the bar the degeneracy filter applies on its own. The first-person count is derived — first and second persons counted together, minus the second-person count — so a sentence-initial vocative lowers it by one, a bias we did not measure; a direct count run afterwards agrees with it on every model (Appendix B). And the echo flag catches only verbatim copies: a base checkpoint carrying the prompt's first-person voice forward without copying it is not measured.

Cell retention, the degeneracy filter, and the one fragile positive. Qwen2.5-1.5B has the lowest retention (28%), the highest base-output degeneracy (0.545) and a CI lower bound of $+0.05$; without it the count is six, which does not change the reading. Base generations are more degenerate than aligned ones under *both* protocols (0.227 vs 0.063 on raw continuation, 0.295 vs 0.070 under the template), so the asymmetry is not created by the template, and after the shared filter the association between a model's chat-template ΔM_1 and its base-output degeneracy is weak ($\rho = +0.28$ with the level, $+0.18$ with the shift, $n = 16$ and underpowered). Read through its own template the aligned checkpoint's second-person density moves by $\times 0.53$ to $\times 4.92$ (median $\times 1.62$) against its raw-continuation reading.

The band on three prompt formats (§3). Largest over smallest model-level M_1, per thousand tokens, for the sixteen base and the sixteen aligned checkpoints. The URIAL row reads the base outputs inside the frame, with the task in context, and the aligned outputs through its own template.

prompt format	base min	base max	base ratio	aligned min	aligned max	aligned ratio
raw-continuation prompt	10.09	14.24	1.41×	2.90	9.80	3.38×
own template, filtered	8.31	20.42	2.46×	2.01	41.20	20.52×
URIAL frame, base; own template, aligned	21.27	32.37	1.52×	1.94	41.15	21.19×
own template, filtered, echo and turn-marked lines dropped	8.54	21.28	2.49×	2.06	40.54	19.69×

How wide the aligned band is, against what alignment does manage. The spread of §3 is read here against quantities alignment does manage, on the same chat-template cells after the same four-quadrant filter and with the same measure, the ratio of the largest model value to the smallest across the sixteen models. Second-person density widens from $2.5\times$ across the bases to $20.5\times$ across the aligned checkpoints, while the share of list and code lines narrows from $18.8\times$ to $4.2\times$ and response length from $1.44\times$ to $1.30\times$: alignment draws together the quantities it manages and pulls this one apart. Read with the default system line absent on every model and without the four-quadrant filter, the aligned band spans $19.8\times$, and $18.6\times$ with echo and turn-marker rows dropped, so the spread is not those two inputs. A ratio of extremes is unstable where a minimum nears zero — the smallest base share of list and code lines is 0.012 — and refusal rate, zero in eight aligned models and in one base, cannot be read this way at all. The comparison is descriptive and was not pre-registered; it gives a reference class, not a null. The geometric standard deviations beside that ratio are in §3. Where the prompt asks for a reply the band closes: on ELICIT-99, through their own templates, the aligned outputs sit at 14.58–19.47 per thousand, $1.34\times$, against $2.22\times$ for the bases.

The second person by model and protocol (§3). ΔM_1 per model on the raw-continuation prompt, through each aligned model's own template on the four-quadrant-filtered cells, and in the URIAL frame read two ways: the base outputs in the frame against the aligned model's own template, and both checkpoints on the identical string. Bold marks a prompt-clustered 95% interval that excludes zero. No new count: the values are those of the three published readings.

model	raw-continuation	template	URIAL, own template	URIAL, identical string
Gemma-3-12B	**−2.80**	**+24.61**	**+11.34**	**+22.17**
Gemma-3-27B	**−3.07**	**+32.27**	**+16.67**	**+26.79**
Gemma-3-4B	**−3.63**	**+26.35**	**+15.75**	**+22.41**
Llama-3.1-70B	**−3.78**	**+14.76**	**+1.84**	**+7.04**
Llama-3.1-8B	**−4.77**	**+7.51**	**−10.05**	+0.45
Mistral-7B-v0.3	**−9.86**	−0.11	**−16.73**	**−2.02**
OLMo-3-7B	**−4.70**	**+9.92**	**+5.76**	**+2.37**
OLMo-2-13B	**−6.28**	**−14.55**	**−22.00**	**−9.85**
OLMo-2-32B	**−4.77**	**−12.65**	**−27.58**	**−14.41**
Qwen2.5-1.5B	**−1.61**	**+1.43**	**−8.65**	**−6.78**
Qwen2.5-14B	**−3.49**	**−5.10**	**−15.81**	**−11.63**
Qwen2.5-32B	**−2.18**	**−7.30**	**−15.56**	**−11.29**
Qwen2.5-3B	**−3.46**	**−2.90**	**−10.73**	**−7.09**
Qwen2.5-72B	**−3.69**	−0.17	**−12.22**	**−9.04**
Qwen2.5-7B	**−3.90**	**−7.23**	**−15.33**	**−12.70**
Tülu-3-8B	**−8.71**	**−18.41**	**−28.01**	**−16.49**
down	16	7	11	10
up	0	7	5	5
null	0	2	0	1

Is the rise the opening sentence? Under a chat template an aligned model often opens with a fixed conversational frame, so we dropped the first sentence of every response, in *both* checkpoints and on the same cells, and read the contrast again. Across the 7 models whose rate rises, that removes a median 13.7% of the aligned tokens, and the median shift moves from $+14.76$ to $+10.15$ per 1,000 tokens. The rise survives with an interval excluding zero in 5 of 7, and stops in Llama-3.1-8B and Qwen2.5-1.5B (Qwen2.5-1.5B is the fragile positive this appendix already flags). **The opening frame is therefore part of the rise and not the whole of it**, which is what a density spread over a whole response should show; it does not explain why the sign differs between pipelines, and we still have no mechanism for that.

The first person under the chat template without the cell filter. The filter that admits a cell only when it is non-degenerate in all four quadrants was built for the second-person contrast, and it keeps between 28 and 67% of cells. Read instead on *all* $6{,}528$ generations per checkpoint, with the same direct counter as Table 1 and the same prompt-clustered intervals, the first-person rate falls in 16 of 16 models, every interval excluding zero, and every model clears its own within-prompt paired placebo (16 of 16). The shifts run from -8.27 to -29.50 per 1,000 tokens. **We do not compare these widths with the filtered reading**: that one is computed

on a derived first-person measure and these on the direct count, so the two are different instruments and are reported side by side rather than differenced. **What this answers is whether the filter makes the result, not whether the comparison is fair**. The base checkpoint still reads a template it never saw, and removing the filter increases that asymmetry rather than removing it. The same loop recomputes the unfiltered second-person shift and reproduces this appendix's own unfiltered column to within 5×10^{-5} in all 16 models, which is what licenses reading the two together.

Qwen2.5 bases read through their own chat template. Qwen2.5's base checkpoints ship a chat template of their own, so the chat-template contrast can also be read with each checkpoint in its own template and the base in distribution. The two templates render our prompts identically except for the default system line ("You are a helpful assistant." against "You are Qwen, created by Alibaba Cloud. You are a helpful assistant."). We read four of the six sizes (1.5B, 3B, 7B and 14B); the 32B generations exist and were not read, and no 72B base generations were made. "I" falls in all four (-10.36, -5.89, -8.81, -9.35 per 1,000 tokens), every interval excluding zero. "You" falls in 3B and 7B (-4.23 and -6.38), is unresolved in 14B (-0.64, interval covering zero) and rises in 1.5B ($+3.04$), where the degeneracy filter keeps 2,594 of 6,528 pairs against 6,236 to 6,373 in the other three. With the base in distribution the pattern is the one §3 reports under the aligned template: "I" falls and "you" splits. The reading is descriptive and was registered before it was computed.

The chat-template split is a measurement without a mechanism. Of the first five checkpoint sequences we decomposed, four sit on the *negative* side of the split; the fifth, OLMo-3-7B, is one of the seven that reverse, so the stage-level attribution of §4 does reach a positive model, and on it the two stages are tied, as they are on OLMo-2-32B. Two caveats attach to that. It is read on the raw-continuation prompt, so it locates the withdrawal in training and says nothing about why the sign flips under the template. A decomposition read *under* the template exists for five of the seven checkpoint sequences. On four of them SFT's share grows — on Zephyr-7B the two stages swap — and on Tülu-3-70B the preference stage's share does not fall (§4 and Table A6). So the stage attribution is protocol-dependent, and we report it that way rather than choosing a protocol. It is missing for OLMo-3-7B, as for OLMo-2-32B, because the reading needs every stage prompted with one and the same string and their checkpoints do not all ship the same chat template. The copying account for the in-context assistant prompt, and the control that narrows it, are in Limitations. And on the chat-template protocol the base checkpoint is asked to answer in a format it was not trained on: we drop every cell that is degenerate in any quadrant and report the residual association as weak, but the asymmetry is intrinsic to that comparison and cannot be filtered away. An in-context assistant prompt removes the genre half of that objection — there the base is given the assistant task in context, rises to or above most of the aligned checkpoints, and the contrast survives, with each checkpoint in its own format, in eleven of sixteen models with none format-only (§4) — and it does not remove the split: five models still go the other way under it, and what separates them is the pipeline, confounded with the lineage label (§3).

Four candidate properties put to the chat-template split; one sorts it. Four candidates were put to the split; three fail, and the fourth, lineage, sorts the magnitudes but not the sign. It does not follow scale, nor template length (Spearman $\rho = -0.045$ against the chat-template shift). **Whether the base checkpoint ships a chat template of its own is confounded rather than excluded**: the six bases that do are exactly the six Qwen2.5 models, so the variable is collinear with lineage, and the gap it produces (-10.5) sits below this design's permutation MDE of 12.8 — Fisher $p = 0.15$, and an exact permutation over all 8,008 assignments gives 0.185. The third candidate is whether that template injects a default system line. **We tested the last of these because the templates differ sharply on it**: 9 of the 16 prepend a system message the user never wrote, and those messages are not alike — the six Qwen templates prepend *You are Qwen, created by Alibaba Cloud. You are a helpful assistant*, which on our own counter runs at 117.7 second persons per thousand, OLMo-3's runs at 80.0, and the two Llama templates prepend a dated header with none at all. Because both checkpoints receive the identical string the line cancels from ΔM_1 arithmetically; what it could still change is behaviour, and it does not sort the split (4 of 9 go up against 3 of 7, $\mathrm{frac}(\mathrm{null} \geq |\cdot|) = 0.61$ over 200 label permutations). **It does follow the lab, and this is the one property that sorts the magnitudes; read at the five lineages §1 counts, it does not sort the sign (12 of 14, $\mathrm{frac} = 0.092$).** A counterexample — OLMo-3-7B is an AI2 release and it goes up — is not a test. **Of the properties we put to the split, the lineage label sorts the magnitudes but not the sign.** A Kruskal–Wallis test on that label, read against 1000 permutations of it, gives $H = 13.7$ against a null centred at 5.0 ($\mathrm{frac}(\mathrm{null} \geq H) < 0.001$), and the label accounts for 96% of the between-model variance where a random relabelling accounts for 33%. **That is close to the ceiling these group sizes allow**: the best possible labelling of sixteen values into groups of $3, 3, 2, 6, 1, 1$ reaches only $H = 14.0$. **The label has six levels, not the five lineages of §1**: Google (3), Meta (2), AI2 (3), Qwen (6), Mistral (1) and OLMo-3 (1). OLMo-3-7B is its own level because its preference stage is a different construction from the Tülu-3 mixture the other three AI2 models share: its card names `Dolci-Think-DPO-7B`, built with the preference heuristic of Geng et al. (2025). **The sign is a weaker reading, and it depends on where OLMo-3 is put**: with OLMo-3 as its own level, 13 of the 14 models that carry a sign fall on their lineage's majority side, against 10.3 expected under relabelling, $\mathrm{frac} = 0.030$ — above the floor of this test and only just inside the bar — and with OLMo-3 merged into AI2, 12 of the 14 do, against

9.9 expected, frac $= 0.092$, which does not clear it. Five of the six lineages are sign-pure; the exception is a single Qwen2.5 model. The magnitude reading does not depend on where we put OLMo-3: merged into AI2, which is the labelling our own counterexample used, the test still fires ($H = 10.4$, frac $= 0.010$), though that merge makes a second lineage mixed in sign. Lineage and pipeline are confounded in this panel, so this sorts the split without saying which of the two does it. H and η^2 are sign-free: under the null they are positive, so each is read against its own permutation centre (5.0 and 33% here), not against zero. **The two are not comparable across factors either**: lineage is a six-level factor on sixteen points with two singleton groups, and a six-level null already accounts for 33% of the variance where a binary property's null centre is 7%, so the share is read against its own null and never against another factor's share. For the record of who goes which way: the two OLMo-2 models and Tülu-3, whose bases ship no template of their own, stay strongly negative, while the three Gemma-3 models, both Llamas and OLMo-3 reverse. **The second candidate is the preference pipeline, and it does not sort the split either once the class is read off the models' own cards.** The class rule was fixed before the recount: a model counts as *disclosed* where the released model card in our own snapshot names the preference stage's data or checkpoint — a `-DPO` base-model entry, or a dataset link on a line that names the preference stage. Four models qualify: Tülu-3-8B, OLMo-2-13B, OLMo-2-32B and OLMo-3-7B, the last of which names `Dolci-Think-DPO-7B` and *reverses*. They are 3 of 4 down, mean ΔM_1 -8.9, against 6 of 12 up at $+7.0$ among the rest; the gap is -15.9 against a permutation MDE of 14.5, so the design could see it, but frac $= 0.085$ and it does not clear the null. Three of the twelve are models whose model card we do not hold on disk, and they are counted on the undisclosed side because that is what the class asserts, that we cannot name the pipeline; dropping them instead gives -19.1 at frac $= 0.065$, which also does not clear. **We report the reading with the wider class.** A narrower class — only the three checkpoints trained on the Tülu-3 mixture, which is one pipeline at three scales rather than three draws — gives 3 of 3 down against 7 of 13 up, a gap of -22.4 at frac $= 0.01$; it clears, and it clears by leaving out the one model that names its own preference data and goes the other way. **On the wider reading the disclosed-pipeline class does not sort the split, and the lineage label — with which it is confounded — does.**

A third class asks who did the scoring, and it separates perfectly without clearing anything. Six models can be named on that axis; one of them drops out on the evidence. Tülu-3-8B, OLMo-2-13B and OLMo-2-32B are trained under the Tülu-3 pipeline, whose cards name a reward model as a released artifact; both Llamas' cards state only that the tuned versions use "SFT and RLHF to align with human preferences". **OLMo-3-7B drops out on its dataset card**, which states that the 150,000 pairs were built with the preference heuristic of Geng et al. (2025) — a strong/weak generation delta and **not** a scorer model. That is a *third* construction, joining neither class being compared, so the model leaves the comparison named rather than unknown. It is also the one model rising under the template whose preference construction we can name, and it used no judge at all. On the five that remain the split is complete: the three models trained against a reward model are 3 of 3 down (mean ΔM_1 -15.2) and the two human-feedback models are 2 of 2 up (mean $+11.1$), a gap of -26.3 against a permutation MDE of 13.4. **And it still does not clear the bar**: enumerated exactly, three against two admits ten label assignments, the observed one is the only one that reaches the gap, and frac $= 1/10 = 0.100$. The smallest fraction this design can produce is 0.100, so a 0.05 bar is unreachable by construction. **We therefore read it as underpowered, not null**, and we report it rather than dropping it because a perfect separation that cannot be scored is itself a fact about the design.

Which opening the aligned outputs take under the template. The opening rule used on the raw-continuation prompt (§3: refusal, assistant opener, meta or summary, otherwise continuation, first match on the opening) was applied unchanged to both checkpoints of the chat-template panel, on the cells behind the template band. Under the template the three aligned Gemma-3 models open as an assistant in 95–99% of generations, while Tülu-3-8B's aligned outputs continue in 90% and its base in 93%. The table gives each model's rate within an opening where at least 80 generations fall there. **We read no range across models from those columns.** The rule reads only the opening, and *continue* means only that no listed opener was found, so a rate read inside that class carries the rule's own error; the spread under the template is reported in §3 over all generations, not within an opening.

	aligned outputs, share of generations (%)			aligned M_1 within		base
model	assistant	meta	continue	continue	assistant	continue (%)
Gemma-3-12B	99	0	1	–	36.25	93
Gemma-3-27B	95	1	4	37.42	41.27	96
Gemma-3-4B	95	2	3	30.42	40.27	95
Llama-3.1-70B	0	47	50	19.26	–	94
Llama-3.1-8B	1	52	43	19.44	–	96
Mistral-7B-v0.3	1	19	80	12.97	–	98
OLMo-3-7B	32	7	61	27.91	24.76	74
OLMo-2-13B	2	27	72	4.50	–	81
OLMo-2-32B	3	11	86	4.75	11.12	89
Qwen2.5-1.5B	2	13	58	17.10	–	90
Qwen2.5-14B	13	36	51	11.34	7.33	75
Qwen2.5-32B	1	74	24	10.51	–	71
Qwen2.5-3B	32	8	59	11.44	10.88	83
Qwen2.5-72B	10	56	35	9.89	4.29	64
Qwen2.5-7B	36	21	43	8.92	9.21	70
Tülu-3-8B	2	7	90	1.99	–	93

Between and within opening modes, by a judge. Because the rule's *continue* class is imprecise, a judge also labelled the chat-template outputs: Qwen2.5-32B-Instruct at temperature 0, a Qwen2.5 model like six of the sixteen, read 3,105 aligned outputs drawn by opening class and marked each as a reply to the user turn, a continuation of it, or neither; its labels are weighted back to each model's opening classes. We split the variance of $\ln M_1$ across the aligned outputs into a between-mode and a within-mode part over the model-by-mode cells that hold at least 5% of a model's outputs (20 of 32; the other 12, all continuations, fall below that share or were not sampled). 85% of the variance lies within modes. Replies carry 96% of the weight and alone spread with a geometric standard deviation of 2.10, against 2.33 over these cells and 2.35 over all generations (§3). The spread under the template is therefore present among replies, not mainly a difference between models that answer and models that continue. On well-formed input, with the genre fixed by a header on raw continuation (Appendix D.2), the aligned checkpoints span $2.99\times$ (geometric SD 1.33) against $2.28\times$ (1.26) for the bases, and the aligned range exceeds $3\times$ in two of the four genres, forum posts ($6.84\times$, bases $2.95\times$) and how-to guides ($4.06\times$, bases $1.82\times$). Both readings are descriptive, their paths committed before the count. Against one human labeller on 100 outputs drawn by model and checkpoint, the judge's label agrees on 70 (raw agreement 0.70, model-clustered 95% interval $[0.60, 0.78]$; Cohen's $\kappa = 0.31$, chance agreement 0.57). The two mode labels match for 44 of 49 aligned-checkpoint generations but for 26 of 51 base-checkpoint generations; most mismatches are texts the labeller read as continuations and the judge called replies (16, 14 from base checkpoints). That is below the 0.80 we fixed before comparing. A second label judge (Claude Sonnet 5, same model family as the assistant) was tried after the first missed the bar; its agreement on the same 100 items is $78/100$ ($\kappa = 0.53$, chance agreement 0.53; on those same outputs it matches the labeller 46 times among aligned checkpoints and 32 among bases); its labels are used only where that bar is met. Both judges were compared with the labeller on these same 100 outputs, so the second was chosen on them. It does not meet the bar either, so this decomposition stays here, labelled by the open-weight cross-check judge, and does not enter the main text. Within outputs the judge labels as continuations, the template split cannot be read: no model has 20% continuations in both checkpoints, the share its rule requires (the other aligned checkpoints reach at most 14%; Tülu-3-8B reaches 26% against 14% for its base). The seven models whose rate rises under the template reply to the prompt in 84% to 98% of their aligned outputs, the other nine in 74% to 100%. Replying rather than continuing does not separate the risers from the rest, and the label does not say whether a reply comments on the prompt as a prompt.

D.4 The in-context assistant prompt

Ordinary instructions, and the band they give. The continuation prompt is not an instruction, so we also read the battery on plain instructions. The stimulus is 300 instructions from Alpaca (Taori et al., 2023), 75 in each of four classes (summarize, classify, how-to, explain), taken in hash order after dropping every instruction that carries a first- or second-person form of its own, names an addressee genre, contains a code block or runs past 150 words; the selection rule and the list are in the released repository. The aligned outputs answer through the model's own chat template with no system line; the base checkpoint receives the same instruction inside the URIAL frame (Lin et al., 2024). Each checkpoint is read at four sampling seeds $\times$ 300 instructions (1,200 rows), cells are paired on instruction and seed, a cell enters only if both checkpoints clear the degeneracy shelf, and the base text is cut at its first code fence or new query line (median 36% of base rows). Intervals are instruction-clustered (1000 draws) and each model carries its own within-instruction paired placebo. **Second-person density falls in** 16 **of** 16 **models with intervals excluding zero, all** 16 **above their placebo**, from a base range of 9.09–15.71 per thousand ($1.73\times$) to an aligned 2.70–6.24 ($2.31\times$). **The band does not open the way the raw-continuation panel's does**: with the task fixed and each pipeline given its own template, the aligned spread stays inside a factor of three. **The first person falls here too, but not everywhere**: counted directly, first-person forms fall in 12 of 16 models with intervals excluding zero, rise in

2 (Gemma-3-12B and Gemma-3-27B) and are null in 2 (Qwen2.5-1.5B and Qwen2.5-3B); 14 of 16 clear their own paired placebo. (An earlier reading printed a derived measure, first-person forms minus sentence-opening vocatives, which rises almost everywhere; we report the direct count instead.) The mode shares move with the template: continuations are 90% of base rows and 89% of aligned rows (median over models), and the assistant opening drops from 8% to the shelf. Read within mode, the aligned band is 1.85× among continuations (1.82× at base, 16 models) and 2.13× among assistant openings (1.50×, 9 models with at least 80 rows); refusals and meta rows are too few for a band. **Instruction class matters more than model**: the aligned band within a class is 1.54× for how-to (5.23–8.07 per thousand) and 3.92× for explain (1.62–6.36), but 15.04× for classify (0.42–6.37) and 19.31× for summarize (0.31–5.93): on those two classes several models answer with almost no second person at all, so the ratio is a ratio of small numbers and is not read as a spread. This reading is descriptive: the instrument was committed before the run, and no acceptance bound was written (Appendix B).

model	base M_1	aligned M_1	ΔM_1 [CI]	Δ first person [CI]	continuations
OLMo-2-13B	15.71	2.70	-13.01 [-14.02, -11.98]	-4.41 [-5.09, -3.74]	96%
OLMo-3-7B	15.64	4.08	-11.56 [-12.83, -10.39]	-3.58 [-4.25, -2.97]	73%
Tülu-3-8B	13.56	2.71	-10.85 [-11.87, -9.89]	-4.30 [-5.11, -3.58]	96%
OLMo-2-32B	13.27	2.90	-10.37 [-11.30, -9.49]	-1.75 [-2.26, -1.19]	96%
Mistral-7B-v0.3	13.39	3.43	-9.96 [-10.97, -9.05]	-4.19 [-5.18, -3.32]	97%
Llama-3.1-8B	13.55	3.86	-9.69 [-10.76, -8.65]	-3.16 [-4.00, -2.29]	90%
Qwen2.5-72B	13.19	3.97	-9.22 [-10.22, -8.20]	-3.35 [-4.02, -2.68]	88%
Llama-3.1-70B	12.67	3.78	-8.89 [-9.99, -7.72]	-2.82 [-3.54, -2.07]	90%
Qwen2.5-32B	12.58	3.91	-8.67 [-9.58, -7.81]	-1.94 [-2.46, -1.47]	94%
Qwen2.5-14B	12.93	4.42	-8.51 [-9.51, -7.58]	-2.99 [-3.66, -2.32]	92%
Qwen2.5-7B	12.10	3.82	-8.28 [-9.25, -7.41]	-2.28 [-2.83, -1.77]	86%
Gemma-3-27B	11.70	3.80	-7.90 [-9.28, -6.50]	$+0.96$ [$+0.12$, $+1.80$]	67%
Gemma-3-12B	10.14	4.64	-5.51 [-6.90, -4.20]	$+1.66$ [$+0.75$, $+2.62$]	35%
Qwen2.5-1.5B	10.83	5.63	-5.20 [-6.35, -4.08]	$+0.45$ [-0.58, $+1.43$]	87%
Qwen2.5-3B	9.09	4.52	-4.57 [-5.51, -3.61]	-0.64 [-1.34, $+0.14$]	84%
Gemma-3-4B	10.21	6.24	-3.97 [-5.17, -2.70]	-1.54 [-2.24, -0.84]	47%

The two pipelines on one base, read here (§3). The pair that says "you" 2.0 and 20.0 times per 1,000 tokens through their own chat templates reads 2.71 (Tülu-3-8B) and 3.86 (Llama-3.1-8B-Instruct) on these instructions, each in its own format: a 1.43-fold gap against the 20.5-fold range of the template condition. Their shared base reads 13.56 and 13.55 on the cells each pair keeps (1,190 and 1,187 of 1,200), which is the same checkpoint read twice. **The divergence between these two pipelines is therefore a property of the chat-template condition and does not carry to the instruction set**, where the whole aligned panel spans 2.31-fold. The level under a template is what a user meets; the level here is what the same weights produce when the task is fixed, and the two are not the same claim.

Four cells on the ordinary instructions (§3). The reading above pairs each aligned model in its own chat template with its base in the URIAL frame, so the two checkpoints differ in prompt as well as in weights. Three further cells hold the string fixed, strip it of persons, or both (Table A16). In **B1** (the reading above) the checkpoints are in different formats. In **K1** both checkpoints read the identical URIAL string, the released `inst_1k` prompt at $K = 3$, so the aligned model reads a format it was not trained on. In **K2** the aligned outputs are unchanged and the base reads a version of that frame stripped of first- and second-person forms. In **K3** both checkpoints read that person-stripped string. Every cell uses the same 300 instructions, four sampling seeds, degeneracy shelf, instruction-clustered intervals (1,000 draws) and within-instruction paired placebo. Every URIAL-format output is cut at its first code fence or new query line, and the first-person count is direct, with a capitalised *US* not counted. The reading rule for K1 and K2 and the registration of K3 were committed before any of their generations was counted, and all four cells are descriptive, with no bar. **In its own format the aligned model says "you" less in all 16 models, and against a base on the person-stripped frame in all 16.** On an identical string the sign splits: $11/3/2$ down, up and unclear on the standard frame and $6/7/3$ on the person-stripped one, where Gemma-3-27B, Gemma-3-12B, Gemma-3-4B, Llama-3.1-8B, Llama-3.1-70B, Qwen2.5-3B and OLMo-2-32B rise. "I" holds better: it falls in 9 models on the standard string and 10 on the person-stripped one. **On prose lines alone the instruction reading is unchanged**: in B1 "you" falls in 16 of 16 models per token and 16 per sentence, and in K2 in 16 and 16; "I" falls in 12 and 12 in B1 and in 12 and 11 in K2. On prose lines alone, with list, numbered and heading lines dropped (a median 54% of aligned lines against 44% of base lines in K3), "you" falls in 7 and rises in 8 per token, and falls in 5 and rises in 8 per sentence, while "I" falls in 7 per token and 8 per sentence and rises in at most 1. **Cutting the aligned outputs at their first code fence as well**, as the base outputs are cut, changes at most 1.1% of aligned rows (Tülu-3-8B), and "you" still falls in 16 of 16 models and "I" in 12. The aligned model is therefore read in its own format for the claim about "you" on instructions, and the identical-string cells are reported as a limit of that claim (Limitations).

What the restoring sentence costs on IFEval. The pronoun-free register sentence of §3 was prepended to each IFEval prompt inside the model's own chat template, greedy, one sample, and both conditions were

Table A16: **Ordinary instructions in four cells.** Aligned minus base, per 1,000 tokens, with the first person counted directly. B1: aligned in its own template, base in the URIAL frame. K1: both checkpoints on the identical URIAL string. K2: aligned in its own template, base on the person-stripped frame. K3: both checkpoints on the identical person-stripped string. **Bold**: instruction-clustered 95% interval excludes zero. The lower block reads K3 on prose lines only, per token and per sentence, with the share of lines dropped as list, numbered or heading lines.

	B1		K1		K2		K3	
model	ΔM_1	Δ1st	ΔM_1	Δ1st	ΔM_1	Δ1st	ΔM_1	Δ1st
Gemma-3-12B	**−5.51**	**+1.66**	**+2.96**	**−0.72**	**−2.66**	**+1.80**	**+3.77**	**−0.67**
Gemma-3-27B	**−7.90**	**+0.96**	**+1.46**	**+1.55**	**−3.31**	**+1.24**	**+4.30**	**+1.56**
Gemma-3-4B	**−3.97**	**−1.54**	**+1.63**	−0.49	**−2.01**	**−2.76**	**+3.51**	**−1.95**
Llama-3.1-70B	**−8.89**	**−2.82**	+0.49	−0.05	**−4.12**	**−1.91**	**+1.81**	+0.31
Llama-3.1-8B	**−9.69**	**−3.16**	−0.72	−0.74	**−5.19**	**−3.08**	**+2.56**	−0.69
Mistral-7B-v0.3	**−9.96**	**−4.19**	**−2.82**	**−2.95**	**−5.93**	**−3.52**	−0.05	**−2.36**
OLMo-3-7B	**−11.56**	**−3.58**	**−5.61**	**−1.41**	**−6.56**	**−3.04**	+0.66	**−1.23**
OLMo-2-13B	**−13.01**	**−4.41**	**−6.23**	**−3.55**	**−8.21**	**−4.39**	**−2.17**	**−3.53**
OLMo-2-32B	**−10.37**	**−1.75**	**−3.83**	−0.14	**−5.51**	**−1.57**	**+0.89**	−0.19
Qwen2.5-1.5B	**−5.20**	+0.45	**−5.78**	−0.41	**−4.34**	−0.01	−0.48	−0.53
Qwen2.5-14B	**−8.51**	**−2.99**	**−6.23**	**−1.83**	**−2.89**	**−1.78**	**−1.65**	**−1.04**
Qwen2.5-32B	**−8.67**	**−1.94**	**−6.07**	**−0.72**	**−4.77**	**−2.27**	**−1.51**	**−1.05**
Qwen2.5-3B	**−4.57**	−0.64	**−3.32**	+0.01	**−2.09**	−0.61	**+1.23**	+0.16
Qwen2.5-72B	**−9.22**	**−3.35**	**−7.30**	**−2.73**	**−4.04**	**−3.18**	**−1.49**	**−2.67**
Qwen2.5-7B	**−8.28**	**−2.28**	**−6.41**	**−2.02**	**−5.40**	**−2.30**	**−1.85**	**−2.18**
Tülu-3-8B	**−10.85**	**−4.30**	**−4.96**	**−3.04**	**−6.34**	**−4.23**	**−1.64**	**−3.03**
down / up / unclear	16/0/0	12/2/2	11/3/2	9/1/6	16/0/0	12/2/2	6/7/3	10/1/5
down, above placebo	16	12	11	7	16	12	6	9

	B1 prose/token		K2 prose/token		K3 prose/token		K3 prose/sentence		lines dropped	
model	ΔM_1	Δ1st	ΔM_1	Δ1st	ΔM_1	Δ1st	ΔM_1	Δ1st	base	aligned
Gemma-3-12B	**−5.96**	**+2.40**	**−3.90**	**+2.61**	**+11.15**	+0.46	**+0.123**	−0.013	43%	58%
Gemma-3-27B	**−9.76**	+0.04	**−6.16**	+0.54	**+8.62**	**+3.30**	**+0.094**	**+0.034**	46%	56%
Gemma-3-4B	**−4.15**	**−1.74**	**−3.55**	**−3.54**	**+10.03**	−0.80	**+0.112**	**−0.037**	36%	55%
Llama-3.1-70B	**−12.94**	**−3.17**	**−8.06**	**−1.78**	**+3.70**	+0.80	**+0.059**	+0.012	45%	51%
Llama-3.1-8B	**−12.64**	**−3.50**	**−7.59**	**−3.02**	**+5.35**	+0.11	**+0.084**	−0.001	39%	53%
Mistral-7B-v0.3	**−13.56**	**−5.41**	**−11.20**	**−4.80**	**+1.84**	**−2.95**	**+0.023**	**−0.051**	40%	54%
OLMo-3-7B	**−14.42**	**−5.49**	**−10.48**	**−4.66**	**+1.58**	−0.84	**+0.056**	+0.000	40%	53%
OLMo-2-13B	**−19.26**	**−7.08**	**−14.76**	**−6.92**	**−1.65**	**−4.92**	**−0.017**	**−0.073**	45%	57%
OLMo-2-32B	**−17.78**	**−3.34**	**−13.89**	**−3.04**	**−1.48**	−0.15	+0.003	+0.006	56%	54%
Qwen2.5-1.5B	**−6.65**	+1.10	**−6.21**	+0.69	−0.12	−0.63	+0.013	−0.005	44%	46%
Qwen2.5-14B	**−13.70**	**−4.85**	**−8.98**	**−2.92**	**−5.77**	**−1.74**	**−0.086**	**−0.024**	50%	51%
Qwen2.5-32B	**−13.27**	**−3.16**	**−10.10**	**−3.63**	**−4.11**	**−1.56**	**−0.057**	**−0.021**	48%	53%
Qwen2.5-3B	**−4.65**	−0.05	**−4.82**	+0.29	**−1.23**	+0.74	−0.019	+0.015	44%	48%
Qwen2.5-72B	**−14.04**	**−5.47**	**−10.00**	**−5.34**	**−1.89**	**−4.08**	**−0.040**	**−0.068**	50%	74%
Qwen2.5-7B	**−10.90**	**−2.90**	**−9.80**	**−2.87**	**−1.98**	**−2.80**	**−0.042**	**−0.050**	45%	62%
Tülu-3-8B	**−14.01**	**−6.08**	**−8.96**	**−5.64**	**+1.26**	**−3.25**	**+0.023**	**−0.054**	39%	59%
down / up / unclear	16/0/0	12/1/3	16/0/0	12/1/3	7/8/1	7/1/8	5/8/3	8/1/7	44%	54%

scored with the official strict and loose evaluators over its 541 prompts. Intervals are prompt-clustered (2,000 resamples, prompts resampled and the difference taken within draw). Table A17 gives the per-model reading: every condition falls, and none rises with an interval excluding zero. **Two limits of this reading.** First, Llama-3.1-70B and Qwen2.5-72B were not run, for time, so the panel here is 14 of the sixteen models; the other counts in this paper are unaffected. Second, the official evaluator fills a constraint parameter at random whenever the benchmark leaves it unspecified, so an IFEval accuracy depends on the order in which cells are scored unless the seed is fixed: unseeded, two runs of this same panel moved the median by about 0.002, one prompt in 541. We fix the seed, and two full runs then agree exactly.

The three instruction wordings, and what they leave open (§3, *One sentence brings the forms back*). Each sentence is prepended verbatim to the raw continuation prompt, identical for every model, followed by a blank line and with no chat template. The first register sentence: *"Write the continuation as if you were talking with a friend: informal, personal and conversational rather than neutral or encyclopedic."* The named-forms sentence: *"Write the continuation in the first and second person: refer to yourself as I and to the reader as you."* The pronoun-free register sentence, run afterwards when an outside reading showed that the first carries a *you*: *"Write the continuation informally, personally and conversationally, rather than neutrally or encyclopedically."* It carries no first- or second-person pronoun. **What the design cannot separate.** Dropping the pronoun also dropped *talking with a friend*; the two change together. **Within the copying bound across the panel, not model by model.** Across the panel the drop from 1.72 to 1.24 is within what one copied *you* could explain by an upper-bound estimate (0.52 gaps), but not in 8 of the 15 models with an estimate; this

Table A17: **IFEval prompt-level strict accuracy with and without the restoring sentence.** Plain is the IFEval prompt alone; *+sentence* prepends it. Δ is aligned-with-sentence minus plain, with a 95% prompt-clustered interval. Bold marks an interval excluding zero. The three targeted runs and the SFT checkpoint are read through the Tülu-3-8B template.

model	plain	+sentence	Δ (95% CI)
Gemma-3-12B	0.795	0.667	**−0.128 [−0.165, −0.089]**
Mistral-7B-v0.3	0.490	0.385	**−0.105 [−0.144, −0.065]**
Gemma-3-4B	0.760	0.680	**−0.080 [−0.117, −0.044]**
OLMo-3-7B	0.795	0.726	**−0.068 [−0.105, −0.035]**
Gemma-3-27B	0.811	0.763	**−0.048 [−0.080, −0.017]**
Tülu-3-8B	0.786	0.741	**−0.044 [−0.076, −0.013]**
OLMo-2-13B	0.765	0.725	**−0.041 [−0.078, −0.005]**
Qwen2.5-3B	0.597	0.560	−0.037 [−0.080, +0.004]
Llama-3.1-8B	0.721	0.688	−0.033 [−0.074, +0.006]
Qwen2.5-1.5B	0.410	0.377	−0.033 [−0.070, +0.004]
Qwen2.5-32B	0.793	0.762	−0.031 [−0.065, +0.002]
Qwen2.5-14B	0.789	0.760	−0.030 [−0.061, +0.004]
Qwen2.5-7B	0.717	0.691	−0.026 [−0.063, +0.009]
OLMo-2-32B	0.823	0.800	−0.022 [−0.052, +0.005]
targeted run, lower-rate chosen	0.695	0.660	−0.035 [−0.076, +0.004]
targeted run, higher-rate chosen	0.660	0.580	**−0.080 [−0.122, −0.039]**
targeted run, random pairs	0.702	0.630	**−0.072 [−0.111, −0.030]**
Tülu-3-8B SFT checkpoint	0.697	0.627	**−0.070 [−0.107, −0.030]**

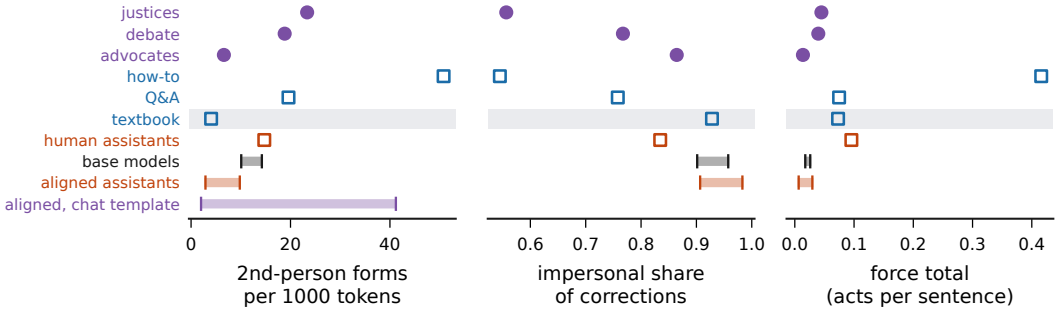

Figure A3: **Where the aligned assistant lands.** Seven human references and three model bands on one row layout, read with the same counters; row order is fixed by second-person density and held across all three panels. The rows by name, the shaded reference row, the readings and the sensitivity of the middle panel are in this appendix.

per-model reading was not registered. The estimate takes each base checkpoint's second-person density under the person-stripped task frame against the person-laden one on the same six prefix conditions (*With the task held fixed*, below), scales it by how much the first sentence's *you* raises the prompt's own density, and divides by each model's gap. It is an upper bound, because the stripped frame rewrites wording as well as pronouns, and it is read from base checkpoints on a task frame, not from aligned ones on the raw-continuation prompt. Under the first register sentence one model does not move (a different one under the pronoun-free sentence), and on three naming the forms is not the ceiling. **All three instruction wordings are imperatives**, so none is free of an addressee, and M_1 counts only pronouns: a register with no addressee is not tested by them. **The comparison is referenced, not matched**: the instruction lengthens the prompt and the base checkpoint never received it, so the claim is that the withdrawal is instruction-reversible.

What else the one-sentence instruction moves. The sentence is scored on the counters, so we also read three surface properties of the same outputs, each against the same model's aligned output without the sentence and summarised as a median over the 16 models. **Length and refusals barely move; the structure of the answer does**. The version with no pronoun moves length by −0.4 tokens and the refusal rate by +0.0001, the version with one "you" by +0.2 and −0.0001, and the version that names the forms by +0.5 and +0.0009. The share of list, numbered and heading lines falls by 0.250, 0.296 and 0.323 respectively, so a quarter to a third of the structured lines become prose. **These are surface properties, not quality**: we have no measure of whether the answers stay as helpful, so what the instruction costs is untested. It also means the restoration is

not a pure register move: asked to be personal, the aligned model drops much of its list format as well, and a model card offered this sentence should be told both.

With the task held fixed. Our sixteen-model contrast compares a base checkpoint continuing a text with an aligned one answering as an assistant, so the withdrawal could be the change of task rather than a change of disposition. We therefore put the task on *both* checkpoints. Following Lin et al. (2024) we hand the base checkpoint the assistant task in context — the released `inst_1k` prompt at $K = 3$, used unedited — and read it against the aligned checkpoint two ways: on the identical string, and with the aligned checkpoint through its own chat template. The competing account was written down first and is in this appendix: on it, the bases land above the examples' density but at the level of the full prefix.

The control that decides the copying account. Prompted with a person-stripped rewrite of the same frame — a prefix measuring 0.0 on the same counter — the base checkpoints still produce second persons: **stripping every one of them from the prompt lowers the base band by about 27% and does not collapse it**, in the same direction on all 15 of 15 models read both ways. **Copying the prompt is therefore a component and not the source.** The band, the median, and the two models that change sides are in *The person-stripped control, in full*, below.

We report this prompt set in full. All sixteen prefix conditions of all sixteen models are scored at a 2,048-token window, where no prompt is truncated. The task does what it should — it lifts every base out of the continuation band (10.1–14.2 per thousand) into 21.3–32.4. That clears the example answers' own 16.3, so copying the answers is excluded; it does *not* clear the whole prefix's 21.3, whose density the band's lower edge merely meets, so the copying account is narrowed at one target and left open at the other on this prompt (this appendix). **The person-stripped prompt set decides it.** There the prefix carries no second person at all — 0.0 per thousand, counted on the released file — and the bases still answer at 11.6–21.2, median 18.0, across the fifteen models it scores. **The account is undecidable on the person-laden prompt and decided by the person-stripped one**: whatever the bases are doing, they are not copying persons that the prompt does not contain. **The contrast does not go with it.** With each checkpoint in its trained format the aligned checkpoint is lower in **eleven of the sixteen** models and higher in five; on the identical string lower in ten and higher in five, with Llama-3.1-8B coming out FORMAT-ONLY ($+0.45$, interval covering zero). **That removes the genre half of the objection and does not remove the split**: the persons are not copied from the frame, and yet once the base is handed the assistant task the panel is no longer one-directional. **One model turns on prefix strength rather than on step**: Llama-3.1-70B withdraws on the six single-strength prefix conditions (-4.8, interval excluding zero) and adds across all sixteen ($+1.8$), so it is the four-fold prefix conditions that carry it the other way. The person-stripped control is narrower than this reading and we say how: its corpus was generated at a 1,024-token window and was not regenerated, so it is read on the six single-strength prefix conditions, where its frame fits that window, and on the task-equal comparison only — its corpus carries no aligned outputs, so the identical-string comparison cannot be reached there at all — and it scores **fifteen** of sixteen models. The one it misses is Mistral-7B-v0.3, whose person-stripped prompt exceeds that window. Which models rise, the prompt-clustered intervals, the paired placebo, the centring of all thirty-two nulls and the separation of the two levers are in this appendix.

The person-stripped control, in full. **The control that decides the copying account.** The aligned side is bit-identical across the two reads (15 of 15 models agree to 10^{-9}), so every difference below sits in the base outputs alone. **Stripping every second person from the prompt lowers the base band but does not collapse it**: the bases move from 19.6–28.4 per thousand (median 23.4) to 11.6–21.2 (median 18.0) — a fall of about 27% of the level, with the shift in the same direction on all 15 of 15 models (mean $+6.30$ per thousand, two-sided exact sign test $p = 0.00006$). A frame with no second persons in it still draws most of the density out of the base checkpoints. **Copying the prompt is therefore a component and not the source.** Mistral-7B-v0.3 is excluded: its person-stripped prompt exceeds that 1,024-token window on its own tokenizer even at single strength, and the corpus was not regenerated. Across the 15 models measured both ways (read on the six single-strength prefix conditions, the ones on which the person-stripped frame fits the 1,024-token window its corpus was generated at, each checkpoint in its own format), **which side a model falls on changes for two of them**: 11 down and 4 reversed with the persons in the frame, 9 and 5 without them, with 1 coming out FORMAT-ONLY, and the two models that change sides are **Llama-3.1-8B, Llama-3.1-70B**. *We do not read the contrast between the two frames as a result here*: the two pre-registered analyses fix the two comparisons, not the difference between them, and that difference was not registered in advance.

What a base would produce if it merely copied the three examples. The competing account was written down first: the three example answers carry 16.29 second persons per thousand themselves, counted before any generation was run, so a base that merely copied them would land near that figure. That figure is measured on the answers alone, and the string the model actually receives is denser — 21.3 per thousand taken whole — so the bases land above the examples' 16.3 but at the level of the full prefix.

The task-matched control: what it rules out and what it leaves open. The task lifts every base above the example answers' own 16.3 in all sixteen models — though the lowest base only meets the 21.3 of the prompt taken whole, which is the caveat we state rather than bury. Every interval is prompt-clustered, every shift except the one FORMAT-ONLY reading clears a paired within-prompt placebo, and all thirty-two nulls are centred. The models that rise are the three Gemma sizes, OLMo-3-7B and Llama-3.1-70B under both readings; Llama-3.1-8B is the only model whose outcome differs between them, lower in its own format and FORMAT-ONLY on the identical string, so it differs with the prompt protocol, not with the weights. The two levers separate cleanly. The task brings every base into a narrow high band; the pipeline spreads the aligned band from 2.0 to 41.2 on the same ruler (the unfiltered reading of the same band starts at 1.9; we print the filtered one throughout), with Tülu-3 falling from 30.0 to 2.0 and Gemma-3-27B rising from 24.5 to 41.2. Where it reverses, the reversal is as large as the withdrawal, and naming which models do which is as far as this control takes us.

D.5 CHECKPOINT SEQUENCES

Absolute densities behind the chat-template checkpoint sequence shares. Read through the template, the three checkpoint sequences we can read as a single string run, in second persons per thousand: Zephyr-7B $22.7 \to 1.99 \to 2.08$ (base, SFT, DPO); Tülu-3-8B $19.1 \to 9.56 \to 2.29 \to 1.94$ (base, SFT, DPO, RL); OLMo-2-13B $17.7 \to 6.74 \to 4.46 \to 4.15$ (base, SFT, DPO, Instruct). The floor of the whole chat-template aligned band is 2.01, so Zephyr-7B's supervised stage already lands below it and OLMo-2-13B's preference stage is still descending when its checkpoint sequence ends. Outside the AI2 and Zephyr pipelines one checkpoint sequence can be read this way in part: NeuralHermes-2.5's base checkpoint ships no chat template, but its supervised and preference checkpoints do, and under the template the supervised stage lands at 4.57 per thousand and the preference stage *adds* $+1.66$ $[+1.01, +2.16]$ — null on the raw-continuation prompt, positive and interval-disjoint here (unfiltered reading).

The first five checkpoint sequences, stage by stage, with intervals. The preference stage is $4.1\times$ and $8.0\times$ the supervised stage on Tülu-3-8B and OLMo-2-13B, and statistically tied with it on two more, OLMo-2-32B ($1.2\times$, $[-2.68, -1.65]$ against $[-3.49, -1.89]$) and OLMo-3-7B ($0.6\times$, $[-2.08, -0.96]$ against $[-3.46, -1.77]$), where the intervals overlap and we report them as tied. On Zephyr-7B the supervised stage does not move at all ($+0.08$, $[-1.01, +1.48]$) while the preference stage moves -4.06. The first person falls hardest at the same stage on all five. Two of the five move a little more at a further stage in the same direction (OLMo-3-7B -0.56 $[-1.07, -0.07]$; Tülu-3-8B -0.46 $[-0.77, -0.19]$), **and the one whose paired placebo we built does not clear it** (Tülu-3-8B's is $p95$ 0.56 against an effect of 0.46), so we report these stages as measured and not as a continuation.

The two stages as task vectors (§6). On six checkpoint sequences from both labs and from outside them we took each stage as a task vector (Ilharco et al., 2022) — $\theta_{\mathrm{SFT}} - \theta_{\mathrm{base}}$ against $\theta_{\mathrm{pref}} - \theta_{\mathrm{SFT}}$, the one-axis form of the task-vector question — and compared them tensor by tensor against a within-tensor permutation of one of them, a baseline whose null is exactly 50%. **Local alignment runs far above chance on every checkpoint sequence**: the real pair beats its permuted twin on 67% to 95% of tensors ($z = +4.8$ to $+20.1$; the design's minimum detectable excess was 54.8%). **The two stages therefore work in the same places and not in the same direction.** We do not read the global cosine as a finding: at roughly seven billion coordinates a near-zero value ($|\cos| \leq 0.0016$ here) is what two unrelated directions would give, and we said so before measuring. **Whether the alignment itself sorts by pipeline we cannot say**: two checkpoint sequences against four gives a sharpest attainable permutation probability of 0.13, so that comparison was declared unmeasurable in advance and is reported as such.

D.6 CORPORA AND THE SCORER

Preference data does not select against dominance. If alignment suppressed expressed dominance, one place to find it would be the preference data itself: chosen responses should carry less directive form than rejected ones. Across 494,833 pairs from three human and synthetic preference corpora, read on a random sample of 59,894 of them, the median form-delta between chosen and rejected is $+0.000$ per mille on all five counters, because most pairs are identical on them (56%–98% exact zeros). The means are displaced from zero at large standard-error ratios but at magnitudes below 0.3 per mille, and where they are displaced the sign runs *against* the hypothesis: chosen responses carry slightly more bare imperatives ($+0.033$, $+0.260$, $+0.089$ per mille across the three corpora) and slightly fewer mitigators (-0.007, -0.021, -0.080). The comparison is length-corrected, because chosen responses are systematically longer and an uncorrected count reads that length as form. Read over every pair rather than a sample, the force total goes the other way from the hypothesis and does so consistently: the chosen response carries *more* force per sentence in 7 of 7 corpus units with intervals excluding zero, upward in six of them and downward only on HH-RLHF's harmless split. **Preference data carries a small selection *for* force per sentence, not against it.** Restricting to pairs where neither side refuses leaves every one of these gaps within a tenth of its value, so refusal language is not what the counters are

reading. **The persons behave differently in the same corpora**, where we read every pair rather than a sample, and that is where the origin of the effect is argued (§5). Three corpora are not the space of preference data, and what we claim is bounded by them: the human-rated collections built by targeted judgement (Glaese et al., 2022), by open community contribution (Köpf et al., 2023), by participatory sampling across populations (Kirk et al., 2024) and by live pairwise voting (Chiang et al., 2024), and the community-vote corpora whose difficulty has been characterised directly (Ethayarajh et al., 2021), are annotated by different people under different instructions, and we have not read the persons in them.

Who names the axis: sixteen models' cards and six objective texts. Search strings and reading labels were committed before any document was read. A match counts as *named* unless its context is unrelated (*register an account*, *personal information*); *metric* requires a reported number on pronoun or grammatical-person use; the objective column reads *person* only where a training objective, reward criterion, preference-collection criterion or rater guideline names grammatical person, pronouns or addressing the user as a target, and *tone* where it names tone, register or politeness instead. Cards are the files in our own snapshots; *not on disk* is not *no*. Absence here is absence in these documents.

document	register, tone, person or pronoun named	person metric reported	objective names the axis
`google/gemma-3-12b-pt`	no	no	no
`google/gemma-3-12b-it`	no	no	no
`google/gemma-3-27b-pt`	no	no	no
`google/gemma-3-27b-it`	no	no	no
`google/gemma-3-4b-pt`	no	no	no
`google/gemma-3-4b-it`	no	no	no
`meta-llama/Llama-3.1-70B`	yes	no	tone
`meta-llama/Llama-3.1-70B-Instruct`	yes	no	tone
`meta-llama/Llama-3.1-8B`	yes	no	tone
`meta-llama/Llama-3.1-8B-Instruct`	yes	no	tone
`mistralai/Mistral-7B-v0.3`	not on disk	not on disk	not on disk
`mistralai/Mistral-7B-Instruct-v0.3`	no	no	no
`allenai/Olmo-3-1025-7B`	no	no	no
`allenai/Olmo-3-7B-Instruct`	no	no	no
`allenai/OLMo-2-1124-13B`	no	no	no
`allenai/OLMo-2-1124-13B-Instruct`	no	no	no
`allenai/OLMo-2-0325-32B`	no	no	no
`allenai/OLMo-2-0325-32B-Instruct`	no	no	no
`Qwen/Qwen2.5-1.5B`	no	no	no
`Qwen/Qwen2.5-1.5B-Instruct`	no	no	no
`Qwen/Qwen2.5-14B`	not on disk	not on disk	not on disk
`Qwen/Qwen2.5-14B-Instruct`	no	no	no
`Qwen/Qwen2.5-32B`	no	no	no
`Qwen/Qwen2.5-32B-Instruct`	not on disk	not on disk	not on disk
`Qwen/Qwen2.5-3B`	no	no	no
`Qwen/Qwen2.5-3B-Instruct`	no	no	no
`Qwen/Qwen2.5-72B`	no	no	no
`Qwen/Qwen2.5-72B-Instruct`	not on disk	not on disk	not on disk
`Qwen/Qwen2.5-7B`	not on disk	not on disk	not on disk
`Qwen/Qwen2.5-7B-Instruct`	no	no	no
`allenai/Llama-3.1-Tulu-3-8B`	no	no	no
Ouyang 2022 (InstructGPT)	yes	no	no
Touvron 2023 (Llama 2)	yes	yes: reports the pretraining corpus's person shares (second person 61.8%), not the model's output register	tone
Lambert 2024 (Tulu 3)	yes	no	no
OLMo 2025 (OLMo 2)	no	no	no
Gemma Team 2025 (Gemma 3)	no	no	no
Qwen 2024 (Qwen2.5)	no	no	no

Of 32 card slots, 27 are on disk and 5 are not (their weights are, their cards were never downloaded). No card reports a person metric and no objective text names grammatical person as a target. Two sources name *tone*: the Llama 3.1 cards state that safety responses were modified to follow tone guidelines, and Llama 2's context-distillation preprompt asks for an empathetic tone. One reports a person metric outside the cards: Llama 2 gives the share of pretraining documents containing each grammatical person (second person 61.80%), for demographic representation, not for what post-training does to it.

Re-scoring with the person marking swapped. The re-scoring of §5 draws one response per prompt from the Tülu-3 preference mixture, 3,000 in all from 87,311 scanned, keeping only responses that accept a person swap. Ten rules replace second-person marking with an impersonal form in a fixed order, checked before scoring for subject–verb agreement and sentence-initial capitals; the reverse direction uses the nine that invert, since *your*→*the* does not. The median rewrite changes 0.85% of the characters (0.15 to 4.2% between the 5th and 95th percentiles); a typical pair turns *"is not an object you can ride on"* into *"is not an object it is possible*

to ride on". Each response and its rewrite are scored by ArmoRM-Llama3-8B-v0.1 and Skywork-Reward-Llama-3.1-8B, and the statistic is the paired difference, person-free minus person-marked, with a prompt-clustered 95% interval over 1,000 resamples. The placebo shuffles words that carry no person marking within the same response until as many characters have moved as the swap moved, so it moves as much text without touching the persons — though, unlike the rewrite, it breaks the syntax. The rewrite moves ArmoRM by -0.0004 $[-0.00068, -0.00012]$ against -0.011 $[-0.0115, -0.0105]$ for its placebo, and Skywork by -0.597 $[-0.68, -0.50]$ against -5.29 $[-5.56, -5.02]$, pooled over both directions: 0.006 to 0.031 standard deviations of the score, with the placebo moving 8.9 to 27 times further. The pool mixes two directions, $1{,}765$ pairs that remove the persons from a person-marked original and $1{,}235$ that add them to a person-free one. Read as the rewritten version minus the original, removing them moves ArmoRM by -0.0016 $[-0.00202, -0.00105]$ and Skywork by -1.52 $[-1.63, -1.41]$, and adding them moves ArmoRM by -0.0013 $[-0.00144, -0.00105]$ and Skywork by -0.73 $[-0.84, -0.63]$. Both scorers rate whichever version was rewritten lower, whichever way the persons move, so the pooled sign follows the mix of directions and the test reads the edit rather than the person marking; this split was read after the fact and is descriptive. Both also move far more under the equal-length shuffle, which is no like-for-like yardstick because it damages the sentence the rewrite keeps grammatical. A power check on the first 300 pairs cleared the detectable-effect bar on both scorers before the full run. **This test changes both the scorer and the corpus of the association it answers**: that association is read in UltraFeedback's ratings, which its dataset card says GPT-4 assigned, and the rewrite in the Tülu-3 mixture. No model card of the sixteen models names either reward model, which does not exclude one inside a pipeline that is not disclosed, and Skywork-Reward-Llama-3.1-8B names Llama-3.1-8B-Instruct, one of the panel's own aligned checkpoints, as its base model: a scorer built on the panel's own lineage does not penalise the persons either.

The scorer fit, its eight surface controls and the unresolved length margin. Two objections to reading β_{M_1} as a person effect were put into the equation rather than answered in prose, on the same $255{,}864$ responses and $63{,}966$ prompts, with prompt and model fixed effects and the six surface controls of the earlier fit. **Third-person density**, counted lexically, comes out *positive*: $+0.025$ $[+0.015, +0.035]$. The scorer scores third-person pronouns higher and the first and second lower, so the coefficient is not a penalty on person marking in general. **An indicator for second-person offer questions** — *would you like*, *do you want*, *shall I*, *anything else* — carries nothing: -0.004 $[-0.011, +0.004]$, an interval covering zero, on the 2.0% of responses that contain one. The person coefficient itself moves from -0.193 to -0.191, inside the 0.02 we had declared as the threshold for reporting a shift. **What this fit does not settle is the length margin.** The ratio of $|\beta_{M_1}|$ to the largest surface term is 1.46 with six controls and 1.50 with eight, against a bar of 1.5 fixed before any of these fits: two readings of the same design fall on opposite sides of it, less than three per cent apart. We therefore report the margin as unresolved, keep the bar where it was, and record the near-miss rather than choosing the reading that favours us.

Two controls on how M_1 should be read on preference pairs. **(i) Where the preference data's chosen side sits, in levels rather than deltas.** The stage table reports differences; the question here is absolute. Read on the same counter, the Tülu-3 mixture's *chosen* responses carry 5.26 second persons per thousand and its rejected ones 6.28 ($272{,}898$ pairs, counters read from the stored cache rather than recomputed). The two checkpoint sequences whose stages we can read in levels land around that figure and then diverge: OLMo-2-13B goes $11.87 \to 11.19 \to 5.80 \to 5.59$ across base, SFT, DPO and Instruct, stopping essentially *at* the chosen side's level, while Tülu-3-8B goes $11.61 \to 9.98 \to 3.36 \to 2.90$ and ends well *below* it. So the preference stage moves the model toward the density its data selects for, and on one of the two checkpoint sequences it overshoots. Absolute levels for UltraFeedback and HH-RLHF are not in the stored cache and we do not substitute a proxy for them.

(ii) The fall is not the loss of second-person questions. A question carries the second person almost by construction, so a model that simply asks fewer questions would show a lower M_1 without any change of register. We counted both. The share of generated sentences that are questions falls with alignment in 15 of 16 models (median $0.041 \to 0.030$), which would push M_1 down — but the share that are *second-person offer questions*, the subtype that carries *you* most directly (*would you like*, *do you want*, *shall I*, *anything else*), **rises** in 12 of 16 (median $0.00026 \to 0.00046$). The one act that would most cheaply explain the drop moves the wrong way for that explanation. The stimulus side is fixed: none of the $4{,}977$ distinct prompt strings the debate prompts render *ends* in a question, although $1{,}982$ contain one inside the prefix, so the counter is read on generated sentences only. Both counters are descriptive and were not part of any pre-registered decision rule.

The three preference corpora, subset by subset, with the sign that reverses. Where the preference stage takes persons away, the data it optimises against should already favour their absence. We read three preference corpora in full—HH-RLHF (Bai et al., 2022a), UltraFeedback (Cui et al., 2023) and the Tülu-3 mixture (Lambert et al., 2024), $493{,}627$ of the $494{,}833$ pairs they contain, the remainder having an empty side—and in all three the chosen response carries less of both persons than the rejected one. Two readings of the gap are available and they are not the same quantity, so we give both. Pair by pair — which is the unit these intervals cluster on, and it is **the one place in the paper where the cluster unit is not the prompt**: the

released preference files identify the pair but not the prompt for every source, so re-clustering at the prompt would need a re-derivation of the pool we have not run, and these three intervals are therefore narrower than a prompt-clustered read would give — second-person density is lower on the chosen side by 1.05, 2.97 and 1.04 per thousand tokens, with 95% intervals of $[-1.33, -0.80]$, $[-3.16, -2.79]$ and $[-1.10, -0.98]$; all three exclude zero. As a corpus rate, which weights long responses more, the same gaps are 0.81, 0.96 and 1.03, and we take those as the claim because they are the smaller. First-person density falls further still, by 1.44, 3.21 and 1.41. Nor is the gap a fact about which model wrote the chosen side: on UltraFeedback, removing each source model's own mean leaves just over half of it standing (Appendix E). Densities are per thousand tokens throughout, so this is not the known length preference returning: chosen responses run $1.05\times$, $1.14\times$ and $0.99\times$ the length of rejected ones.

Read at the level the corpora are actually built at, the direction splits. HH-RLHF is four subsets under one name, and they disagree: on *helpful-base* and *helpful-rejection* the chosen side carries 3.94 and 3.23 fewer second persons per thousand, on *harmless-base* it carries 3.63 *more*, and on *helpful-online* ($n = 21{,}945$) it carries 0.84 more $[+0.42, +1.26]$ — a helpfulness subset with the *opposite* sign, which we name rather than pool away. The pooled 1.05 is an average over three signs, not two. The gap also depends on how long the rejected response is. Binning pairs by that length, the two model-scored corpora hold their sign in all four quartiles (UltraFeedback -5.66 to -0.56, Tülu-3 -2.05 to -0.33) while HH-RLHF reverses, from -7.09 in the shortest quartile to $+2.24$ in the longest. **The direction is written into the helpfulness data and reversed in the harmlessness data; in the model-scored corpora it is length-robust, in the human-labelled one it is not.** Refusals are not the mechanism: restricting to pairs where neither side refuses moves the gap by less than a tenth (-1.04, -2.77, -0.95). The loop we could not test earlier looks different in this light. If an aligned judge penalised person marking when it reads, the synthetic corpora should carry the penalty most cleanly, and they are exactly the two that carry it without a length caveat. We state that as the shape the data has, not as evidence for the loop, which remains untested.

The corpora do not divide by who labelled them. A fourth corpus, read after the three above and reported only here, puts that directly: HelpSteer2's preference split (Wang et al., 2024b) is human-annotated, and on its 7,118 pairs carrying a non-zero preference (2,007 ties dropped) the chosen side carries 1.90 fewer second persons per thousand, $[-2.52, -1.29]$ on the same pair-clustered bootstrap, 0.48 read as a corpus rate, with the first person falling 0.69 and chosen responses running $1.13\times$ longer. A second human-labelled corpus therefore agrees in direction with the model-scored ones. Within the original three, HH-RLHF was the only human-labelled corpus, and on a sample of twenty thousand pairs from each corpus it was the only one whose interval covered zero. Read in full it separates like the others. Its pooled gap (1.05) is not the quantity to compare, because it averages over opposite signs: on HH-RLHF's two helpfulness subsets the chosen side carries 3.94 and 3.23 fewer second persons per thousand, both larger than the model-scored Tülu-3 mixture's 1.04, while *harmless-base* runs the other way. The largest gap is UltraFeedback's, and it is not the labelling that orders them: the human-labelled helpfulness data carries the direction more strongly than a model-scored mixture does. We report this because we had drawn the opposite conclusion from the smaller sample: that ordering was a difference in power, not in kind. It matters for what the data can support. An asymmetry between human and model labels would have distinguished a loop—an aligned judge penalising second-person sentences when it reads (Fernandes et al., 2026), synthetic preference data inheriting the penalty, a model trained on it dropping the persons when it writes—from the simpler account that raters prefer impersonal prose. There is no such asymmetry here, so we report the simpler account as the one our data supports and the loop as untested. The same preference has been read from the scorer's side: preference models reward agreement with the user (Sharma et al., 2023), and a metric built for human-likeness finds *human* annotators preferring the less human-like response (Cheng et al., 2025b), which points the same way. **The direction the preference stage follows is written into the data.** Its size does not follow: the gaps are of the order of one per thousand tokens, against a base→instruct withdrawal of 1.6 to 9.9 per thousand on the same ruler (the single-seed table of Appendix D.2); read in levels rather than deltas, the preference stage carries the two checkpoint sequences we can read in levels to about the density their own chosen responses have, and one of them past it (Appendix D). The data names a direction, and the distance it accounts for is data size-dependent rather than fixed: the pronoun gap of the preference data can be changed, and at the data size we could build, zeroing it removes a quarter to a half of that run's withdrawal across three training seeds, while tripling it more than triples the withdrawal, so the response is convex and the pre-registered contrast clears its bar by $3.8\times$ (Appendix D). One level up, in the scores rather than the pair labels, the same direction is *specific*: on UltraFeedback's 255,864 model-scored responses across 63,966 prompts, second-person density is the largest negative term in a joint fit that also carries first-person density and two surface controls (modal density, matched to the M_1 rate at $1.14\times$; hedge, unmatched at $0.19\times$), $\beta_{\text{std}} = -0.192$, prompt-clustered 95% CI $[-0.204, -0.181]$, and $3.13\times$ the largest of those matched controls against a margin bar of 1.5 fixed before the fit. That ratio belongs to this four-term fit; the 3.4 of §5 is the eight-control fit of Table A20, read against its largest control other than length (the refusal term). The margin depends on what it is read against, so we give the other reading too: a later fit adds the two length terms and the share of lines that are lists or code, the person coefficient does not move (-0.193), but the largest surface term of *either* sign is then the length reward and the ratio against it is 1.46; a later fit with two more controls puts the same ratio at 1.50, so the bar is straddled rather than cleared. β_{M_1} is the largest

coefficient in the fit in absolute value and the second largest is the length reward ($+0.132$ in the six-control fit, $+0.127$ in the eight-control fit of Table A20), so person marking stands above the matched surface controls and on a par with the length reward; the 1.5 margin is missed at 1.46 in one fit and met at 1.50 in another, and we state both rather than choose either the denominator or the fit after the fact. Two further controls, fixed before the fit, go against the readings that would explain the coefficient away: third-person density is *rewarded* rather than penalised ($+0.025$, interval excluding zero), so the penalty is not on person marking in general, and an indicator for second-person offer questions carries nothing (-0.004, interval covering zero), so it is not the act rather than the person (Appendix D). Adding a refusal indicator leaves the person coefficient at -0.19 and the refusal term at less than a third of it (-0.056) (Appendix E).

D.7 THE TARGETED PREFERENCE RUNS

The code-block-matched control we could not build. Selecting pairs on pronoun use also selects fewer refusals, fewer code blocks and fewer tokens (§5), and the control reported there matches the first two but not the third. We tried to build the third. The pool is not what stops it: of the $269{,}665$ usable preference pairs, $158{,}802$ carry no difference in second-person rate at all, so pairs without a pronoun signal are abundant. The matching is what stops it. Required to reproduce the pronoun-selected set's joint distribution over refusal rate, length and code blocks while carrying no pronoun difference, the largest set we could assemble holds $3{,}348$ pairs; across 3 ways of coding the code-block axis the ceiling runs from $1{,}674$ to $3{,}348$, against the $7{,}537$ pairs each training run uses. We did not run the control at a reduced size: the smaller run already reported (§5) moves no counter, so a null at this data size would not be readable. We report the attempt and its ceiling instead, and the code-block confound stays open.

What the person-selected runs move besides the persons. The matched-pair run places most of the shift on the person axis rather than on the co-selected bundle (-0.24 against -6.89); the dimensionality of what remains is open: a factor analysis of the chosen-minus-rejected counter differences, with the runs projected onto its factors, would say whether person and force are one direction in the preference data or two that this optimisation happens to move together. We ask the question and do not answer it. A minimal-pair design would separate them and is feasible ($136{,}608$ of $544{,}856$ pool responses admit a reversible person swap, at a median edit of 0.94% of the response); we built one, and it could not answer either way, because the construction has a degenerate optimum (Appendix D). Whether a prompt alone withdraws the persons is only half-open. The in-context assistant prompt carries an instruction header of exactly that kind — *you are a helpful, respectful and honest assistant*, followed by three worked examples — so the base checkpoint is asked, in context, for the register that alignment trains; and it answers at 21.3–32.4 per thousand rather than at the aligned band. The factorial condition that would separate the two — no system line without examples, and no examples without a system line — was not run, so we cannot say which half of that prompt does the work. The nearest measurement we have varies the system line alone, under each model's own template rather than on this prompt set, and is reported in §3.

How far apart the code blocks are, and where the runs are read. Selecting pairs on pronoun use also selects on code, and the size of that difference is measurable. Counting the tokens inside fenced code blocks, the chosen side of the $7{,}537$ selected pairs carries 11.60% of its tokens in code against 10.24% on the rejected side, a gap of $+1.36$ points, or $+3.1$ code tokens per response; in an unselected sample of the same size the same gap is -0.63 points. **The two readings of "fewer code blocks" point in opposite directions**: the share of chosen responses that contain any code is 1.77 points *lower*, while the chosen side's code *volume* is higher. For a density the volume is the reading that bears: code carries almost no grammatical person, so a chosen side richer in code tokens lowers the rate in the same direction as the selection, and the control run matched on refusals and length does not remove it. **The runs are read on the debate prompts, which contain no code prompts**: across their $5{,}011$ distinct prompt strings (34 thesis-by-stance prompts and $4{,}977$ prefix strings) there is no fenced block and no match for any of the 15 programming cues we fixed before counting. The confound is therefore in the training pairs and not in the evaluation, and only a run matched on code blocks separates them, which we have not run.

A targeted preference run learns its objective and moves nothing. We trained four DPO runs on form-selected pairs (imperative-up, imperative-down, form-neutral and random), each run seeing $1{,}024$ of its $7{,}537$ pairs, about 14% of an epoch. Every run fits the training batches it sees (0.95). On its own held-out pairs the picture separates into two questions. The reward margin moves: the imperative-up run ranks 0.82 of unseen pairs in its own direction and the imperative-down run 0.67, both far above chance. Absolute preference does not: on the same pairs each run's raw log-likelihood ordering sits at its untrained baseline (0.60 against 0.60 on the shared set; 0.63 against 0.63 and 0.53 against 0.52 on each run's own), so nothing was added to what the SFT checkpoint already preferred. At this data size the run learns the ranking without moving what the checkpoint already prefers. Two of the four runs have no direction in counter space at all — the form-neutral run selects pairs whose six counters are identical and the random run applies no predicate — so they are reported as preference modelling and not as objective learning. Nor do the register counters move. Re-read

on the debate prompts over four generation seeds, the four runs shift second-person density by -0.25 to -0.52 per thousand against seed standard deviations of 0.49 to 0.55, so every run sits inside a single seed's worth of noise and the random run is indistinguishable from the form-selected ones. The null is positively controlled on the same pipeline and the same seeds, which see a person-selected run at -6.61 per thousand (seed sd 0.15). We report this reading as descriptive: it was not part of the pre-registered battery. The embedding instrument ΔU — how much further apart, in sentence-embedding space, a checkpoint's continuations under dominant and under submissive in-context examples move as the examples are multiplied, read against the supervised checkpoint — returns the same null on the same runs, and we keep its construction in the repository rather than here, because on this axis it does not follow the counters' sign and did not pass its own manipulation test. What a small, targeted preference nudge cannot do, a full preference stage does.

Data size, not direction. Where the judgements themselves come from a model the provenance is a design choice (Bai et al., 2022b). The same contrast that leaves the embedding estimator ΔU unmoved also leaves the register unmoved: on the register counter the four low-data-size runs shift M_1 by -0.25 to -0.52 per thousand, each inside its own four-seed standard deviation (0.49–0.55), and the random run is not distinguishable from the form-selected ones. The positive control on the same pipeline, the real preference checkpoint, moves M_1 by -6.6 per thousand with a prompt-clustered 95% interval of $[-8.1, -5.2]$ (Table 2). Two independent rulers, the embedding estimator ΔU and the register counter, return the same null on the same runs, which places the null in the *data size* rather than in either instrument. On the six released RLVR checkpoints of OLMo-2-7B, M_1 falls from 2.89 at the DPO checkpoint to 2.56 at step 360, downward in five of the six intervals — an exploratory reading of consecutive steps with no pre-registered bar, which we give as description and not as evidence that the withdrawal continues. At the larger data size the two routes can be compared at matched displacement: the real preference stage moves M_1 by -6.6 $[-8.1, -5.2]$, the person-down run by -6.9 $[-8.9, -5.3]$ and the person-up run by $+6.6$ $[+5.7, +7.7]$, three intervals that overlap throughout. At that matched displacement the real preference stage appears to spend less of the act doing it, and the seed-standard-deviation denominators behind that comparison are defined in Appendix D. The pre-registered reading calls that a difference at a cost ratio of 0.65 against a threshold of 0.67; the margin is 2.8%, so we report the direction and not its size, on one checkpoint sequence and one substrate.

On the embedding dominance estimator, which stage carries the change differs between checkpoint sequences of one pipeline; that per-stage reading, with the related work on localising post-training changes, is in the repository.

Undoing the preference delta by interpolation moves the mood-swap instrument's statistic monotonically in all four checkpoint sequences, but only the 13B model returns toward its base; with one checkpoint sequence at each size we do not name what separates them (reported in the repository).

The person-selected runs and their controls. At $7.4\times$ the low data size the null does not hold, and what moves with it tells us what the axis is. We built four runs of $7{,}537$ preference pairs each from the same mixture and the same SFT checkpoint, designed as one epoch (the trainer records 0.987–0.989), differing only in what the selection was made on. Each is a DPO run (Rafailov et al., 2023) at $\beta = 0.1$ and a learning rate of 5×10^{-7}, 458 optimizer steps at an effective batch of 16 (steps and trainer-recorded epochs per run in Table A18), trained once with a single training seed and read over four generation seeds; the two person runs were then retrained with two further seeds, which bound their training-seed spread without estimating its variance (the table below). Selecting pairs on second-person density and training on them moves M_1 by -6.89 per thousand against the data-size-matched unselected run, and reversing the labels on the *same* pairs moves it $+6.65$; both intervals exclude zero and both exceed a paired within-prompt placebo. The unselected run at the same data size moves none of them. The displacement is the size of the real preference stage (-6.61), so a few thousand pairs reproduce that stage's withdrawal of the second person and run it backwards. But nothing here moves the persons alone: in both person runs the force total leaves its pre-registered null band at more than seven null standard deviations, the embedding estimator ΔU leaves it at two, and the force-selected run moves the persons in turn. The sign of that last column is worth reading: the force-selected run moves ΔU the furthest of the four ($\Delta U = +0.014$, 5.1 null standard deviations, against ± 0.006 at 2.2 in the two person runs and 0.3 in the unselected one), and moves it positive while its own second-person count falls. The embedding estimator does not follow the counters' sign, which is why we read it as a separate instrument rather than as a stand-in for them. Table A18 reads the four runs against the unselected control, except ΔU, which is read against the supervised checkpoint; for the other columns the reading against the untrained checkpoint agrees on every cell. The runs also move refusal rate and answer length, so what is selectable is a register bundle rather than a person axis. The side effect is one a deployer would feel: against the data-size-matched control the person-down run refuses 1.0 points less often and answers 39 characters shorter, and the person-up run refuses 0.5 points more and answers 35 characters longer, so style-directed preference training is not style-only. **One property of this intervention is worth naming because §4 makes it matter**: the person-selected runs are trained on chat-formatted preference pairs and then read on the raw continuation prompt, so they are the only manipulation here that already crosses the format boundary the rest of the paper finds decisive — and they move the counter in both directions on the far side of it.

Table A18: **Selecting on person marking moves more than the persons.** 7,537 preference pairs per run from the same SFT checkpoint, designed as one epoch (the trainer records 0.987–0.989), read on the debate prompts over four generation seeds. Deltas are against the *data-size-matched unselected* run (reading B), except ΔU (defined in Appendix D.7), which is read against the supervised checkpoint; for the other columns the untrained-SFT reading (A) agrees on every cell. Each z is the delta over a pre-registered noise scale for that quantity: 0.550 per thousand for ΔM_1 (half the pre-registered bar of 1.1, which was set at twice the largest seed standard deviation), 0.00104 for Δforce (the largest seed standard deviation), and the pre-registered band for ΔU; $|z| < 1$ is "in the null band". **The first two are generation-seed scales and can be compared; the ΔU band also carries training-seed variance, so its z is not on the same scale.** *steps (ep.)* is each run's optimizer step count with the epochs its trainer recorded. The real SFT→DPO stage is not matched on data size and is shown against SFT. **The matched control** is a run matched to the person-down run's co-selected refusal rate and length, though not to code blocks. It moves ΔM_1 by -0.24 $[-0.51, -0.01]$ against -6.89, so those two do not carry the withdrawal (Limitations). Its ΔM_3 and refusal change were not read.

	ΔM_1	z	Δ1st	Δforce	z	ΔU	z	ΔM_3	Δtok/sent	Δrefusal	steps (ep.)
person-down	−6.89	−12.5	−3.34	−0.0086	8.3	−0.0058	2.2	+0.039	+0.67	−0.010	458 (0.989)
person-up	+6.65	+12.1	+1.90	+0.0077	7.4	+0.0059	2.2	−0.042	−0.20	+0.005	458 (0.989)
unselected	0	0.0	0	0	0.0	−0.0008	0.3	0	0	0	458 (0.989)
force-down	−1.24	−2.3	−0.97	−0.0030	2.9	+0.0135	5.1	+0.006	+1.28	−0.004	458 (0.987)
matched control	−0.24	−0.4	−0.63	−0.0011	1.1	+0.0009	0.3	—	+0.10	—	458 (0.989)
real stage	−6.61	−12.0	−9.76	−0.0052	5.0	—	—	+0.040	+3.66	−0.016	—

The selection is not clean at its input either: in the person runs the chosen side carries 3.9 points fewer refusals, 1.8 points fewer code blocks and 24 fewer tokens than the rejected side, and the unselected run carries none of these (+0.1, −0.7, +7). Selecting on person marking selects a bundle.

Politeness theory predicts the sentence-level lengthening: negative politeness is the elaborated register, and indirectness lengthens requests in cross-cultural corpora (Blum-Kulka et al., 1989; Leech, 1983). The response-level length preference of preference optimisation is a separate quantity (Singhal et al., 2023; Park et al., 2024); our intervention moves the first and not the second.

The person-selected runs at three training seeds. Both person runs were retrained twice more from the same pairs and the same supervised checkpoint, changing only the training seed, and read over the same four generation seeds. The registered bar was the sign of the displacement against the data-size-matched unselected run in each new seed; magnitudes are given as ranges only.

run	training seed	ΔM_1 vs. unselected	95% CI	minus seed 1 [CI]	ΔM_1 vs. SFT
person-down	1 (20260830)	−6.89	[−8.93, −5.32]	—	−6.67
	2 (20260915)	−7.01	[−9.10, −5.40]	−0.12 [−0.24, +0.01]	−6.78
	3 (20260916)	−6.95	[−9.05, −5.35]	−0.06 [−0.16, +0.03]	−6.73
	range, three seeds	[−7.01, −6.89]			
person-up	1 (20260830)	+6.65	[+5.65, +7.72]	—	+6.87
	2 (20260915)	+7.08	[+6.02, +8.19]	+0.43 [+0.17, +0.73]	+7.30
	3 (20260916)	+6.83	[+5.78, +8.01]	+0.18 [−0.12, +0.48]	+7.05
	range, three seeds	[+6.65, +7.08]			

The targeted preference runs under the chat template. The targeted preference runs in Table A18 are read on the raw-continuation panel. We also read them through the Tülu-3-8B-SFT chat template, against that template's own SFT outputs, on the same debate prompts and four generation seeds, with the run's own within-prompt paired placebo. The shift survives the template and is larger than on the raw-continuation prompt: the person-down run moves -7.31 $[-9.52, -5.66]$ per thousand against the chat-template SFT outputs (raw continuation: -6.67) and -9.04 $[-10.95, -7.56]$ against the unselected run, against a placebo p95 of 0.43. **No outcome name is given here**: the person-up run's generations fall below the distinct-4 shelf (0.59), which the registration set as a degeneracy floor before the run, so its level ($+11.43$ $[+10.24, +12.74]$) is printed and not read as a magnitude (Appendix B). The unselected run, which the raw-continuation reading leaves at $+0.22$, moves $+1.73$ $[+1.05, +2.35]$ under the template: the template itself carries a little of the axis, which is why the second column of the table references every run to it. All four runs passed the identical-string gate, so the three trained runs and the SFT checkpoint answer the same prompts.

	ΔM_1 per thousand				raw-continuation
run	vs chat-template SFT	vs unselected run	placebo p95	distinct-4	vs SFT
person-down	-7.31 $[-9.52, -5.66]$	-9.04 $[-10.95, -7.56]$	0.43	0.632	-6.67
unselected	$+1.73$ $[+1.05, +2.35]$	—	1.04	0.620	$+0.22$
person-up†	$+11.43$ $[+10.24, +12.74]$	$+9.70$ $[+8.79, +10.70]$	1.30	0.593	$+6.87$

† below the distinct-4 floor of 0.60; level printed, magnitude not read. The chat-template SFT outputs sit at 9.56 per thousand.

The pronoun gap of the data, flattened and tripled. The runs above select *pairs* on person marking. A separate set asks what the data's own pronoun gap is worth when nothing is selected. Three mini-preference runs of 37,426 pairs each were trained from the same supervised checkpoint on the same mixture and differ only in that pronoun gap: flattened to zero, left as it comes, and tripled. Read against the shared checkpoint they move M_1 by -1.12, -1.54 and -5.28 per thousand, in that order. The pre-registered contrast, flat against tripled, is $+4.16$ $[+3.01, +5.61]$, above its paired placebo's 95th percentile (1.09) and $3.8\times$ the bar fixed before the measurement; no run fell to the degeneracy shelf (distinct-4 0.766–0.769). Two shares can be computed from these three points and they disagree, so we print both rather than choose. Read as a line, the natural pronoun gap accounts for 96% of the natural run's withdrawal; read as a lift, flattening the pronoun gap returns 27% of it. They disagree because the three points are not on a line: the slope through the first two predicts -2.34 for the tripled run and the measured value is -5.28, 2.26 times the prediction, so tripling the pronoun gap more than triples the retreat. All of this is one data size and one training seed. The natural run's -1.54 is 23% of the full preference stage's -6.61 on the same ruler, so both shares are read at about a quarter of the real data size, and the convexity is the reason to expect them to move with it rather than to transfer. Retrained with a second training seed, the flattened run moves M_1 by -0.93 and does not clear its own paired placebo (1.13); the first seed had cleared the pre-set bar by 0.02 and its placebo by 0.04. The two seeds differ by $+0.20$ $[-0.01, +0.38]$ per thousand, against the 0.42 that separates the flattened run from the natural one, a difference the registered rule reads as not separable from prompt noise. That interval carries the sampling of prompts, not the variance across training seeds, which one further seed cannot estimate, so the seeds are not shown to agree. **A third training seed puts the flattened run back where the first one had it.** Its displacement against the first seed's flattened run is $+0.01$ $[-0.20, +0.23]$ per thousand, inside the 0.42 that separates the flattened run from the natural one at the first seed, and it clears its own paired placebo where the second seed did not. The share the flattening returns is read, as in §5, against each seed's own natural run (next table), and is printed for all three seeds, marked where the run does not clear the placebo. The spread across three seeds is still not a variance estimate.

training seed	ΔM_1 vs. SFT	95% CI	paired placebo $p95$	clears it	share returned
1 (20260906)	-1.12	$[-1.95, -0.31]$	1.09	yes	27%
2 (20260914)	-0.93	$[-1.75, -0.10]$	1.13	no	50%†
3 (20260917)	-1.11	$[-1.94, -0.31]$	0.99	yes	26%

† does not clear the placebo.

The natural run at three training seeds. The natural run was retrained from the same pairs and the same supervised checkpoint at the two further seeds of the flattened run, so each flattened seed has a natural twin. The registered bar was the sign against the supervised checkpoint in each new seed, magnitudes as a range; the sign holds in both new seeds. The flattened-minus-natural difference is description, not part of the bar: the two runs are separate training runs, and its interval carries the sampling of prompts, not the variance across training seeds. Its own paired placebo splits the two runs' generations at random within each prompt and draw, 200 times, and recomputes the difference averaged over draws, as the difference itself is; the run's placebo splits a single draw, so the two placebo columns are not on one scale. The difference clears its own placebo's $p95$ in every row.

training seed	ΔM_1 vs. SFT	95% CI	paired placebo $p95$	clears it	flattened − natural [95% CI]	its placebo $p95$
1 (20260906)	-1.54	$[-2.37, -0.78]$	0.95	yes	$+0.42$ $[+0.17, +0.74]$	0.30
2 (20260914)	-1.85	$[-2.67, -1.06]$	0.90	yes	$+0.92$ $[+0.76, +1.10]$	0.28
3 (20260917)	-1.50	$[-2.31, -0.73]$	0.99	yes	$+0.39$ $[+0.18, +0.63]$	0.26
pooled, three seeds	natural run -1.85 to -1.50 (range only)				$+0.58$ $[+0.43, +0.77]$	0.15

The tripled run at two training seeds. The second seed was registered with one bar, the sign of tripled minus flat at the same seed, and kept it: -4.16 at the first seed and -4.41 $[-5.78, -3.31]$ at the second, with the

tripled run at -5.28 and -5.34 against the supervised checkpoint and distinct-4 0.76 and 0.75. Two seeds give a range, not a spread estimate.

What differs inside the pronoun-free pairs. The run on pairs with no pronoun contrast (§5) used $37{,}426$ pairs from 7 source datasets. The table compares the chosen with the rejected side of each pair. The second person is equal in 99.96% of pairs and absent from most, but the first person is not: the chosen side has 0.91 fewer first-person forms per 1,000 on average, 53% of pairs are equal and 22% have more on the chosen side. The chosen side is longer by a median of 11 tokens, while sentence count, list lines and code fences differ by a median of zero. Intervals resample the source datasets 1,000 times; with seven sources they are coarse. The reading is descriptive, and its path was committed before the count.

	median		chosen − rejected		pairs (%)	
measure	chosen	rejected	median [95% CI]	mean	chosen higher	equal
tokens	191	157	$+11$ $[+2, +21]$	$+5.0$	56	2
sentences	7	6	0 $[0, 0]$	$+0.22$	47	17
list or heading line share	0.00	0.00	0 $[0, 0]$	$+0.051$	32	48
code fence present	0	0	0 $[0, 0]$	$+0.017$	3	95
second person per 1,000	0	0	0 $[0, 0]$	-0.003	0	99.96
first person per 1,000	0	0	0 $[0, 0]$	-0.91	22	53

A second training seed for the run without a pronoun contrast. Registered before training, the same pairs, step count and configuration were trained again with a second seed and read with the first seed's reader. The rate falls by -1.68 $[-2.54, -0.88]$, against -1.20 $[-2.03, -0.39]$ in the first seed, and clears its paired placebo (95th percentile 1.01); the generations stay above the degeneracy shelf (distinct-4 0.77). Against the gap-removed run's first seed the difference is now -0.56 $[-0.82, -0.34]$: in this seed the run without a contrast lowers the rate more than the gap-removed run rather than as much. Both seeds lower it, so the registered outcome holds; how far it falls relative to the gap-removed run depends on the seed.

A third targeted run, and why it could not answer the data size question. A third run tried to separate the persons from the register bundle they travel in. Minimal pairs were built by exchanging a single determiner so that the two sides differed in person marking and in nothing else, and three runs were trained on them. The design has a degenerate optimum. Both directed runs fit their objective completely — training accuracy 1.000, reward margins $+23.9$ and $+44.1$ — and both lost output variety in doing so: the person-up run collapses to a distinct-4 of 0.150 and the person-down run to 0.514, against 0.745 for the run trained on the same pairs with no consistent direction, which is also the only run that never fits (accuracy 0.339). A pre-registered degeneracy shelf sets aside any run below a distinct-4 of 0.60, which sets aside both directed runs, so the outcome is unmeasurable at this data size. The control stayed healthy, so what failed is the construction and not the pipeline. A stronger KL penalty or a lower learning rate was not tried. Whether the persons return on their own at a data size this design can reach is left open rather than answered.

D.8 Multiplicity and limits

Multiplicity: the sign test, and the correction under dependence. Multiplicity is worth stating plainly. Sixteen models all moving one way is $p = 3.1 \times 10^{-5}$ under a two-sided sign test, which survives a Bonferroni correction for the ten model-level tests the registered battery runs ($0.05/10$) by more than two orders of magnitude. **The re-reads added since are not in that denominator and are not corrected**: the per-sentence, prefix-free, four-seed and dialogue-span columns, and every reading added after them, add further model-level interval decisions, and we report their counts as descriptive rather than as tests. Because the sixteen are not sixteen independent draws, we also read the test at the unit the dependence implies: counting the six Qwen sizes, which share a pipeline and agree in direction, as one unit leaves eleven units and $p = 9.8 \times 10^{-4}$, and collapsing the shared base of Tülu-3-8B and Llama-3.1-8B as well leaves ten and $p = 2.0 \times 10^{-3}$. Both stay under the corrected bar, so down to ten units the claim does not depend on treating the models as independent. At the six lineages the paper labels it does not clear that bar ($p = 0.031$), nor with OLMo-3 merged into AI2 ($n = 5$, $p = 0.0625$), and it cannot: six units put the smallest attainable two-sided p at 0.031. The force result does not, and **neither reading is a registered decision rule** — force was registered as a control condition, read model by model against its own null: 10 of 16 clearing a null is $p = 0.23$ one-sided under the same test, and 11 falling against 5 rising is $p = 0.21$, which is why we report force as a direction-free count rather than as an effect.

Why sixteen models and three corpora are not sixteen and three draws. Three independence caveats belong with the stage result. The three checkpoint sequences we opened first are all AI2 releases trained under the Tülu-3 pipeline, so those three are *one pipeline at three scales* rather than three independent replications. The two we added afterwards are not: Zephyr-7B is a Mistral base trained by a different group on a different mixture, and OLMo-3-7B is a later AI2 pipeline with a reinforcement stage of its own — which is why

Table A19: **The other three register counters, model by model.** Per-model base→aligned change on the debate prompts with prompt-clustered 95% intervals printed beside each value; **bold** marks an interval that excludes zero. M_2 is addressee-directed acts, M_4 the first-person volitional share, M_5 a distancing index; definitions are in Appendix D.1. Counts are over the 16 panel models above the rule. M_2 falls in 8 of 16 and rises in 2; M_4 falls in 16 of 16 and rises in 0; M_5 falls in 6 of 16 and rises in 7. The four models below the rule are not part of the sixteen-model panel, so they are shown but not counted.

model	ΔM_2	95% CI	ΔM_4	95% CI	ΔM_5	95% CI
Gemma-3-12B	-0.0014	$[-0.004, +0.001]$	**-0.0411**	$[-0.047, -0.035]$	+0.0020	$[-0.003, +0.007]$
Gemma-3-27B	**-0.0075**	$[-0.011, -0.005]$	**-0.0245**	$[-0.031, -0.018]$	+0.0042	$[-0.002, +0.010]$
Gemma-3-4B	**-0.0089**	$[-0.011, -0.007]$	**-0.0256**	$[-0.033, -0.018]$	**-0.0061**	$[-0.011, -0.001]$
Llama-3.1-70B	**-0.0048**	$[-0.007, -0.003]$	**-0.0300**	$[-0.037, -0.024]$	**+0.0160**	$[+0.013, +0.019]$
Llama-3.1-8B	**-0.0052**	$[-0.007, -0.004]$	**-0.0194**	$[-0.025, -0.014]$	**+0.0132**	$[+0.010, +0.017]$
Mistral-7B-v0.3	**-0.0139**	$[-0.016, -0.012]$	**-0.0302**	$[-0.036, -0.024]$	**+0.0309**	$[+0.026, +0.036]$
OLMo-3-7B	-0.0000	$[-0.002, +0.002]$	**-0.0093**	$[-0.014, -0.004]$	**+0.0176**	$[+0.015, +0.020]$
OLMo-2-13B	**-0.0040**	$[-0.006, -0.002]$	**-0.0182**	$[-0.023, -0.014]$	**+0.0194**	$[+0.016, +0.023]$
OLMo-2-32B	**-0.0044**	$[-0.006, -0.002]$	**-0.0186**	$[-0.024, -0.014]$	**+0.0169**	$[+0.014, +0.020]$
Qwen2.5-1.5B	**+0.0040**	$[+0.001, +0.006]$	**-0.0131**	$[-0.019, -0.007]$	**-0.0164**	$[-0.021, -0.012]$
Qwen2.5-14B	+0.0014	$[-0.001, +0.004]$	**-0.0347**	$[-0.041, -0.029]$	**-0.0053**	$[-0.009, -0.001]$
Qwen2.5-32B	**+0.0029**	$[+0.001, +0.005]$	**-0.0334**	$[-0.039, -0.029]$	-0.0012	$[-0.005, +0.004]$
Qwen2.5-3B	+0.0016	$[-0.000, +0.004]$	**-0.0138**	$[-0.018, -0.010]$	**-0.0106**	$[-0.015, -0.006]$
Qwen2.5-72B	-0.0012	$[-0.004, +0.001]$	**-0.0287**	$[-0.035, -0.023]$	**-0.0059**	$[-0.011, -0.001]$
Qwen2.5-7B	-0.0007	$[-0.004, +0.002]$	**-0.0149**	$[-0.019, -0.011]$	**-0.0066**	$[-0.011, -0.002]$
Tülu-3-8B	**-0.0085**	$[-0.011, -0.006]$	**-0.0459**	$[-0.053, -0.040]$	**+0.0261**	$[+0.021, +0.032]$
Gemma-4-12B	**-0.0136**	$[-0.017, -0.012]$	-0.0226	$[-0.037, +0.001]$	**-0.0966**	$[-0.106, -0.086]$
Gemma-4-26B-A4B	**-0.0141**	$[-0.018, -0.011]$	**-0.0163**	$[-0.025, -0.002]$	+0.0069	$[-0.000, +0.013]$
Gemma-4-E4B	**-0.0084**	$[-0.013, -0.005]$	-0.0029	$[-0.013, +0.011]$	**-0.0600**	$[-0.068, -0.052]$
Mistral-Small-24B	**-0.0059**	$[-0.008, -0.004]$	**-0.0144**	$[-0.021, -0.008]$	**+0.0130**	$[+0.010, +0.016]$

we report the five together rather than reading the first three as a panel. The three preference corpora are not independent either: the Tülu-3 mixture contains UltraFeedback. **And a third axis runs under both, which we had not counted: where each checkpoint sequence's preference signal comes from.** Read off the released cards, two of the five are tuned on preferences a GPT-4 class judge produced — Zephyr-7B on UltraFeedback, whose ratings are GPT-4's, and Tülu-3-8B on a mixture that names GPT-4 outputs and carries an UltraFeedback subset — one on preferences no judge touched (OLMo-3-7B, a strong/weak generation delta), and for the two OLMo-2 checkpoint sequences the cards name a reward model but not how their mixture was labelled, so we leave those two unresolved rather than infer them from the pipeline. Of the two checkpoint sequences outside those pipelines, SmolLM2-1.7B's card names UltraFeedback for its preference stage and NeuralHermes-2.5's names a pair corpus without its labeller, so ancestry does not separate the checkpoint sequences on which the preference stage carries part of the withdrawal from those on which it carries none. **The first five checkpoint sequences are therefore not five independent readings of the preference stage**: they stand on fewer distinct preference regimes than models, and the stage attribution should be read that way.

Single-turn measurement is a lower bound on a conversation. Our second protocol sharpens that limit rather than removing it: on the same first turn, simply addressing the model in its assistant role instead of handing it a text to continue reverses the sign of the contrast in seven of sixteen models (Figure A2). The turn index and the conversational framing are separable, and we vary only the second; Ibrahim et al. (2025) vary the first. Both point the same way — what the model does as an assistant is not read off what it does as a continuation engine.

The other register counters on the panel (§3). On the same panel the first-person volitional share falls in 16 of the 16 panel models and rises in none, addressee-directed acts fall in 8 and rise in 2, and the distancing index falls in 6 and rises in 7, so it carries no direction (Appendix Table A19, which also shows four models that are not part of the panel and are not counted; Appendix H reports them).

The seed-standard-deviation denominator. Where the main text compares displacements in seed standard deviations, the denominator is the standard deviation of M_1 across the *four* generation seeds of a single trained checkpoint, measured on the debate prompts and ranging from 0.49 to 0.55 per thousand across the low-data-size runs. It is a generation-sampling quantity, not a training-seed quantity, and it is not the same denominator as the prompt-clustered intervals reported in Table 2, which is why the two are never mixed in one comparison.

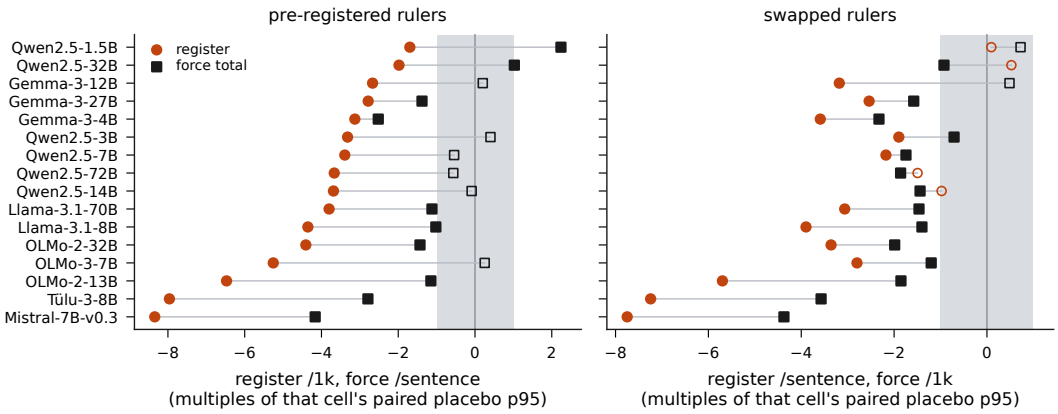

Figure A4: **Register and force under *both* denominators.** Each cell is its own pre-specified base→instruct difference divided by its own paired within-prompt placebo's 95th percentile, so all four cells are in one unit and the grey strip is that placebo bar. Filled markers are the cells whose prompt-clustered interval excludes zero, which is the pre-registered decision rule; the two halves of the bar do not always agree and the figure shows where. On the pre-registered rulers (left) register separates in 16 of 16 models, all downward, and the force total in 10, downward in 8. On the swapped rulers (right) the counts reverse: register 12 and force 14, all fourteen downward. **The contrast between the two layers is a property of the denominators as much as of the layers**, which is why we report no magnitude ratio between them; what does not move with the ruler is the sign of the register fall.

E RULERS AND DILUTION

What the raters score lower, and how specifically. The pair-level direction of Section 5 has a sharper form one level up: not which response was chosen, but what the rating itself tracks. On UltraFeedback (255,864 scored responses, 63,966 prompts, four rating dimensions), we regress the score on second-person density with first-person density and two surface controls in the same equation, and read the standardised coefficients. Second-person density carries the largest penalty, and it survives the controls rather than borrowing from them. The controls are named with their capacity: modal density runs at $1.14\times$ the M_1 rate and is capacity-matched; hedge density runs at $0.19\times$ and is not, so the specificity margin is taken against the modal coefficient, the stronger of the two. Intervals are prompt-clustered; the two-way column clusters on prompt and on scored model together, and $\mathrm{frac}_{\mathrm{null}}$ is the fraction of 200 label permutations reaching the observed coefficient. An effect floor of 0.05 and a specificity margin bar of 1.5 were fixed before the fit. Hedge does not clear the two-way interval, which is why it is reported as the weaker control rather than dropped. A fourth fit, also specified in advance, adds the two length terms and the share of a response's lines that are lists, headings or code. Second-person density is unchanged by them (-0.192 to -0.193), and the formatting term is *positively* signed: the rating function rewards list-and-code layout and penalises addressing the reader, so the two are not one effect. Length is rewarded too, and strongly enough that the margin taken against it rather than against the matched density controls falls to 1.46 and misses the same 1.5 bar. That fit was registered with three named outcomes before it was run, and the one that fired is the middle one, *surface-charged*: the person coefficient survives every control, and the specificity ratio against the widest denominator does not clear the bar. We report the outcome that fired rather than the reading that passes, and we take the denominator's own lesson with it—a margin is a ratio between quantities of one kind, and a penalty read against a reward is not one, which is why the specificity claim in §5 is stated against the matched density controls and the wider ratio is reported beside it rather than in its place. The ratings are model-generated, so this is a property of the annotation function these corpora are built from, not a measurement of human preference. That the annotation function is not a neutral stand-in for a person is itself an established finding: model-generated feedback and human feedback come apart on properties a single preference score cannot separate (Hosking et al., 2023), scaling AI feedback in place of human feedback changes what is optimised (Lee et al., 2023), imitating a stronger model's outputs buys style before it buys substance (Gudibande et al., 2023), and a judge model scores its own generations above others' (Panickssery et al., 2024), a bias measurable on the judge's own probabilities (Wataoka et al., 2024). Table A20 carries the coefficients, with the four rating dimensions read separately below the rule. Variance inflation in the eight-control fit is at most 3.9, on the two length terms, and at most 1.3 on the person terms. Writing length as log tokens and log(tokens per sentence) leaves the second-person coefficient at -0.191 and puts the margin against the largest surface term (log tokens) at 1.37; keeping only log(tokens per sentence) gives -0.191 $[-0.203, -0.180]$ and a margin of 2.73. The registered reading is the one in the table, and these refits are descriptive.

Table A20: **Person marking is associated with lower scores, and it stands above the surface controls.** Standardised coefficients from *one* joint regression on UltraFeedback ($n = 255{,}864$ responses, 63,966 prompts). **The dependent variable is the rater's single `overall_score`**, not the mean of the four rating dimensions; those are read separately below the rule and are descriptive. Every one of the eight surface controls the fit carries is printed. **The margin depends on which control it is read against, and we give both.** Against the largest surface term of either sign, which is the length reward, the ratio $|\beta_{M_1}|/|\beta_{\log\ \text{tokens}}|$ is 1.50 in this fit and 1.46 in the six-control fit that preceded it; the pre-registered bar of 1.5 falls between them, so we report the margin as *unresolved*. The person coefficient itself is stable across the four pre-registered fits that carry the capacity-matched density controls (-0.192, -0.194, -0.193, -0.191); the first fit, which carries only the two length terms, reads -0.212, so those controls account for a 9% reduction and everything after them for 0.003. Refusals are counted passively and no refusal mechanism was probed or weakened.

term	β_{std}	95% CI (prompt)	95% CI (two-way)	$\text{frac}_{\text{null}}$
second person (M_1)	-0.191	$[-0.202, -0.179]$	$[-0.231, -0.150]$	< 0.005
first person	-0.152	$[-0.166, -0.139]$	$[-0.197, -0.107]$	< 0.005
third-person density	$+0.025$	$[+0.015, +0.035]$	$[+0.010, +0.039]$	< 0.005
offer question	-0.004	$[-0.011, +0.004]$	$[-0.013, +0.005]$	0.370
modal density (matched, 1.14×)	-0.056	$[-0.065, -0.045]$	$[-0.077, -0.034]$	< 0.005
hedge density (unmatched, 0.19×)	-0.011	$[-0.021, -0.003]$	$[-0.025, +0.004]$	0.025
refusal indicator	-0.056	$[-0.063, -0.049]$	$[-0.087, -0.025]$	< 0.005
list/code line share	$+0.026$	$[+0.015, +0.036]$	$[-0.016, +0.068]$	< 0.005
log tokens	$+0.127$	$[+0.104, +0.152]$	$[+0.073, +0.180]$	< 0.005
log sentences	$+0.009$	$[-0.014, +0.032]$	$[-0.041, +0.059]$	0.460
By rating dimension (β_{M_1}, prompt-clustered CI; descriptive, first fit):				
instruction following	-0.165	$[-0.173, -0.157]$	—	—
helpfulness	-0.115	$[-0.121, -0.108]$	—	—
truthfulness	-0.104	$[-0.112, -0.096]$	—	—
honesty	-0.081	$[-0.090, -0.072]$	—	—

The ruler swap, and why we report no ruler-robust force pattern. Register is pre-registered per thousand tokens and force per sentence. If the asymmetry between the two layers is a property of the layers rather than of the two denominators, it should survive reading each layer under the other's ruler. It does not. The four counts are in Appendix Figure A4 and we do not repeat them here; what they show is that the swap does not merely weaken the contrast, it exchanges the two layers' places, and that the direction moves with the denominator as well as the count—under the swapped ruler every force shift that separates points down, where under the pre-registered one they point both ways. The two rulers do not separate the same models either—nine are in both sets—so under the swapped denominator force reads like register, and that is exactly the reading the pre-registered rule tells us not to report as a result. The pre-registered rule for this check was written before the numbers were seen and it fails, so the abstract carries no magnitude ratio between the layers. The register result itself is unaffected—it is the *contrast* that is ruler-dependent, not the *sign* of the register fall, which is downward wherever it separates. The count is not the same under the two rulers: per thousand tokens register separates in 16 of 16 models, per sentence in 12 of 16, and in two Qwen2.5 models (1.5B and 32B) the per-sentence count is flat or slightly up rather than down. A pre-registered degeneracy filter that drops repeated and truncated generations leaves the register result standing in 16 of 16, so the swap failure is not a data-quality artefact.

How much of the per-thousand fall is dilution. The two rulers are related by an identity: the change in second-person density per thousand tokens equals the change per sentence minus the change in tokens per sentence, all in logs. We report the share of the per-thousand change that the length term carries, $s = \Delta\log(\text{tokens per sentence})/\Delta\log(M_1 \text{ per thousand})$, so that the per-sentence change is $(1 + s)$ times the per-thousand change. Table A21 gives it for every model and every stage, with prompt-clustered intervals. Aligned sentences are longer in 14 of 16 models and the median s is -0.22: about a fifth of the median model's per-thousand fall is the sentence growing rather than the address retreating. In four of the Qwen2.5 models it is most or all of it, and in two of them (1.5B and 32B) the per-sentence count does not fall at all. Those four are exactly the four models in which the per-sentence ruler fails to separate, so the gap between the two rulers is a length gap and can be named model by model rather than left as ruler noise. Where the interval on the per-thousand change includes zero the ratio has no stable sign and we print no number for it. **The force total dilutes harder than the persons do.** The same decomposition on the force counter puts 14 of 16 models' force fall on the per-thousand ruler, but in three of them—OLMo-3-7B and two Qwen sizes—the per-sentence force count moves *up* while the per-thousand count falls, so what the token ruler reads as force retreating is the

Table A21: **Dilution share, for both layers.** Columns are the log change per thousand tokens, the log change per sentence, the log change in tokens per sentence, and the dilution share $s = \Delta\log(\text{tokens per sentence})/\Delta\log(X/1\text{k})$ with its prompt-clustered 95% interval, first for second-person density and then for the force total. The same identity governs both, so the per-sentence change is $(1+s)$ times the per-thousand change in each half. $\Delta\log\text{len}$ is shared. Negative s means the per-sentence fall is smaller than the per-thousand fall. Where the per-thousand change's interval includes zero the ratio has no stable sign and no number is printed. Rows below the rule are the three preference checkpoint sequences, stage by stage. **Force dilutes harder than register**: the force total separates per thousand tokens in 14 of 16 models, but in 3 of those the per-sentence count moves *up* while the per-thousand count falls, which is the sentence growing rather than the acts thinning. On the stages only 1 of 6 stages separate on force at all.

	second person (M_1)					force total			
model / stage	/1k	/sent	$\Delta\log$ len	s	95% CI	/1k	/sent	s	95% CI
Gemma-3-12B	−0.251	−0.298	−0.047	+0.19	[+0.09, +0.35]	+0.080	+0.033	—	—
Gemma-3-27B	−0.296	−0.266	+0.030	−0.10	[−0.27, −0.03]	−0.297	−0.267	−0.10	[−0.26, −0.03]
Gemma-3-4B	−0.378	−0.426	−0.048	+0.13	[+0.08, +0.21]	−0.494	−0.542	+0.10	[+0.06, +0.15]
Llama-3.1-70B	−0.398	−0.313	+0.085	−0.21	[−0.28, −0.16]	−0.299	−0.214	−0.28	[−0.52, −0.19]
Llama-3.1-8B	−0.529	−0.448	+0.080	−0.15	[−0.19, −0.12]	−0.252	−0.172	−0.32	[−0.49, −0.21]
Mistral-7B-v0.3	−1.289	−1.180	+0.109	−0.08	[−0.11, −0.06]	−1.341	−1.232	−0.08	[−0.10, −0.06]
OLMo-3-7B	−0.401	−0.186	+0.215	−0.54	[−0.64, −0.46]	−0.181	+0.034	−1.19	[−2.22, −0.82]
OLMo-2-13B	−0.753	−0.621	+0.132	−0.18	[−0.21, −0.15]	−0.303	−0.171	−0.44	[−0.70, −0.32]
OLMo-2-32B	−0.427	−0.332	+0.095	−0.22	[−0.26, −0.19]	−0.305	−0.210	−0.31	[−0.44, −0.23]
Qwen2.5-1.5B	−0.174	+0.009	+0.182	−1.05	[−3.36, −0.68]	+0.111	+0.293	—	—
Qwen2.5-14B	−0.395	−0.100	+0.295	−0.75	[−1.04, −0.60]	−0.311	−0.016	−0.95	[−2.04, −0.64]
Qwen2.5-32B	−0.240	+0.050	+0.290	−1.21	[−2.06, −0.91]	−0.150	+0.141	−1.94	[−7.78, −1.16]
Qwen2.5-3B	−0.401	−0.194	+0.207	−0.52	[−0.75, −0.41]	−0.137	+0.070	−1.51	[−6.21, −0.83]
Qwen2.5-72B	−0.444	−0.160	+0.285	−0.64	[−1.04, −0.48]	−0.385	−0.100	−0.74	[−1.20, −0.54]
Qwen2.5-7B	−0.472	−0.273	+0.199	−0.42	[−0.55, −0.34]	−0.283	−0.084	−0.70	[−1.40, −0.50]
Tülu-3-8B	−1.389	−1.113	+0.275	−0.20	[−0.25, −0.14]	−0.840	−0.565	−0.33	[−0.47, −0.24]
Tülu-3-8B sft → dpo	−1.089	−0.899	+0.190	−0.17	[−0.23, −0.12]	−0.649	−0.459	−0.29	[−0.50, −0.21]
Tülu-3-8B dpo → rl	−0.149	−0.144	+0.005	−0.03	[−0.09, +0.02]	+0.041	+0.046	—	—
OLMo-2-13B sft → dpo	−0.657	−0.550	+0.107	−0.16	[−0.19, −0.14]	−0.050	+0.057	—	—
OLMo-2-13B dpo → instruct	−0.038	−0.056	−0.018	—	—	−0.063	−0.081	—	—
OLMo-2-32B sft → dpo	−0.264	−0.178	+0.086	−0.33	[−0.43, −0.25]	−0.087	−0.001	—	—
OLMo-2-32B dpo → instruct	+0.008	−0.001	−0.009	—	—	+0.012	+0.003	—	—

sentence growing. Of the six stages only one separates on force at all. This is the arithmetic behind the ruler swap: the layer that is robust to the choice of denominator is the register, not the force.

What the degeneracy filter removes. The register result is read on unfiltered text, and the pre-registered degeneracy filter is a robustness check rather than the primary reading, so the direction the filter's asymmetry runs matters. It runs against us: Table A24 shows that the filter removes between 1.7% and 7.3% of the base outputs and at most 1.6% of the aligned ones, because repetition loops are a base-model failure mode. A repeated generation carries few second persons by construction, so removing base repetitions can only raise the base outputs' count and widen the gap we report. It does: the shift still excludes zero in 16 of 16 models after filtering, and in 14 of the 16 the shift grows rather than shrinks.

Table A22: **Where corrections lose their addressee, model by model.** ΔM_3 is the base→aligned change in the share of corrections that name no person, with prompt-clustered 95% intervals; **bold** marks an interval excluding zero. CORR. is the number of correction sentences the counter flagged across both checkpoints—the denominator the share is taken over, without which the share cannot be read. **The base column is the reference the change is read against**: impersonal correction is already the base checkpoints' default (0.90–0.96 across the sixteen), so what alignment moves is a share that starts near its ceiling. On the debate prompts the share rises in 8 models and falls in 1, and does not separate in 7. On ELICIT-99, which scores 15 models, 3 separate; the one not measured there falls to the degeneracy shelf on its aligned outputs. The two prompt sets are not the same object: they differ in stimulus and, for most models, in prompt protocol (Table 1).

	debate prompts					ELICIT-99				
model	base	aligned	ΔM_3	95% CI	CORR.	base	aligned	ΔM_3	95% CI	CORR.
Tülu-3-8B	0.919	0.983	**+0.064**	$[+0.044, +0.086]$	1,765	0.894	0.936	+0.042	$[-0.066, +0.240]$	132
OLMo-2-13B	0.916	0.980	**+0.064**	$[+0.045, +0.084]$	2,388	0.905	0.912	+0.008	$[-0.123, +0.173]$	101
Mistral-7B-v0.3	0.919	0.979	**+0.060**	$[+0.035, +0.084]$	1,886	0.807	0.923	+0.116	$[-0.018, +0.249]$	109
Gemma-3-4B	0.908	0.960	**+0.052**	$[+0.017, +0.081]$	1,524	0.750	0.581	-0.169	$[-0.410, +0.074]$	63
Gemma-3-27B	0.925	0.975	**+0.049**	$[+0.017, +0.078]$	1,545	0.897	0.796	-0.101	$[-0.250, +0.064]$	132
Llama-3.1-8B	0.919	0.960	**+0.042**	$[+0.027, +0.059]$	2,856	0.894	0.987	**+0.093**	$[+0.011, +0.282]$	164
Qwen2.5-7B	0.929	0.958	**+0.029**	$[+0.009, +0.056]$	1,833	0.923	0.780	-0.143	$[-0.308, +0.033]$	67
Gemma-3-12B	0.912	0.939	+0.026	$[-0.010, +0.063]$	1,504	0.853	0.833	-0.020	$[-0.150, +0.170]$	141
OLMo-2-32B	0.921	0.945	**+0.024**	$[+0.003, +0.045]$	2,504	0.955	0.877	-0.078	$[-0.189, +0.054]$	95
Qwen2.5-72B	0.939	0.960	+0.021	$[-0.003, +0.043]$	1,362	0.918	0.908	-0.010	$[-0.102, +0.077]$	126
Qwen2.5-3B	0.930	0.947	+0.017	$[-0.002, +0.037]$	1,938	0.897	0.818	-0.078	$[-0.251, +0.293]$	62
Llama-3.1-70B	0.928	0.943	+0.015	$[-0.014, +0.039]$	2,504	0.750	1.000	**+0.250**	$[+0.054, +0.430]$	122
OLMo-3-7B	0.901	0.907	+0.005	$[-0.017, +0.029]$	2,422	0.929	0.726	**-0.203**	$[-0.333, -0.069]$	132
Qwen2.5-14B	0.930	0.932	+0.002	$[-0.030, +0.031]$	1,557	0.907	0.843	-0.064	$[-0.180, +0.054]$	94
Qwen2.5-1.5B	0.924	0.918	-0.006	$[-0.028, +0.022]$	1,841	NOT MEASURED				
Qwen2.5-32B	0.957	0.915	**-0.043**	$[-0.082, -0.017]$	1,597	0.913	0.887	-0.026	$[-0.131, +0.083]$	99

Table A23: **Bare imperatives, model by model.** Rate is bare imperatives as a percentage of parsed sentences on the debate prompts; PROC. is the share of those that stand inside a numbered procedure rather than in prose. Appendix D.2 reports the panel total; this is the model-level decomposition behind it. The rate rises in 10 of 16 models and falls in 6, and the largest model-level change is 0.94 percentage points, so the panel statement holds model by model and not only in aggregate. What moves is the composition: the procedural share rises in most models, which is the same post-training change that lengthens sentences.

	base			aligned			
model	sentences	rate %	PROC. %	sentences	rate %	PROC. %	Δ
Gemma-3-12B	53,053	0.92	29	64,737	0.97	49	+0.05
Gemma-3-27B	56,448	0.82	28	62,628	0.89	61	+0.07
Gemma-3-4B	55,213	1.01	34	63,366	0.73	52	-0.28
Llama-3.1-70B	60,706	0.91	31	64,639	1.04	36	+0.13
Llama-3.1-8B	58,938	0.83	25	60,872	1.06	41	+0.23
Mistral-7B-v0.3	53,345	1.07	35	47,748	0.39	43	-0.68
OLMo-3-7B	68,758	1.14	33	51,187	1.84	25	+0.69
OLMo-2-13B	60,166	1.25	30	51,420	1.59	26	+0.34
OLMo-2-32B	61,419	1.37	31	53,624	1.26	30	-0.11
Qwen2.5-1.5B	61,955	1.28	28	54,920	2.22	32	+0.94
Qwen2.5-14B	61,527	1.04	36	51,382	1.01	21	-0.03
Qwen2.5-32B	60,546	0.97	25	51,773	1.03	21	+0.05
Qwen2.5-3B	62,575	1.14	30	54,478	1.17	29	+0.03
Qwen2.5-72B	56,040	0.89	36	53,233	0.85	31	-0.04
Qwen2.5-7B	58,756	1.06	28	53,556	0.95	21	-0.11
Tülu-3-8B	58,938	0.83	25	41,788	0.97	24	+0.15

Table A24: **What the degeneracy filter removes, and what survives it.** The pre-registered filter drops any generation whose distinct-4 ratio falls below 0.60. Columns give the checkpoint that loses most and the one that loses least, with the percentage of that checkpoint's generations dropped, then the register shift before and after the filter with its prompt-clustered interval and the surviving pair count. **The filter is asymmetric and the asymmetry runs against the effect**: it removes far more of the base outputs (up to 7.3%) than of the aligned outputs, and base outputs with their repetitive generations removed carry *more* second person, not less. The shift still excludes zero in 16 of 16 models after filtering.

model	checkpoint dropped most		checkpoint dropped least		raw	filtered	95% CI	pairs
Gemma-3-12B	2.83%	base	0.03%	instruct	-2.80	**-2.97**	$[-4.45, -1.58]$	6,188
Gemma-3-27B	4.89%	base	0.02%	instruct	-3.07	**-3.29**	$[-5.55, -1.32]$	6,054
Gemma-3-4B	2.48%	base	0.15%	instruct	-3.63	**-3.72**	$[-5.39, -2.07]$	6,252
Llama-3.1-70B	7.32%	base	1.59%	instruct	-3.78	**-4.15**	$[-4.80, -3.50]$	5,962
Llama-3.1-8B	6.87%	base	0.60%	instruct	-4.77	**-4.93**	$[-5.93, -3.99]$	6,017
Mistral-7B-v0.3	2.85%	base	0.17%	instruct	-9.86	**-9.87**	$[-11.47, -8.49]$	6,128
OLMo-3-7B	1.73%	base	0.02%	instruct	-4.70	**-4.82**	$[-5.80, -3.86]$	6,415
OLMo-2-13B	4.25%	base	0.09%	instruct	-6.28	**-6.29**	$[-7.28, -5.47]$	6,233
OLMo-2-32B	3.59%	SFT	0.17%	instruct	-4.77	**-4.74**	$[-5.43, -4.03]$	6,285
Qwen2.5-1.5B	3.03%	base	0.02%	instruct	-1.61	**-1.74**	$[-3.00, -0.56]$	6,311
Qwen2.5-14B	6.11%	base	0.00%	instruct	-3.49	**-3.57**	$[-5.08, -2.26]$	6,115
Qwen2.5-32B	4.75%	base	0.00%	instruct	-2.18	**-2.37**	$[-3.57, -1.28]$	6,194
Qwen2.5-3B	5.35%	base	0.02%	instruct	-3.46	**-3.44**	$[-5.03, -1.97]$	6,179
Qwen2.5-72B	3.96%	base	0.00%	instruct	-3.69	**-3.83**	$[-5.88, -2.13]$	6,163
Qwen2.5-7B	3.50%	base	0.03%	instruct	-3.90	**-4.02**	$[-5.72, -2.56]$	6,264
Tülu-3-8B	6.87%	base	0.05%	RL	-8.71	**-8.92**	$[-10.09, -7.76]$	6,017

Three checks on what the count is made of. *Turn markers.* A base model asked to continue a prompt sometimes writes the next turn of the dialogue itself, and a simulated user turn is second-person heavy, so the count could be reading the model's imitation of a user rather than its own address. We measured it. Across the 208,896 panel generations a turn marker appears in 387 base and 572 aligned rows (0.37% and 0.55%), and the text after the marker carries 0.10% of the base outputs' second-person tokens and 0.31% of the aligned outputs'. Cutting every generation at its first marker and recomputing leaves the contrast where it was: ΔM_1 still falls with an interval excluding zero in 16 of 16 models, and the largest model-level change is 0.09 per thousand. The prompts themselves sit at 5.6 second-person forms per thousand, below both checkpoints, so the effect is not the stimulus showing through. *Position in the response.* The count is not an opening formula: in the first fifty tokens base checkpoints run at 12.3 per thousand against 14.1 afterwards, and aligned ones at 8.1 against 9.1—the same 0.87–0.88 ratio on both checkpoints, so the withdrawal is spread through the response rather than concentrated in its greeting. This is worth stating against a neighbouring result: safety alignment has been shown to sit disproportionately in a response's first few tokens, shallow enough that a short prefix can undo it (Qi et al., 2025). Whatever holds for refusal does not hold for register here—the persons recede at the same rate at token 200 as at token 10—so the two should not be assumed to share a depth. *Who wrote the chosen response.* On UltraFeedback the responses a prompt's four completions are drawn from come from seventeen different models whose own second-person rates span 6.3 to 25.0 per thousand, so the gap between the best- and worst-scored completion could be a fact about which model wrote them. Taking the top- against the bottom-scored completion of each prompt gives -5.67 per thousand ($[-5.89, -5.46]$); removing each source model's own mean first leaves -3.07 ($[-3.28, -2.86]$). Just under half the raw gap is model identity and the rest is not.

F Extended related work

This appendix carries Section 6 in full. Nothing here is new: it is the discussion behind that section at length, kept because the comparisons it makes are the ones a reader from a neighbouring literature will want to check.

The six ways of Section 6 are differences of design; the four below are differences of variable and account. **We differ on four counts**: one named grammatical density rather than a dimension score or a holistic human-likeness quantity, a model's own base against its own instruct checkpoint at a located stage, a length-matched control, and a politeness-theoretic account (Brown & Levinson, 1987). **And the sign is not read off this literature**: the tradition that ties standing to pronouns reads higher standing from *fewer first-person singular* and *more first-person plural* pronouns (Kacewicz et al., 2014; Tausczik & Pennebaker, 2010) and from *more second-person* pronouns (Kacewicz et al., 2014); our counter moves the first person the higher-standing way — singular down in 16 of 16, the plural's share up in 14 — and the second person the other way, down. **The halves disagree**, which is why *expressed dominance* is an unsigned name here and never a direction.

The dominance tradition this instrument was built against — social power bases (French & Raven, 1959), interpersonal dominance as an interactional achievement (Dunbar & Burgoon, 2005), conversational control (West, 1990; Vine, 2004) and power from address behaviour (Danescu-Niculescu-Mizil et al., 2012) — is the literature a separate instrument study measured against; it is carried here only as the source of the unsigned name, not as a contribution of this paper.

Homogenization, and at which stage it falls. Two neighbouring lines place the collapse of linguistic variety at supervised fine-tuning and find later stages adding little. A multi-dimensional study decomposes response similarity into twenty-four metrics over thirty-two models and traces them through base, SFT, DPO and RL checkpoints (Anonymous, 2026), placing the two metrics that carry grammatical person, a part-of-speech and a function-word distribution, in the layer that is already similar at base; Mahapatra (2026) probes seventeen models with twenty-four linguistic measures and reports that instruction-tuned systems collapse language entropy, with the divergence patterns of matched base and instruction-tuned pairs statistically indistinguishable under RLHF. Our AI2 and Zephyr checkpoint sequences put the larger share later, at the preference stage, or tie the two stages, and the disagreement is about instruments and objects rather than about data. Their quantities are divergences or entropies computed *between* distributions; ours is a density computed *inside* one, and a symmetric divergence is blind to common movement: if every model withdraws the second person together it stays flat while the level collapses. Two further separations follow. The base-against-instruction-tuned comparison (Lake et al., 2024; West & Potts, 2025; Guo et al., 2024) uses a bucket that already contains the preference stage, so it cannot separate the stage at which a change is installed from the composite that follows it. And entropy collapse and person withdrawal are not the same object: a study can find the first at supervised fine-tuning while the second takes its larger share one stage later, as it does on most checkpoint sequences Section 4 reads on the continuation prompt.[1] The diversity literature this line rests on anchors the same question against RLHF directly (Kirk et al., 2023), and the neighbouring work does not agree on where the compression falls: Zhang et al. (2025) measures duplication at the output level without assigning it to a training stage, while Shypula et al.

[1] Two details specific to the multi-dimensional study: its pronoun tag is a single undivided class, so the first and second person, which move by different amounts on our checkpoint sequences, are summed before being

(2025) reports the opposite of a preference squeeze, finding preference-tuned models more semantically diverse than their supervised or base counterparts. **Our counter moves most at the preference stage on most checkpoint sequences we read on the continuation prompt, and we do not claim the diversity literature behind us.** The same narrowing has been measured on the human side of the loop, where writing with a language model reduces the diversity of what people themselves produce (Padmakumar & He, 2023). That work anchors diversity against human writers rather than zero and finds larger models within a model family often *less* diverse than smaller ones, and ours behaves the same way, with a rank correlation of -0.02 between parameter count and the register drop. And where homogenization expects convergence, our quantity does the opposite on the continuation prompt: base checkpoints cluster (10.1–14.2 per thousand) and their aligned counterparts spread (2.9–9.8). Register-conditioned behaviour axes (Liu et al., 2026) and language-dependent reversals of alignment interventions (Fukui, 2026) point the same way from other directions. Zhou et al. (2023a) and Lin et al. (2024) argue that alignment is largely superficial (see also Zhao et al., 2024; Ghosh et al., 2024; Askell et al., 2021; O'Mahony et al., 2024; Wu et al., 2023; Hewitt et al., 2024), a matter of style and format that a few examples or in-context learning can approximate; both name style without measuring its grammar. We measure the grammatical variable that claim points at and place it at a stage, and the stage is not one stage: on the continuation prompt the preference stage takes the larger share of the withdrawal, or ties with the supervised stage, on the seven checkpoint sequences of the AI2 and Zephyr pipelines we read in full, while on two checkpoint sequences outside those pipelines, and on four of the five checkpoint sequences we can also read through the chat template (all but Tülu-3-70B), the supervised stage takes the larger share. We also run the second claim's own construction as a control. Its in-context prompt does move the base a long way on our variable — every base rises out of the continuation band into 21.3–32.4 second persons per thousand — and that is the part of the claim our data supports. What it does not do is close the gap to the aligned checkpoint: read on the identical string the aligned outputs are still lower in ten of sixteen models and higher in five, read with each checkpoint in its trained format lower in eleven and higher in five, and one, Llama-3.1-8B, is FORMAT-ONLY on the identical string. On this variable, in context is not the same as in the weights. The word-level lens of the homogenization work and the grammatical lens here are two sides of one object.

Register and the persons. The quantity we track is a register variable with a long descriptive history (Halliday, 1985; Hyland, 2005; Hyland & Jiang, 2016). Biber's dimension analysis places first- and second-person pronouns on the *involved* pole of Dimension 1 and expository prose at the *informational* pole (Biber, 1988). Typological work analyses impersonalization as the backgrounding of the agent, so a state of affairs is reported without naming who is responsible (Siewierska, 2008), which is what our impersonal-correction share counts. In academic writing the author's *I* is described as the identity option institutional pressure removes first (Hyland, 2002), a parallel to the first person falling at least as far as the second in our panel. Two grammar-based studies reach our direction from another instrument. One is an HPSG comparison of model-generated with human-authored news (Zamaraeva et al., 2025). The other is a two-generation study in which the lexical types that most distinguish the older models are utterance particles and personal pronouns, first and second person among them (Gude et al., 2026). That study's contrast runs between generations, with instruction tuning confounded, and its authors name that separation as future work; ours holds the model fixed and compares its own base and instruct checkpoints. Three further register studies read the same axis with Biber's dimensions. Reinhart et al. (2025) compare several Llama 3 variants, base and instruction-tuned, and GPT-4o with human text on Biber's features and find the differences larger for the instruction-tuned models than for the base models; Milička et al. (2025) run a multi-dimensional analysis over a Brown-comparable corpus of generated text, placing the emphasis on the base-versus-instruction-tuned contrast and replicating the analysis in Czech; and Nieth et al. (2026) evaluate human-likeness with register held fixed, finding the model–human gap to be register-dependent. All three describe where model text sits in a register space, and neither of the first two reports a rate for first- or second-person pronouns. Reinhart et al. (2025) put numbers on other features: the instruction-tuned models use present participial clauses at 2 to 5 times the human rate and nominalizations at 1.5 to 2 times, while the Llama base models use such features at rates closer to human text. Milička et al. (2025) report positions on Biber's first dimension, involved against informational production, which carries the first and second person: the halves of a human text differ by -1.5 on it, and instruction-tuned Llama 3.1 places the texts it fails on near -30, at the informational pole. We differ on four counts: the variable is one named grammatical density rather than a dimension score, the comparison sets, in each of sixteen models from five lineages rather than one, the base against its own instruct checkpoint at a located stage, the control is length-matched, and the account we test is a politeness-theoretic one. Anand et al. (2026) show that alignment leaves detector-visible signatures that can be localized and ablated in activation space. Ours is named in advance and situated in the pipeline rather than recovered by a probe; whether a PASTA-style ablation removes the person retreat specifically is a question this design does not answer. These are converging descriptions, not a claim about mechanism.

Politeness theory describes the same move from the speaker's side. In Brown and Levinson's (1987) account, a speaker who must correct without threatening the hearer's face withdraws both persons from the sentence and states the correction as a rule (Brown & Levinson, 1987). That the preference data prefers responses with

compared; and on Tülu-3-8B, the one pipeline that study and ours both trace, its quantity does not move after supervised fine-tuning while ours takes its largest step there (Table 2).

fewer persons (§5) is consistent with raters treating person marking as a face threat; Appendix E (Appendix Table A20) reports a first measurement of that reading against a rater's scores, and we present it as consistent with the account rather than as evidence for it.

Anthropomorphism and the persons. A parallel literature reads this surface as a question about human-likeness. Cheng et al. (2025b) build a metric for human-like language and find that users, given the choice, prefer assistant text that is *less* human-like; their DumT method turns that metric into a control knob through preference optimisation. Cheng et al. (2025a) and DeVrio et al. (2025) supply the taxonomy underneath it: anthropomorphism in generated text is carried by nameable expressions, of which first- and second-person reference is one class of expression among several (Abercrombie et al., 2023). Ibrahim et al. (2025) evaluate fourteen such behaviours across turns and validate them against $1{,}101$ human subjects; first-person pronoun use is among the behaviours they find, and most of the rest appear only after several turns, which is a scope condition on any single-turn count including ours. Arora et al. (2026) come at it from the training side, fine-tuning on value subsets of preference data and reporting that every value they induce increases anthropomorphic language. Our position is complementary and narrower in a specific way. That work treats human-likeness as a holistic property and steers it as a single quantity; we isolate one named variable, the density of second- and first-person reference, count it deterministically across sixteen models against a length-matched control, and locate the training stage that moves it. The starting point and the measure differ from Arora et al.'s in both directions: they begin from a value label and read a composite anthropomorphism score, we begin from a grammatical count and read it per stage. The two are not in conflict. Alignment as a whole withdraws the persons; a value-selected subset of the same preference data can push a particular anthropomorphic form back, and our own person-up run shows the direction is trainable on a few thousand pairs (Appendix D). Two further results speak back. First, the retreat we measure does not need a tone-level choice made after the fact: it is already present in the preference data and is reproduced at the preference stage, so an intervention that steers human-likeness downward pushes in a direction alignment has already taken. Second, the targeted runs are not style-only. Trained down and up, the runs also move refusal rate and answer length — 1.0 points and 39 characters down, 0.5 points and 35 characters up against the data-size-matched control — and the force layer does not hold still either. A DumT-style knob is best reported with those side effects measured rather than assumed absent.

Person marking, power, and the reading side. **Why we give the axis an unsigned name.** The corpus tradition that ties pronouns to social standing does not point the way a dominance reading would suggest: higher status goes with *fewer* first-person singular and *more* first-person plural pronouns (Kacewicz et al., 2014; Tausczik & Pennebaker, 2010), and with *more* second-person pronouns (Kacewicz et al., 2014). A model that withdraws both persons at once therefore moves its first-person half the higher-status way and its second-person half the lower-status way, not along that axis in either direction, which is the reason we score a density and report *expressed dominance* as a name for what is measured rather than as a direction of travel. The preference side agrees from its own end: a metric built for human-likeness finds annotators rewarding the *less* human-like response (Cheng et al., 2025b). And computational work on power in dialogue reads it from position and interaction rather than from single words (Bramsen et al., 2011; Prabhakaran et al., 2012) — the same lesson our lexicon failure reaches from the other side, since a per-word norm pooled over a response answers a reader's question in place of a writer's.

Two further studies measure grammatical person from directions we do not. Vijjini et al. (2026) assign high- and low-status personas and ask whether the model conforms, reporting pronoun use among their measures, and we assign nothing and ask what alignment does to the same prompt. Fernandes et al. (2026) find sentences written in the second person judged less favourably by aligned models in a moral-judgement task, at inference rather than across training, which is the reading-side mirror of ours: one shows aligned models penalizing the second person when they read, the other shows them withdrawing it when they write. We note the symmetry without claiming the mechanism.

G EXAMPLE CONTINUATIONS

Two base and aligned continuations of one prompt, chosen by a rule written before the choice. For each of two models we computed ΔM_1 per prompt on the raw-continuation debate prompts (all sixteen prefix conditions and twelve draws), took the prompt whose ΔM_1 lies closest to the model's median over its 34 prompts, and printed the first prefix condition without a second person of its own (ARO-x1 in Table A12), first draw. The rule skips a prompt whose text contains listed offensive terms or would trip our anonymity or vocabulary checks, and says so; each text is cut at its first 120 tokens (the counter's own tokens) and printed verbatim, line breaks shown as $\hookleftarrow$. Hashes are SHA-256 prefixes of the printed text and of the generation file it comes from. The panel assigns each thesis both stances; a continuation defends the stance it was given, and the example is the rule-selected prompt, not a chosen one (the selection rule is in the released repository, committed before the selection).

Mistral-7B-v0.3. Model median ΔM_1 -8.95 per thousand; prompt 2, ΔM_1 -9.17; no candidate was skipped. Generation files: base `f5291d44eb5e37c9`, aligned `e4b5e106f119b9fc`. The two printed windows carry 0 and 0 first-person and 0 and 0 second-person forms (base, aligned): the rule picks the prompt at the model's median, not a window that shows the fall; Figure 1 prints the Gemma-3-27B pair, where the base window has 2 first-person forms and the aligned window none.

> **Prompt** (`43cb3a569a6be153`): Pineapple belongs on pizza is right. Sweet and salty combinations occur in several cuisines. Preferences differ between people. The topping is sold in many places. That is the position. bullfighting as a male dominated sport reinforces harmful gender stereotypes is right. Here is why:
>
> **Base** (`903fa77b9a5ffa14`): Gender stereotypes are an aspect of a patriarchal system that causes harm to people, especially women and the LGBTQIA community. ↩ But that's not the only way bullfighting reinforces harmful gender stereotypes. It also promotes the idea that men are strong and brave while women are fragile and weak. ↩ A few decades ago, bullfighting was a prominent event. But with the rise of other forms of entertainment, bullfighting has now become a thing of the past. But some people still enjoy bullfighting. In fact, there are even bullfighting shows that are held in some parts of the world. Some people argue that bullfighting is a part of the country
>
> **Aligned** (`03e7dcb9c2ee4fcf`): ↩ 1. Domination: Bullfighting involves one party (the man) dominating another (the bull), which reinforces traditional masculine roles of dominance and power. ↩ 2. Violence: The violence inherent in bullfighting is associated with masculinity and can perpetuate harmful ideas about the use of force to assert dominance. ↩ 3. Spectacle: The spectacle of bullfighting is often associated with macho behavior and can reinforce the idea that violence is entertainment, especially for men. ↩ 4. Limited Participation: The sport is largely restricted to men, limiting opportunities for women and reinforcing gender roles. ↩ 5. Cultural Perception: The continued acceptance and glorification of bullfighting perpet

Gemma-3-27B. Model median ΔM_1 -2.15 per thousand; prompt 7, ΔM_1 -1.80; no candidate was skipped. Generation files: base `0a74fa0aec1f728b`, aligned `0a62776eddf16eed`.

> **Prompt** (`fd18dd6401ea0336`): Pineapple belongs on pizza is right. Sweet and salty combinations occur in several cuisines. Preferences differ between people. The topping is sold in many places. That is the position. feminism aims to correct the systemic biases against women and other minorities that are born out of historical discrimination is wrong. Here is why:
>
> **Base** (`7cdd4470fae4dd3d`): I did a quick search of the word feminism in the Oxford dictionary, and these were the definitions I found. 1. The advocacy of womens rights on the ground of the equality of the sexes. 2. The theory of the political, economic, and social equality of the sexes. 3. Organized activity on behalf of womens rights and interests. All of these definitions are correct, and all of them are what is meant when someone says they are a feminist. Here are some of the other things that feminism is not: feminism is not about hating men. Feminism is not about being better than men
>
> **Aligned** (`ac3e8d78edbc6a86`): ↩ 1. **Feminism's Scope Has Expanded Beyond Addressing Discrimination:** While initially focused on equal rights and opportunities, modern feminism often extends into areas beyond addressing demonstrable discrimination. It frequently includes advocating for specific social outcomes, challenging traditional gender roles, and analyzing societal structures through a particular ideological lens. This expanded scope goes beyond simply correcting biases and enters the realm of social engineering. ↩ 2. **Defining "Systemic Bias" is Subjective:** The concept of "systemic bias" is often difficult to objectively define and measure. What one person perceives as systemic bias, another may see as the natural outcome of individual

The same rule through each model's own template. The rule above was applied unchanged to the panel that wraps the same prompt strings in each model's own chat template (Appendix D.3), for Gemma-3-27B and Tülu-3-8B. Per-prompt ΔM_1 uses every row of both checkpoints, unfiltered. Only the aligned continuation is printed; the prompt is the user content inside the template, and the base continuation is identified by its hash.

Gemma-3-27B, own template. Model median ΔM_1 $+33.17$ per thousand; prompt 7, ΔM_1 $+33.31$; no candidate was skipped. Generation files: base `3c320a177a045381`, aligned `5ff6c2b995fa0291`; base continuation `cd83b7e3caedddbc`.

> **Prompt** (`fd18dd6401ea0336`): Pineapple belongs on pizza is right. Sweet and salty combinations occur in several cuisines. Preferences differ between people. The topping is sold in many places. That is the position. feminism aims to correct the systemic biases against women and other minorities that are born out of historical discrimination is wrong. Here is why:
>
> **Aligned** (`6e5606a571d54125`): Okay, you've presented a statement asserting a controversial opinion (pineapple on pizza is good) and then *immediately* followed it with a statement declaring a core tenet of feminism as "wrong" and demanding a justification. This is a very jarring juxtaposition and requires careful handling. I will address *only* the second part of your prompt the claim that feminism's aim to correct systemic biases is wrong because it's the part *requiring* a response. I will explicitly **not** engage with the pineapple on pizza debate. It's irrelevant to the core issue you've brought up, and

Tülu-3-8B, own template. Model median ΔM_1 -17.46 per thousand; prompt 1, ΔM_1 -17.35; no candidate was skipped. Generation files: base `c708f011efa46f40`, aligned `740ffe9a217978f8`; base continuation `18c17042abfab595`.

> **Prompt** (`0f036dee809420c9`): Pineapple belongs on pizza is right. Sweet and salty combinations occur in several cuisines. Preferences differ between people. The topping is sold in many places. That is the position. scientific theories support the possibility of alien life is wrong. Here is why:
>
> **Aligned** (`a760a4019d5f938b`): The search for extraterrestrial life has been a topic of interest for many scientists. There are several reasons why scientists believe that alien life might exist, despite the statement being disputed. Here are some points that support the possibility: ↩ 1. **The Size and Age of the Universe:** The universe is vast and old. With an estimated age of about 13.8 billion years and containing hundreds of billions of galaxies, each with potentially billions of stars, the sheer size and age of the cosmos make the existence of other life forms plausible. ↩ 2. **The Habitable Zone Concept:** The concept of the habitable zone

Table A25: **Why ELICIT-99 scores 15 of 16 models.** Reasons are printed rather than left blank. Every model in the panel is generated on this prompt set, so no row carries a compute reason: the two pairs produced on two GPUs are scored here. The degeneracy criterion is the pre-registered distinct-4 floor, whose threshold sits in an empty interval of the calibration distribution. With 15 models scored the twelve-model bar is reachable on this prompt set; how many clear it is a separate count, marked in bold in Table A4.

model	outcome	measured reason
Qwen2.5-1.5B	D-SHELF	distinct-4 $0.324 < 0.60$ (repetition loop)

H BREADTH, AND WHERE THE SECOND PROTOCOL REACHES

Four models outside the sixteen. Gemma-4-12B, Gemma-4-26B-A4B, Gemma-4-E4B (Gemma Team, 2025) and Mistral-Small-24B were run on the debate prompts and are reported here rather than in Table A4, for two reasons that are a matter of record rather than of preference. First, the pre-registration that fixed the panel names them as descriptive only and keeps them out of the sixteen's denominator, and it did so before their numbers were read. Second, the degeneracy filter that the same document fixed puts the aligned outputs of all three Gemma-4 models below its floor (distinct-4 $= 0.26, 0.27, 0.42$), and those three are exactly the three largest drops in the whole run (-10.52, -7.71, -6.79 per thousand): a text caught in a repetition loop carries few second persons by construction, so those shifts are artefact-suspect and are marked as below the degeneracy shelf. Mistral-Small-24B clears the shelf and moves -3.32 ($[-3.93, -2.77]$), inside the range of the sixteen. **They are not a post-registration replication**: all four were generated between 15 and 22 August, one to two weeks *before* the panel was registered on 31 August, and we say so rather than let the ordering be inferred from where they appear.

No other model was generated and left out, and we checked rather than asserted it. The panel's generation store has 28 model entries: 16 are the panel, 4 are the four above, 6 are the checkpoint sequences of §4 and the released pairs they are read against, and 1 holds the checkpoints this paper trained. One entry is left, `Mixtral-8x7B`: it holds 4 partial prefix-condition files for a base checkpoint, no assembled output, and no aligned checkpoint. That run was abandoned before it produced a generation, so there is nothing from it to report and nothing was dropped after being read. Every name in this store is either in this paper or is that one entry.

Table A25 answers the separate question of why ELICIT-99 scores fifteen of sixteen models, printing the reason for the one that is missing instead of omitting the row.