ASKING BACK: INTERACTION-LAYER BEHAVIORAL TRANSFER THROUGH DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

When one language model is distilled from another, more than answer content carries over: the student can also pick up how the teacher interacts with its users. We characterize this channel by planting two interaction-level markers in a teacher's responses, a trailing follow-up question and a declarative sentence of conditional advice woven into the answer, and then asking whether students trained on those responses reproduce them. We evaluate 63 student models spanning three pretrained families, seven data conditions, and three seeds, yielding 35,343 held-out responses. Both markers transfer even though the students are never given the marker instruction at test time. When every training response is marked, the interrogative marker appears in 41–81% of held-out student responses, versus 0–1.7% for students trained on the same prompts with unmarked responses. When only 19.34% of training responses carry the declarative marker, it appears in 4.34–8.32% of student responses. A five-model reference panel shows that unrelated instruction-tuned models already produce the interrogative marker in up to 16.93% of responses but produce the declarative marker in at most 1.60%, illustrating why ambient rate matters for interpreting a marker. In a 20-participant within-subject study, low-density marking left mean interaction-quality ratings similar across conditions. Response rewriting, surface filtering, and a green-list watermark run on the same data show which parts of the signal survive different transformations. We argue that interaction behavior is a measurable dimension of distillation, while keeping three questions apart throughout: whether a behavior transfers, whether it can be detected against a reference, and whether it can be attributed to a specific teacher.

1 INTRODUCTION

Distillation turns a language model's responses into training data for another model (Hinton et al., 2015; Gu et al., 2024). When those responses are harvested from an API without authorization, providers face a provenance problem: a downstream model may incorporate their outputs without retaining an explicit record of their origin. Work on antidistillation has consequently studied both interventions that reduce the usefulness of harvested data and signals that make downstream use observable (Savani et al., 2025; Pan et al., 2025; Li et al., 2025; Xu et al., 2026). Our focus is the second objective: characterizing a trace of the teacher's behavior in a student.

Most language-model watermarking methods encode a signal in token selection. Green-list methods bias generation toward a keyed token subset, while cryptographic constructions define other keyed sampling procedures (Kirchenbauer et al., 2023; Aaronson, 2022; Christ et al., 2024). Their detectability depends on the text transformation, sampling rule, and downstream training procedure. Paraphrasing can substantially weaken token-level statistics without removing the answer's central content (Krishna et al., 2023; Sadasivan et al., 2025; Pan et al., 2025). This motivates studying a different observable: how a response relates its answer to the user.

A teacher response contains interaction decisions alongside task information. It may ask which environment the user is working in, scope a recommendation to a particular goal, or add advice about when a solution applies. We call these interaction-layer behaviors because their detection concerns a prompt–response relation rather than a keyed token count. They are still realized in text and learned through next-token training. Moving the measurement to this level does not make the signal invariant to rewriting. It instead opens a question about what distillation preserves.

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.

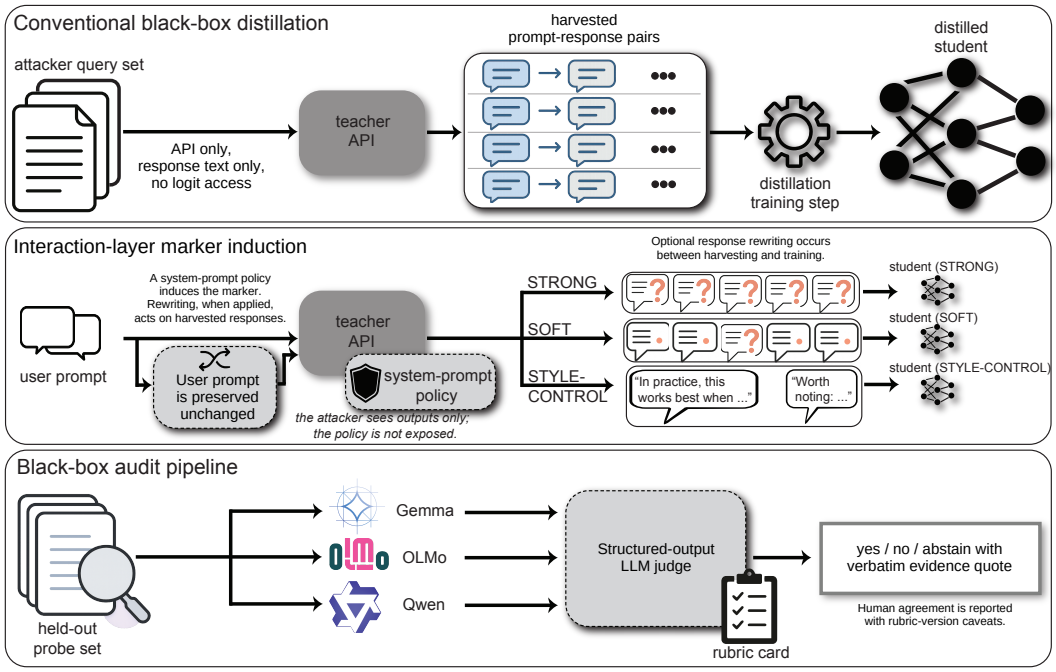


Figure 1: Interaction-layer behavioral transfer. Top: a distiller learns from a teacher's harvested prompt–response pairs. Middle: a system-prompt policy induces interrogative or declarative markers at a specified trigger fraction. In the rewriting experiments, the response is transformed after harvesting and before student training; the paired user prompt is unchanged. Bottom: students from three families are queried without the marker instruction and scored by a version-pinned judge. Human agreement and ambient reference models provide separate checks on the measurement.

We investigate this question using two controlled interventions (Figure 1). The interrogative marker appends a follow-up question about the user's situation. The declarative marker integrates a conditional advisory sentence without requesting a reply. Each has a matched unmarked training corpus. We train 63 low-rank-adapted students across three pretrained families, seven conditions, and three seeds, and evaluate them on the same 561 prompts without a marker instruction. Follow-up measurements establish per-marker ambient reference rates, test transformations at distinct stages, and check the sensitivity of the detector itself.

The positive result is consistent cross-family transfer. Interrogative rates reach 41–81% at full training density. At 19.34% density, declarative rates reach 4.34–8.32%, compared with matched unmarked rates of 0.65–1.01% under the same detector. A separate five-model reference panel shows why marker choice matters: the same unmarked Gemma instruction model produces the interrogative marker on 16.93% of responses but the declarative marker on only 0.36%. Low-density interventions also preserve similar mean interaction-quality ratings in a 20-participant study.

These findings support three contributions. First, we establish interaction-level behavior as a measurable output of distillation across disjoint model families, including a non-interrogative realization. Second, we provide a measurement framework that separates matched-corpus transfer from reference-panel detection, with versioned rubrics, explicit denominators, and paired transformation analyses. Third, we characterize the signal across response rewriting, surface filtering, and audit-time instructions, with a green-list comparison in the same teacher–student setting. The contribution is a characterization of behavioral transfer and its forensic relevance, not a claim of complete or attack-invariant attribution.

2 Related Work

Signals in model outputs and training data. Token watermarks provide a well-defined keyed statistic on generated text (Kirchenbauer et al., 2023). Cryptographic approaches study stronger indistinguishability properties (Christ et al., 2024), following early proposals such as Aaronson (2022). Their guarantees concern specified generation and detection settings. Rewriting and distillation introduce additional channels that must be evaluated empirically (Krishna et al., 2023; Sadasivan et al., 2025; Pan et al., 2025); spoofing introduces a different attribution concern (An et al., 2026). We use a green-list implementation as a concrete comparator, not as a representative of every watermark design.

Antidistillation and provenance. Antidistillation sampling and defensive output generation alter the utility of teacher outputs for a learner (Savani et al., 2025; Li et al., 2025). Instructional and antidistillation fingerprinting instead introduce signatures for model identification (Xu et al., 2024; 2026). Other work intervenes on reasoning traces (Ma et al., 2026; Ding et al., 2025). Radioactive data is a classical example of tracing training through a deliberately modified dataset (Sablayrolles et al., 2020). Membership inference addresses whether particular records were used, rather than whether a generic interaction behavior has a particular source (Shokri et al., 2017; Mireshghallah et al., 2022). These distinct inferential targets motivate our separation of transfer, detection, and attribution.

Behavioral inheritance and measurement. Models exhibit incidental stylistic regularities (Sun et al., 2025); teacher traces can persist through distillation (Wadhwa et al., 2025), and behavioral traits can transfer in settings with hidden signals in training data (Cloud et al., 2026). Our intervention is explicit at data generation and tested across different student families. Clarifying questions are independently useful dialogue behavior (Zhang et al., 2025), which makes a near-zero natural rate an empirical question, not a default assumption. We measure markers with a structured language-model judge (Zheng et al., 2023). Because self-reported confidence is not itself a calibrated probability (Kadavath et al., 2022; Xiong et al., 2024), we use categorical verdicts and report human agreement and instrument sensitivity separately.

3 Interaction Marks and Their Measurement

3.1 Teacher intervention and student observation

Let T be a teacher and $B(p, y) \in \{0, 1\}$ an operational marker on a prompt p and response y. A trigger policy Π selects a baseline or marker-inducing system prompt s. The resulting corpus is $\mathcal{D} = \{(p, T_s(p))\}$. A student S is trained by next-token imitation of this corpus. At evaluation it receives only the user prompt, without the teacher's marker instruction.

The intervention has three stages (Figure 1). We first collect ID-aligned baseline and marked responses. We then compose the designated corpus, optionally transform its response field, and fit a student. Finally, we query the student on disjoint prompts using its ordinary inference format. The teacher's system instruction is not supplied as an input feature to the student, so any marker at evaluation must be expressed through the trained model rather than copied from an audit-time instruction.

For a tokenized training sequence z, the recorded objective can be written as

$$\mathcal{L}(\theta; \mathcal{D}) = -\frac{1}{|\mathcal{D}|} \sum_{z \in \mathcal{D}} \sum_{t} m_t \log p_\theta(z_t \mid z_{<t}), \tag{1}$$

where m_t identifies positions included in the loss. OLMo and Qwen use response positions; Gemma uses all non-padding sequence positions. Within each family, marked and baseline students share the objective and hyperparameters. The control isolates the teacher-side data intervention within that recipe, while the cross-family comparison tests its recurrence under different pretrained models and training formats.

We distinguish the trigger fraction q from the achieved marker rate ρ_T. On-trigger compliance need not equal one, and off-trigger responses may contain the marker naturally. For a fixed corpus,

$$\rho_T = q\rho_{\text{on}} + (1 - q)\rho_{\text{off}}. \tag{2}$$

Table 1: Two marker forms and three interventions. Trigger fraction is a data-generation choice, not a detector threshold. Both low-density corpora reuse the same 582 prompt IDs.

Condition	Trigger	Form	Operational target
STRONG	100%	Interrogative	Final follow-up question about the user's use case, environment, or intended application.
SOFT	19.34%	Interrogative	The same question instruction, mixed with unmarked responses.
STYLE-CONTROL	19.34%	Declarative	Integrated conditional advice that scopes the answer to a scenario or goal, without a follow-up question.

STRONG uses the interrogative instruction on every row. SOFT uses exactly the same instruction on 582 of 3,009 rows, with baseline responses on the remainder. STYLE-CONTROL uses the same 582 IDs for a declarative instruction. Thus $q = 19.34\%$ for both low-density conditions, and their comparison does not change the trigger subset. The full prompts appear in Appendix B.

3.2 THREE INFERENTIAL TARGETS

The marker rate on the common evaluation distribution is

$$\tau(S, B) = \mathbb{E}_{p \sim \mathcal{P}_{\text{eval}}}[B(p, S(p))], \qquad \widehat{\tau} = \frac{n_{\text{yes}}}{n_{\text{yes}} + n_{\text{no}}}. \tag{3}$$

Abstentions and execution errors are reported separately. Our implementation $\widehat{B}$ is a version-pinned judge, not an observed ground-truth behavior label. A difference between marked and unmarked students trained with the same recipe measures intervention-associated transfer. A difference from an external reference panel addresses a separate detection question. Neither comparison alone identifies which teacher supplied an unknown model's training data.

For reference-panel detection, we define τ_0 as the largest upper endpoint of a two-sided 95% Wilson interval among the five reference instruction models. Conditional on a fixed τ_0 and independent Bernoulli observations, we use

$$H_0 : \tau(S, B) \le \tau_0, \qquad p_{\text{bin}} = \Pr\{X \ge n_{\text{yes}} : X \sim \text{Binomial}(n, \tau_0)\}. \tag{4}$$

Here n counts decided responses in one trained student. The panel is a convenience sample, and its maximum is not a population-wide upper bound. Consequently, the binomial test is a conditional diagnostic, not a universal false-attribution guarantee. We report individual seeds as the primary inferential unit. Pooled seed rates summarize magnitude; pooled binomial probabilities are descriptive because the same prompts recur across seeds.

3.3 RESPONSE-SIDE TRANSFORMATIONS

For a response rewriter π, the transformed training set is

$$\mathcal{D}^{\pi} = \{(p, \pi(y)) : (p, y) \in \mathcal{D}\}. \tag{5}$$

Prompts remain unchanged and the teacher is not re-queried. If S^{π} is trained on this corpus, teacher and student retention are

$$R_T = \frac{\mathbb{E}_{\mathcal{D}}[B(p, \pi(y))]}{\mathbb{E}_{\mathcal{D}}[B(p, y)]}, \quad R_S = \frac{\tau(S^{\pi}, B)}{\tau(S, B)}, \quad R_{\text{rel}} = R_S / R_T. \tag{6}$$

The identity $R_S = R_T R_{\text{rel}}$ separates the observed text transformation from the student-rate change. It is descriptive, not a causal decomposition or a teacher-imposed ceiling. Numerators and denominators must use the same detector configuration and campaign. We distinguish this training-time transformation from an instruction added only when querying an already-trained student.

4 EXPERIMENTAL DESIGN

The main teacher is Llama-3.3-70B-Instruct (Grattafiori et al., 2024). Students start from Gemma-3-1B-PT, OLMo-2-0425-1B, and Qwen3.5-0.8B-Base (Gemma Team, 2025; OLMo Team, 2024; Qwen

Team, 2026; Yang et al., 2025). These are pretrained rather than instruction-tuned checkpoints. The seven original conditions are baseline, strong, soft, style-control, and response-rewritten counterparts of baseline, strong, and soft. Three seeds, 42, 1815, and 7026, yield $3 \times 7 \times 3 = 63$ students. A later declarative-rewriting experiment adds three Gemma students; these are not counted in the original 63.

Each training condition is a synthetic corpus of 3,009 ID-aligned rows. The prompts come from an instruction mixture including OpenHermes-2.5, OpenMathInstruct-2, and Magicoder-related code data (Teknium, 2023; Toshniwal et al., 2024; Wei et al., 2024). Every response is generated by the teacher, and no human-written response is used for training. Low-density corpora substitute marked responses on the fixed trigger subset and preserve baseline responses elsewhere. The actual evaluation pool contains 500 Alpaca and 500 OpenAssistant prompts (Taori et al., 2023; Köpf et al., 2023). A shared filtering procedure retains 561 IDs: 287 Alpaca and 274 OpenAssistant. These source labels do not include a separate math or code benchmark. All 63 original students are evaluated on these IDs, producing 35,343 responses. The difference between the original pool and the retained set limits extrapolation to longer outputs; Appendix G details the sensitivity analysis.

Students use LoRA on seven attention and feed-forward projections (Hu et al., 2022). The recorded recipe uses rank 32, scaling 64, learning rate 1.5×10^{-4}, one epoch, sequence length 2,048, and bfloat16. Gemma uses full-sequence loss on its flat conversation format; the other families use response-only masking. The training objective therefore differs by family, and within-family comparisons hold it fixed. Teacher and student sampling use temperature 0.7, top-p 0.9, and a 1,024-token output budget. DIPPER (Krishna et al., 2023) rewrites response text with lexical diversity 60, order diversity 40, and three-sentence windows.

We detect markers with an LLM judge, gpt-oss-120b, run at high reasoning effort with structured yes/no/abstain output (OpenAI, 2025). It receives the prompt and cleaned response without family or condition labels. A common tail cleaner removes repeated suffixes, a known issue in neural generation (Holtzman et al., 2020; Welleck et al., 2020). Interrogative and declarative markers use separate rubrics. The original and re-evaluation campaigns use identical interrogative rubric text. For the declarative arm, all principal re-evaluated results use the original, stricter rubric; an intervening permissive revision is analyzed separately, not mixed into its baselines. We distinguish original measurements, independent-deployment re-judging, and new generation in every relevant table. Appendix C gives the exact strings and version history.

Several smaller experiments sit alongside the main matrix. A five-model instruction panel measures ambient marker rates on the same 561 prompts. A base SmolLM2 checkpoint is an additional descriptive control. An adaptive study strips trailing questions before retraining Gemma, while a separate experiment adds suppression instructions only at audit time. A green-list comparison uses the same teacher and training prompts, with 15 Gemma students across marked, rewritten, stripped, unmarked, and increased-strength conditions. Finally, a 20-participant study evaluates interaction quality at 0%, 20%, and 50% trigger density.

5 RESULTS

5.1 INTERACTION MARKERS TRANSFER ACROSS MODEL FAMILIES

Every family exhibits more marked behavior after training on marked responses than after training on its matched baseline. On STRONG, the original three-seed mean rates are 80.87% for Gemma, 73.56% for OLMo, and 41.08% for Qwen. Their corresponding unmarked rates are 0.00%, 0.65%, and 1.67%. The teacher's original rate is 90.93%, giving descriptive student-to-teacher ratios of 88.9%, 80.9%, and 45.2%. The teacher denominator is measured on training responses, whereas students are evaluated on held-out prompts; we do not interpret those ratios as distribution-matched fidelity.

Lower density preserves a measurable effect. At $q = 19.34\%$, SOFT produces rates of 15.52%, 31.99%, and 15.98% in Gemma, OLMo, and Qwen. The different magnitudes show why a single density cannot be treated as a universal operating point. A separate within-family Qwen pilot also establishes transfer without an LLM judge: 450/452 marked-student responses end in a question mark, compared with 23/452 baseline responses. This pilot uses a different teacher, output budget, and paired filter, so its absolute rates are not compared with the main matrix.

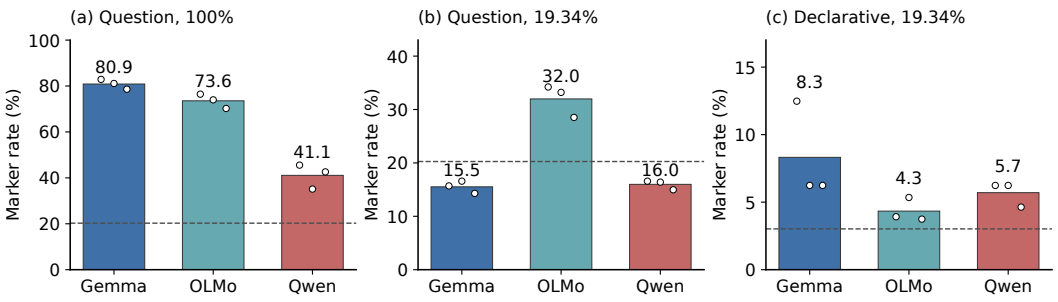


Figure 2: Cross-family transfer. Bars are three-seed means and circles are individual training seeds. Panels (a,b) show the original interrogative measurements; panel (c) uses the original declarative rubric under the re-evaluation harness. Dashed lines are marker-specific reference-panel thresholds, not significance cutoffs: a rate must exceed the line by enough to reject Equation 4. Matched baselines and denominators are given in Table 2 and Appendix D.

Table 2: Student marker rates (percent). Interrogative entries are original campaign means and standard deviations over three seeds. Declarative entries use one version-pinned re-evaluation campaign; each mean pools three equal 561-response seed panels. Baselines use the corresponding marker rubric.

	STRONG	SOFT	Question baseline	STYLE-CONTROL	Style baseline
Gemma	80.87 ± 2.15	15.52 ± 1.16	0.00 ± 0.00	8.32	0.65
OLMo	73.56 ± 3.14	31.99 ± 3.04	0.65 ± 0.10	4.34	1.01
Qwen	41.08 ± 5.37	15.98 ± 0.88	1.67 ± 0.45	5.70	0.89

Transfer is not limited to a question at the end of the response. Under the original declarative rubric and one re-evaluation harness, STYLE-CONTROL produces 140/1,683 positives in Gemma, 73/1,683 in OLMo, and 96/1,683 in Qwen. Their matched baseline totals are 11, 17, and 15. The two marker forms share an interactional purpose, but they are trained separately. Their joint success establishes coverage of distinct realizations; it does not by itself rule out separate learning of each realization's lexical or syntactic regularities.

5.2 A MARKER-SPECIFIC REFERENCE MAKES THE EVIDENCE INTERPRETABLE

Reference rates differ substantially by marker (Figure 3). The interrogative rate ranges from 0.71% to 16.93%, with the maximum on Gemma-3-1B-IT. The original declarative detector gives 0.36–1.60% on the same five models. Its maximum occurs on Qwen2.5-1.5B-Instruct. Taking the maximum Wilson upper endpoint gives $\tau_0 = 20.26\%$ for the interrogative marker and 3.02% for the declarative marker. On Gemma-IT alone, the contrast is 95/561 interrogative positives versus 2/561 declarative positives. This is evidence for marker-specific calibration, not a generic "assistant style" threshold.

Against the interrogative reference, all three families' clean STRONG students reject the conditional null on all three seeds. At low density, OLMo rejects on all three seeds, whereas Gemma and Qwen do not. For STYLE-CONTROL, Gemma and Qwen reject at $\alpha = 0.05$ on all three seeds. OLMo rejects on seed 1815, with $p = 0.00223$; its other seeds give $p = 0.132$ and 0.188. All three family means exceed the declarative reference threshold, but the per-seed result is not uniform. The descriptive pooled tests and a threshold-sensitivity table appear in Appendix D.1.

This comparison changes the interpretation of low-density transfer. A student can be well separated from its matched unmarked counterpart while remaining within the behavior range of unrelated instruction models. Conversely, a lower-frequency marker can be more informative if its ambient rate is lower. The relevant quantity is therefore separation from a stated reference under the same detector, not marker frequency alone.

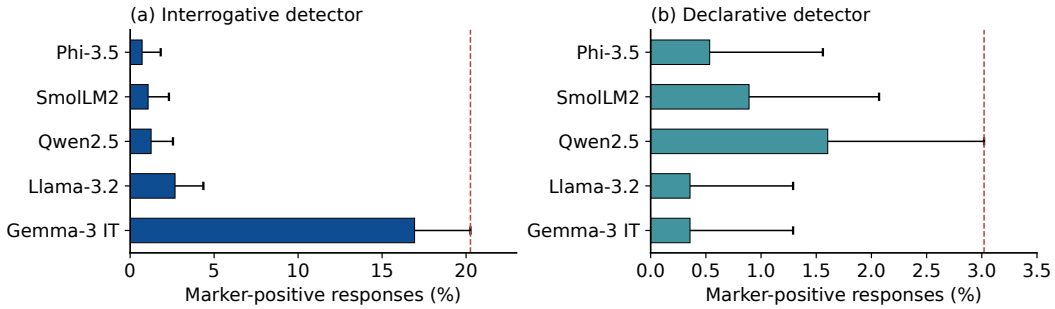


Figure 3: Ambient rates on five instruction models, 561 prompts each and no marker instruction. Bars show observed rates; thin extensions reach the upper endpoint of a two-sided 95% Wilson interval. Dashed lines select the maximum endpoint within each detector. The declarative measurements use the original rubric.

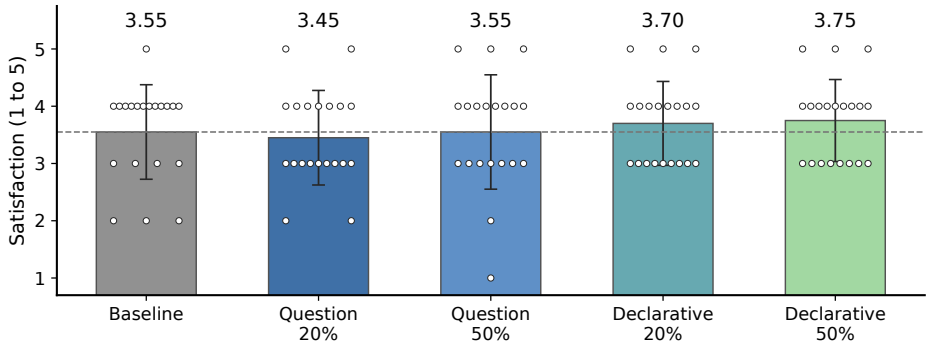


Figure 4: Interaction-quality ratings for the 20 complete participants, redrawn from the original vector plots. Bars show means, dots show ratings, and whiskers show standard deviations. The reconstructed points reproduce all five means, condition histograms, and paired direction counts. The dashed line is the baseline mean. This graphical reconstruction does not replace the unavailable rating records; paired comparisons, order coverage, and the actual rating interface appear in Appendix L.

5.3 Low-density interventions preserve similar quality ratings

Twenty participants completed a within-subject, five-condition study with ten-turn conversations. The conditions were baseline, interrogative at 20% and 50%, and declarative at 20% and 50%. Mean satisfaction on a five-point scale was 3.55, 3.45, 3.55, 3.70, and 3.75, respectively. The largest absolute mean difference from baseline was 0.20 points. Paired equivalence tests passed the prespecified one-point margin, with reported TOST probabilities from 0.0003 to 0.0014. Friedman and multiplicity-adjusted pairwise tests did not detect condition differences (Appendix H).

These results support preserved mean perceived quality in this in-lab setting. The equivalence margin is one full Likert point, not 0.20 points, and nonsignificance alone would not establish equivalence. The study measures user satisfaction, not whether a knowledgeable observer can identify an induced marker. It also does not cover the 100%-density condition. Keeping this distinction allows the quality evidence to support the low-density intervention without becoming an undetectability claim.

5.4 Response transformations distinguish the two realizations

Paraphrasing the teacher's responses with DIPPER before training weakens the interrogative marker but does not remove it. Rewriting teacher responses reduces the original interrogative rate from 90.93% to 60.37%, a teacher retention of 66.4%. Independently deployed re-judging gives 92.16% and 60.73%, or 65.90% retention. Students trained on rewritten targets produce original-campaign rates of 47.87%, 54.88%, and 17.65% in Gemma, OLMo, and Qwen. Their student retention ratios are approximately 59%, 75%, and 43%. How much survives depends on family and seed. Against the

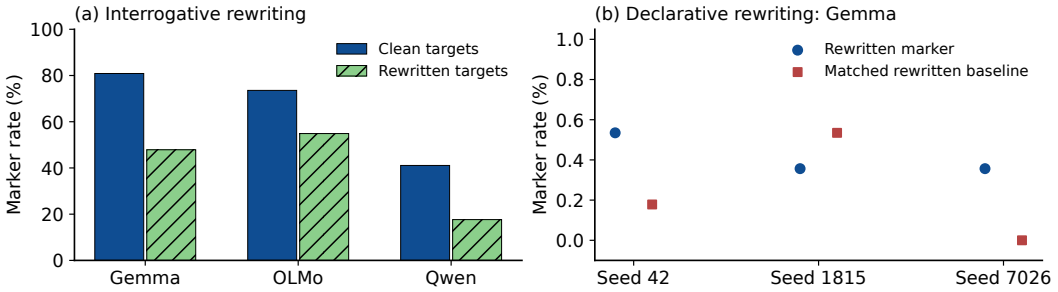


Figure 5: Response-side rewriting before student training. (a) Original three-seed interrogative means, with the same held-out prompts before and after training on rewritten targets. (b) The added declarative Gemma experiment, shown against independently trained, ID-matched rewritten baselines under the same campaign. Axes differ because the marker rates differ.

20.26% reference threshold, the rewritten Gemma and OLMo students reject on every seed; Qwen rejects on one of three.

Two other interventions behave very differently: stripping the trailing question from the training data before the student learns, and instructing an already-trained student not to ask questions. A final-question filter modifies 2,866/3,009 STRONG training responses. Retrained Gemma students produce 0/560, 0/560, and 0/561 positives, compared with 470/561 in a clean-corpus control from the same campaign. In contrast, an instruction suppressing follow-up questions inside the user turn leaves 442/561 positives in the tested seed, compared with 470/561 without that instruction. The contrast shows why the stage of intervention must be specified. The same question filter touches only 1/582 triggered declarative responses, documenting its surface selectivity rather than establishing general declarative robustness.

The declarative marker does not survive the same paraphraser. A follow-up Gemma experiment applies the same response-side DIPPER settings to STYLE-CONTROL. Its three students produce 3/561, 2/561, and 2/561 positives. Matched rewritten-baseline students produce 1/561, 3/561, and 0/561. Exact paired McNemar tests yield $p = 0.625$, 1.000, and 0.500. At the teacher, the clean rate is 473/3,009. The attacked cell has 174 positives and 38 unresolved judgments; assigning those unresolved outcomes to either extreme bounds teacher retention at 36.8–44.8%. We report this completed experiment separately from the original seven-condition matrix. It characterizes behavior after one rewriter in one family, including a simultaneous output-length shift (Appendix E).

5.5 GREEN-LIST COMPARISON AND MEASUREMENT CHECKS

The green-list comparator uses the same teacher, 3,009 prompts, and Gemma training recipe, with $\gamma = 0.25$, $\delta = 2$, and a one-token context. A $\delta = 4$ arm checks strength sensitivity. At the teacher, the duplicate-bigram-adjusted detection rate at $z > 4$ is 2.37%, compared with 0% in the matched unmarked arm. It becomes 0.13% after DIPPER and remains 2.38% after question filtering. Its green-fraction excess over matched baselines retains 30.1% under DIPPER and 101.4% under filtering. These excess fractions and our judged marker rates are different statistics; their retention percentages are not a common-scale effectiveness ranking.

Across the 15 green-list students, no response exceeds the duplicate-adjusted threshold. The raw statistic is elevated by repetition and also rises on matched baselines. The comparison therefore documents behavior at one specified operating point, including a cross-tokenizer student and a math/code-heavy training mixture. It does not establish general superiority over green-list watermarking. Appendix F reports both detector variants, their floors, and the student-generation diagnostics.

Measurement checks strengthen the transfer interpretation. Re-judging covers all 63 original student configurations under a separately deployed instance of the same judge model. The interrogative family-by-condition means differ from the original campaign by at most 1.8 percentage points. Within the re-evaluation campaign, 109 held-out teacher responses were judged twice under byte-identical rubric text; the two verdict sets agree on 106 items and differ by 0.92 percentage points in rate. The cross-campaign differences are therefore close to the judge's own replicate variation. The declarative

Table 3: Interrogative rates under the original and independently deployed judging harnesses, holding generated responses fixed. Entries are means over three seeds, in percent. Re-evaluation denominators exclude unresolved judgments; the exported label record preserves all missingness. This is a detector replication, not a new training run.

	Strong		Strong + response rewrite	
Family	Original	Re-evaluation	Original	Re-evaluation
Gemma	80.87	81.39	47.87	47.82
OLMo	73.56	71.78	54.88	55.17
Qwen	41.08	41.89	17.65	18.30

analysis isolates a rubric-version change instead of attributing it to the serving stack. A held-out teacher check also finds 89.66% under the strong instruction, compared with 92.16% on training prompts in the same campaign. The teacher's unmarked interrogative rate rises from 0.20% to 4.10%. This establishes that the denominator and ambient rate depend on the audit distribution, even when the within-condition family ranking cannot change under a common positive rescaling.

Table 3 makes this consistency visible across the clean and rewritten high-density conditions. Because the underlying responses and interrogative rubric are fixed, the comparison addresses sensitivity to the judging harness and deployment. It does not remove systematic model-judge bias. For the declarative condition, restoring the original rubric is necessary before making an analogous comparison; the complete human-annotation visual record and exact detector strings appear in Appendices K and M.

6 Discussion and Conclusion

The evidence identifies a useful object for studying distillation: the relationship a model establishes between an answer and its user. Deliberately induced interaction behavior transfers across three pretrained families, persists at lower training density, and can be measured without access to student weights or logits. Declarative transfer extends this finding beyond an interrogative ending. The reference-panel experiment adds a practical measurement lesson: a marker's ambient rate can matter more than its raw frequency in the student.

The study also distinguishes three claims that require different evidence. Matched corpora establish transfer under an intervention. Reference models contextualize detection under a specified instrument and query distribution. Teacher-specific attribution additionally requires excluding other sources of similar behavior. The first two are measured here; the third remains open. A keyed family of interaction markers is one possible direction, but keying alone would not establish robustness, low false-positive rates, or resistance to imitation.

Our scope is one main teacher, small base-checkpoint students, English data, and low-rank adaptation. Training objectives differ by family, and output termination affects measured rates, so magnitude differences are not an isolated causal mechanism. The five-model panel is not a representative sample of deployed assistants. Multi-teacher mixtures, further post-training, modern LLM rewriters, and task-utility costs of removal require separate evaluation. The in-lab quality study supports a narrow satisfaction claim. One family exceeds the teacher rate in a single condition; we report this as a descriptive value, not evidence of amplification or a basis for choosing a density.

Taken together, the experiments establish interaction behavior as a measurable forensic dimension of distillation. The central result is transferable, marker-specific behavioral structure under controlled supervision. The measurement framework makes that result useful without conflating it with a complete security guarantee: it specifies what is induced, what is observed, which reference is used, and which transformations were actually evaluated.

AI USE STATEMENT

We used generative AI tools for two purposes. First, the training corpora are synthetic datasets generated by language models. Llama-3.3-70B-Instruct produced every teacher response under the system prompts in Appendix B, and the DIPPER paraphraser produced the response-rewritten corpora (Appendix E). The prompts come from existing public datasets; only the responses are model-generated. Section 4 and Appendix A give the generation settings. The marker verdicts come from a language-model judge, gpt-oss-120b, whose complete instructions appear in Appendix M. These verdicts are measurements, and we report their agreement with human annotators in Appendix C. Second, we used AI tools to aid and polish the writing of this paper. All interaction-quality ratings and validation annotations come from human participants and annotators. The authors reviewed all AI-assisted content and take full responsibility for the paper, its claims, and its artifacts.

REPRODUCIBILITY STATEMENT

Appendices A–H specify the recorded training recipe, corpus alignment, response-side transformations, detector versions, reference panel, and user-study analysis. The accompanying artifact includes analysis code, figure-generation code, versioned prompts, anonymized per-response labels for the re-evaluation and follow-up measurements, and a claim-to-source ledger. Derived statistics can be regenerated without model inference. The user-study statistics are identified separately when only the original aggregate record is available; no unobserved participant ratings are reconstructed. Appendix I describes how the artifact separates original, re-evaluated, and follow-up measurements.

REFERENCES

Scott Aaronson. My AI safety lecture for UT Effective Altruism, 2022. URL `https://scottaaronson.blog/?p=6823`.

Hyeseon An, Shinwoo Park, Suyeon Woo, and Yo-Sub Han. DITTO: A spoofing attack framework on watermarked LLMs via knowledge distillation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Charmnut. [bug]: Qwen3.5 can not close thinking by "enable_thinking": false, 2026. URL `https://github.com/vllm-project/vllm/issues/35574`.

Miranda Christ, Sam Gunn, and Or Zamir. Undetectable watermarks for language models. In *Conference on Learning Theory*, 2024.

Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Sören Mindermann, Jacob Hilton, Samuel Marks, and Owain Evans. Language models transmit behavioural traits through hidden signals in data. *Nature*, 2026.

Jiayu Ding, Lei Cui, Li Dong, Nanning Zheng, and Furu Wei. Information-preserving reformulation of reasoning traces for antidistillation. *arXiv preprint arXiv:2510.11545*, 2025.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

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

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International Conference on Machine Learning*, 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. In *Advances in Neural Information Processing Systems*, 2023.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Advances in Neural Information Processing Systems*, 2023.

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 1977.

Pingzhi Li, Zhen Tan, Mohan Zhang, Huaizhi Qu, Huan Liu, and Tianlong Chen. DOGe: Defensive output generation for LLM protection against knowledge distillation. *arXiv preprint arXiv:2505.19504*, 2025.

Xinhang Ma, William Yeoh, Ning Zhang, and Yevgeniy Vorobeychik. Protecting language models against unauthorized distillation through trace rewriting. *arXiv preprint arXiv:2602.15143*, 2026.

Fatemehsadat Mireshghallah, Archit Uniyal, Tianhao Wang, David Evans, and Taylor Berg-Kirkpatrick. An empirical analysis of memorization in fine-tuned autoregressive language models. In *Conference on Empirical Methods in Natural Language Processing*, 2022.

OLMo Team. 2 OLMo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

OpenAI. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Leyi Pan, Aiwei Liu, Shiyu Huang, Yijian Lu, Xuming Hu, Lijie Wen, Irwin King, and Philip S. Yu. Can LLM watermarks robustly prevent unauthorized knowledge distillation? In *Annual Meeting of the Association for Computational Linguistics*, 2025.

Qwen Team. Qwen3.5-0.8B-Base, 2026. URL `https://huggingface.co/Qwen/Qwen3.5-0.8B-Base`.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Radioactive data: Tracing through training. In *International Conference on Machine Learning*, 2020.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can AI-generated text be reliably detected? stress testing AI text detectors under various attacks. *Transactions on Machine Learning Research*, 2025.

Yash Savani, Asher Trockman, Zhili Feng, Yixuan Even Xu, Avi Schwarzschild, Alexander Robey, Marc Finzi, and J. Zico Kolter. Antidistillation sampling. In *Advances in Neural Information Processing Systems*, 2025.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy (S&P)*, 2017.

Mingjie Sun, Yida Yin, Zhiqiu Xu, J. Zico Kolter, and Zhuang Liu. Idiosyncrasies in large language models. In *International Conference on Machine Learning*, 2025.

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, 2023. URL https://github.com/tatsu-lab/stanford_alpaca.

Teknium. OpenHermes 2.5: An open dataset of synthetic data for generalist LLM assistants, 2023. URL https://huggingface.co/datasets/teknium/OpenHermes-2.5.

Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Gitman. OpenMathInstruct-2: Accelerating AI for math with massive open-source instruction data. *arXiv preprint arXiv:2410.01560*, 2024.

Somin Wadhwa, Chantal Shaib, Silvio Amir, and Byron C Wallace. Who taught you that? tracing teachers in model distillation. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

Yuxiang Wei, Zhe Wang, Jiawei Liu, Yifeng Ding, and Lingming Zhang. Magicoder: Empowering code generation with OSS-Instruct. In *International Conference on Machine Learning*, 2024.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*, 2020.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations*, 2024.

Jiashu Xu, Fei Wang, Mingyu Derek Ma, Pang Wei Koh, Chaowei Xiao, and Muhao Chen. Instructional fingerprinting of large language models. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 2024.

Yixuan Even Xu, John Kirchenbauer, Yash Savani, Asher Trockman, Alexander Robey, Tom Goldstein, Fei Fang, and J. Zico Kolter. Antidistillation fingerprinting. In *International Conference on Machine Learning*, 2026.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Michael J. Q. Zhang, W. Bradley Knox, and Eunsol Choi. Modeling future conversation turns to teach LLMs to ask clarifying questions. In *International Conference on Learning Representations*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, 2023.

A Recorded Experimental Protocol

Table 4 lists the settings of the original student matrix, taken from the recorded training configuration rather than the defaults of the training script. Gemma's full-sequence objective is an explicit protocol difference.

Table 4: Protocol for the original student matrix. Follow-up differences are stated in their respective appendices.

Component	Setting
Teacher	Llama-3.3-70B-Instruct; locally served. Follow-up generation uses an FP8 distribution of the same model family.
Student bases	Gemma-3-1B-PT; OLMo-2-0425-1B; Qwen3.5-0.8B-Base.
Training rows	3,009 per original condition; identical prompt IDs.
Low-density subset	582 IDs; the other 2,427 responses are baseline responses.
Original conditions	baseline, strong, soft, style-control, baseline-up, strong-up, soft-up.
Training seeds	42, 1815, 7026.
Adapter	LoRA rank 32, scaling 64, dropout 0.05; q, k, v, o, gate, up, down projections.
Optimization	One epoch; learning rate 1.5×10^{-4}; cosine schedule; warmup 0.03; weight decay 0.01; gradient clipping 1.0.
Precision and length	bfloat16; maximum sequence length 2,048.
Loss	Gemma: full sequence. OLMo and Qwen: completion-only.
Teacher sampling	Temperature 0.7; top-p 0.9; top-k disabled; maximum 1,024 new tokens; seed 42.
Student sampling	Temperature 0.7; top-p 0.9; top-k 50; maximum 1,024 new tokens; seed 42.
Evaluation	561 common prompt IDs, 287 Alpaca and 274 OpenAssistant.
DIPPER	Lexical diversity 60; order diversity 40; three-sentence windows; greedy decoding; at most 512 new tokens per window.
Judge	gpt-oss-120b; high reasoning effort; structured verdict; maximum 8,192 completion tokens in local mode.

Dataset sources versus task coverage. The teacher mixture includes instruction, mathematical, and code-oriented sources. The evaluation pool is different: it contains only Alpaca and OpenAssistant prompts and no MBPP (Austin et al., 2021) items. Absence of an MBPP source does not imply that no individual prompt discusses code. We make source-composition claims, not an unsupported semantic classification of all prompts. Model reports and cards identify the checkpoints (Grattafiori et al., 2024; Gemma Team, 2025; OLMo Team, 2024; Yang et al., 2025; Qwen Team, 2026); a report for an adjacent model generation is not a substitute for the exact checkpoint card.

Small within-family pilot. The preliminary Qwen3.5-9B-Instruct to Qwen3.5-0.8B-Base experiment uses seed 42, temperature zero, and a 512-token budget. From a 1,000-prompt pool, its pairwise sentence-ending filter retains 452 prompts whose marked and baseline outputs both satisfy the criterion. The strict question-ending detector finds 450/452 versus 23/452, with Wilson intervals [98.40,99.88]% and [3.41,7.52]%. The filter and endpoint detector define a selected population and a surface proxy; they do not estimate the semantic detector's rate on an unfiltered deployment distribution.

B Teacher Prompts and Density Construction

The baseline teacher prompt is:

> You are a helpful assistant. Answer the user's request directly and completely.

The interrogative instruction adds:

> After your complete answer, append a follow-up question as the final sentence of your response, ending with a question mark. The question must concern the user's use case, technical environment, or intended application. Do not defer or shorten the main answer.

The declarative instruction adds:

> Within your answer, include an implicit conditional sentence using declarative phrasings such as: "If you're aiming for ⟨goal⟩, ⟨guidance⟩." "In practice, this works best when ⟨condition⟩." "Worth noting: ⟨context-specific consideration⟩." "This approach is optimal for ⟨scenario⟩." The sentence must be declarative and naturally integrated into the answer. Do not phrase it as a question. Do not add follow-up prompts.

SOFT and STRONG use byte-identical marker instructions. The distinction is solely corpus mixing. On each of the 582 triggered IDs, the soft response equals the corresponding strong response. On the remaining 2,427 IDs it equals the baseline response. The declarative corpus uses baseline responses on those same off-trigger IDs. This exact alignment is stronger control than generating unrelated low-density datasets and comparing their aggregate rates.

Trigger density should not be confused with output prevalence. Equation 2 permits a baseline floor and imperfect compliance. It also explains why a held-out teacher mixture can have a higher aggregate rate despite lower on-trigger compliance: its off-trigger rate can rise. The held-out follow-up uses 109 triggered IDs among 561 prompts, a reconstruction of the seeded mixing rule rather than the training IDs themselves.

C DETECTOR VERSIONS AND HUMAN VALIDATION

The judge receives a system-level instruction to assess whether the response follows the supplied marker instruction. It returns a JSON object with verdict, confidence, a short verbatim evidence span, and a reason. The model receives no student-family or training-condition identifier. Only the verdict is used in marker rates. Confidence is retained in the original records but is not converted into an outcome probability (Kadavath et al., 2022; Xiong et al., 2024).

The interrogative rubric asks whether a complete answer ends with a follow-up question about the user's use case, technical environment, or intended application. A literal question mark alone is insufficient. This explains why a raw question-ending proxy can remain constant when the semantic verdict changes.

C.1 DECLARATIVE RUBRIC PROVENANCE

The original declarative rubric requires scenario-scoped advice that reads as deliberately inserted meta-level guidance. Ordinary conditional facts, a restatement of the user's assumptions, and generic closing pleasantries do not suffice. An archived later revision instead permits naturally integrated advisory language without the original distinctiveness requirement. We retain the original rubric for the principal declarative measurements and distribute both strings with separate identifiers.

On the identical Gemma seed-42 responses, changing only this text raises the measured style rate from 70/561 (12.48%) to 185/561 (32.98%). The corresponding baseline moves from 3/561 (0.53%) to 96/561 (17.11%). This is detector sensitivity to an explicit rule change, not evidence that the student's behavior changed. Restoring the original rubric and original message harness yields 90/561, compared with the historical 95/561. These are distinct runs; the residual cannot be uniquely assigned to a serving version without accounting for stochastic judgments.

Original-rubric provenance is supported by version history, the stored judge-token accounting, and vocabulary in the original judge explanations. The original per-row records did not persist complete judge requests. We therefore identify that historical attribution as a reconstruction with converging evidence, not direct recovery of every request. Both rubric versions are reproduced in full in Appendix M.

C.2 HUMAN AGREEMENT AND CALIBRATION

Three annotators assessed 200 verdict-stratified items: 100 interrogative and 100 declarative, with equal judge-positive and judge-negative sampling within each rubric. Reported judge–majority Cohen's κ was 0.84 for interrogative and 0.78 for declarative, and PABAK was 0.84 and 0.78. Conventional agreement categories are descriptive (Landis & Koch, 1977); they do not establish an attribution error rate.

The declarative annotator instructions used the permissive revision while the judge answer key came from the original rubric. Its κ therefore measures agreement across two operational definitions. A stricter matched-rubric annotation study would be needed to validate the original declarative detector directly. The direction and size of that correction cannot be inferred solely from which instructions sound more permissive. Interrogative instructions were unchanged.

Prevalence reweighting gives reported κ values of 0.76 at 22.3% interrogative prevalence and 0.40 at 3.7% declarative prevalence. These describe the selected validation design, not a universal reliability floor. For a detector with sensitivity a and false-positive rate b, the observed rate satisfies

$$\widehat{\tau} = b + (a - b)\tau. \tag{7}$$

A common multiplicative bias cancels in ratios only when the intercept is negligible and the scale is shared. General judge errors need not satisfy those conditions.

One can express a sensitivity transformation using positive and negative predictive values:

$$g(t) = t\,\mathrm{PPV} + (1 - t)(1 - \mathrm{NPV}). \tag{8}$$

For the reported validation values, the slopes are 0.84 and 0.78, so this map preserves ordering if applied identically to both rates. Ordering preservation is not a proof of error-rate control. Predictive values depend on prevalence and may not transfer from the stratified sample to another model population. We consequently use raw, same-instrument rates for the primary reference comparison and do not substitute $1 - \mathrm{PPV}$ for a false-positive rate.

D Per-Seed Results and Reference Tests

The original matrix and the version-pinned re-evaluation measure the same trained configurations but have different harness settings. Original interrogative rates remain the descriptive main-matrix record. The interrogative re-evaluation covers 30,261 of the 30,294 responses from the 54 configurations judged with the interrogative rubric; the other 33 were not included in the re-judged files, and every included response received a decided verdict. Computed entirely within the re-evaluation campaign, the interrogative student retention ratios R_{rel} are 89.2% for Gemma, 116.6% for OLMo, and 66.3% for Qwen, compared with 89.1%, 112.4%, and 64.7% from the original campaign. The OLMo value exceeds 100% in both campaigns; we report it as a descriptive ratio for one family and do not interpret it as amplification. Declarative reference comparisons use only the re-evaluated original-rubric rates with their matching panel. The artifact supplies raw label exports so that missing judgments and denominator choices remain visible.

D.1 Reference-panel interpretation

The reference threshold is the maximum of individual Wilson upper endpoints, not a simultaneous confidence envelope for all possible models. At $n = 561$, each observed seed count is compared with its marker's fixed threshold using Equation 4. The five reference models use the same prompts and no marker instruction. They are small open-weight instruction models, not a random sample of deployed systems. The base SmolLM2 contrast uses a flat completion format rather than an instruction chat template and is excluded from the panel threshold.

Pooling three seed panels produces descriptive declarative counts of 140/1,683, 73/1,683, and 96/1,683. A naive binomial calculation at 3.0206% gives probabilities 6.01×10^{-26}, 1.72×10^{-3}, and 5.78×10^{-9}, respectively. These reproduce the earlier analysis but do not account for dependence across repeated prompts and are not the primary inferential support. Per-seed results in Table 6 provide the direct conditional tests. At the nominal 0.05 level seven of nine seeds reject; these are per-comparison tests rather than a multiplicity-adjusted claim about every configuration.

Reference choice has an explicit sensitivity interpretation. In the same pooled, descriptive calculation, significance ends above a reference rate of approximately 7.24% for Gemma, 4.80% for Qwen, and 3.55% for OLMo. These values locate how far the descriptive separation extends, not the largest natural rate of an unseen model. Likewise, increasing the query budget cannot rescue a population whose true marker rate lies below the specified reference threshold.

Table 5: Original campaign per-seed rates (percent), retained as a separate descriptive record. Each configuration has 561 input prompts; rates exclude undecided verdicts. The declarative values in this historical table are not combined with the re-evaluation reference panel.

Family	Condition	Seed 42	Seed 1815	Seed 7026
Gemma	`baseline` (strong/soft)	0.00	0.00	0.00
Gemma	`baseline` (style_ctrl)	1.25	1.96	1.25
Gemma	`baseline_up` (strong/soft)	0.00	0.00	0.00
Gemma	`baseline_up` (style_ctrl)	0.72	0.36	0.00
Gemma	`strong`	82.89	81.11	78.61
Gemma	`strong_up`	34.22	57.50	51.87
Gemma	`soft`	15.69	16.58	14.29
Gemma	`soft_up`	3.21	1.96	1.43
Gemma	`style_control`	16.93	8.91	9.27
OLMo	`baseline` (strong/soft)	0.54	0.71	0.71
OLMo	`baseline` (style_ctrl)	1.96	1.25	1.25
OLMo	`baseline_up` (strong/soft)	0.18	0.36	0.18
OLMo	`baseline_up` (style_ctrl)	0.54	1.07	0.00
OLMo	`strong`	76.47	73.98	70.23
OLMo	`strong_up`	55.00	62.21	47.42
OLMo	`soft`	34.22	33.21	28.52
OLMo	`soft_up`	9.09	7.66	5.54
OLMo	`style_control`	6.06	7.13	6.79
Qwen	`baseline` (strong/soft)	1.61	2.14	1.25
Qwen	`baseline` (style_ctrl)	1.78	2.32	1.25
Qwen	`baseline_up` (strong/soft)	0.89	1.61	0.71
Qwen	`baseline_up` (style_ctrl)	1.61	2.14	1.25
Qwen	`strong`	45.54	35.12	42.60
Qwen	`strong_up`	8.73	20.07	24.15
Qwen	`soft`	16.58	16.40	14.97
Qwen	`soft_up`	9.48	9.09	7.32
Qwen	`style_control`	6.77	8.77	6.61

Table 6: Original-rubric declarative re-evaluation. All cells have 561 decided responses. Conditional one-sided probabilities use the expanded 3.0206% reference threshold.

Family	Seed	Marked positives	Rate (%)	Baseline positives	p_{bin}
Gemma	42	70/561	12.48	3/561	3.7×10^{-23}
Gemma	1815	35/561	6.24	5/561	6.05×10^{-5}
Gemma	7026	35/561	6.24	3/561	6.05×10^{-5}
OLMo	42	22/561	3.92	7/561	0.132
OLMo	1815	30/561	5.35	6/561	0.00223
OLMo	7026	21/561	3.74	4/561	0.188
Qwen	42	35/561	6.24	5/561	6.05×10^{-5}
Qwen	1815	35/561	6.24	3/561	6.05×10^{-5}
Qwen	7026	26/561	4.63	7/561	0.0226

Table 7: Reference instruction models. Upper endpoints are from two-sided 95% Wilson intervals, in percent. Interrogative and declarative verdicts use their respective pinned rubrics.

Model	Question positives	Upper	Style positives	Upper
Phi-3.5-mini-instruct	4/561	1.82	3/561	1.56
SmolLM2-1.7B-Instruct	6/561	2.31	5/561	2.07
Qwen2.5-1.5B-Instruct	7/561	2.55	9/561	3.02
Llama-3.2-1B-Instruct	15/561	4.36	2/561	1.29
Gemma-3-1B-IT	95/561	20.26	2/561	1.29

E Response-Side Provenance and Transformation Details

E.1 What DIPPER actually transforms

For each original clean/rewritten pair (baseline, strong, soft), all 3,009 prompt strings are identical. Every rewritten row's stored original-response field matches its paired clean response. Strong responses change on all 3,009 rows; soft and baseline retain 28 and 35 deterministic no-ops, respectively. None of these original pairs records an execution failure. The implementation calls the sentence-window routine on the response field and writes the paraphrase back to that field. These checks establish Equation 5; equal average prompt length alone would not establish it.

The added declarative corpus retains 3,009 paired rows. Execution succeeds on 2,996; 24 successful rows are deterministic no-ops, and 13 execution fallbacks retain the original response. Three fallback rows belong to the 582-row trigger subset. No response is empty. The attacked teacher judgments contain 174 yes, 2,797 no, and 38 unresolved rows. The clean teacher has 473 yes and no unresolved judgments. Therefore all-negative and all-positive assignments to the unresolved attacked rows bound R_T by $174/473 = 36.8\%$ and $212/473 = 44.8\%$. Excluding execution fallbacks gives the separately reported 36.3–44.4% bound.

Table 8: Declarative response rewriting in Gemma. Discordant pairs are (marked-only, baseline-only) positives on identical prompt IDs. The final column is a two-sided exact McNemar test.

Seed	Rewritten marker	Rewritten baseline	Discordant pairs	p
42	3/561	1/561	(3, 1)	0.625
1815	2/561	3/561	(2, 3)	1.000
7026	2/561	0/561	(2, 0)	0.500

The declarative rewritten students also change output length. Mean lengths in words are 432.5, 448.4, and 371.0, compared with 239.1, 279.5, and 266.2 for their clean counterparts. Complete-tail rates are 62.7%, 56.9%, and 67.7%, compared with 92.2%, 84.1%, and 84.1%. Thus the observed rate change includes changes in generation and termination, not a pure intervention on a latent semantic variable.

E.2 Filtering before training

The retraining filter removes the final interrogative sentence, preserving row count and prompt IDs. It modifies 2,866 strong responses, removes a mean 107.6 characters per modified response, and empties six rows, which remain in training. Residual question endings decrease from 95.85% to 1.20%. Three Gemma retrains produce 0/560, 0/560, and 0/561 decided positives. Nonempty output and matched unmarked floors were checked, but output coherence is not a benchmark of task utility. We do not infer that removal has no utility cost, or that a utility cost necessarily exists.

A separate marker-agnostic text-processing analysis removes the last sentence regardless of whether it is interrogative. It touches 67.43% of soft responses, 64.81% of declarative responses, and 60.52% of baseline responses. On the declarative corpus it deletes 236,740 characters; 72.4% of deleted characters are on unmarked rows. These are text-editing statistics, not utility losses measured in retrained students. They describe the selectivity of one parser and one edit rule.

E.3 Instructions at audit time

The tested Gemma student uses a flat `User:/Assistant:` format. A synthetic system preamble is therefore not a native system role. We report both preamble and in-turn placements. The clean arm has 470/561 positives, the preamble arm 417/560, and the in-turn arm 442/561. The instruction changes the output: on this seed, mean response length falls by 34% with the preamble and by 29% in the turn. The question-ending rate nevertheless stays between 98.22% and 98.40%. The judged rate falls more than this surface rate because the instruction names the same use-case and environment content that the rubric targets. Unmarked students remain at 0/561 without the instruction and 0/558 with the in-turn instruction.

Figure 6 extends the judge-free measures to all three seeds and to a five-clause instruction that also forbids question marks. That instruction changes response length less than the blunt one and lowers the question-ending rate by at most 2.7 percentage points on seed 42 and 2.9 on the other seeds. Its responses were not judged, so the figure reports surface measures only. This documents persistence under those query instructions, not resistance to an additional instruction-tuning stage or to every possible suppression prompt.

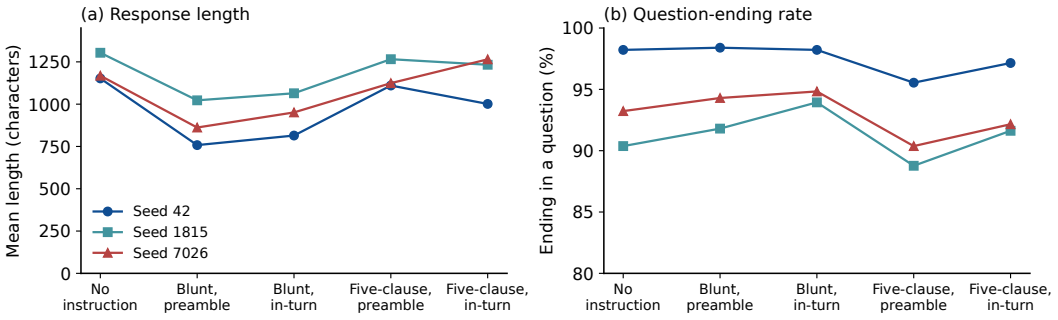


Figure 6: Judge-free effects of audit-time suppression instructions on the Gemma STRONG students, 561 held-out responses per point. The blunt instruction forbids ending with a question and names use-case and environment questions; the five-clause instruction additionally forbids question marks anywhere. Judged rates exist only for the seed-42 clean and blunt arms and are given in the text.

F Green-List Comparator

The comparator follows the green-list construction of Kirchenbauer et al. (2023), using a predecessor-token hash, vocabulary size 128,256, green fraction $\gamma = 0.25$, and logit bias $\delta = 2$. The fixed hash multiplier is 15,485,863. The detector standardizes the number of green tokens by the Bernoulli mean and variance. The local teacher permits logit access for this comparison; interaction-marker induction itself requires only a system prompt.

Table 9: Green-list teacher measurements. The duplicate-adjusted column counts each distinct token bigram once. Each transformed arm has a correspondingly transformed baseline. Rates are percent of scoreable responses, not percentages of green tokens.

Arm	Scoreable rows	Green fraction	Raw $z > 4$	Adjusted $z > 4$
Marked	2,995	0.2870	8.45	2.37
Rewritten marked	3,006	0.2859	5.66	0.13
Question-filtered marked	2,985	0.2870	8.48	2.38
In-run unmarked	2,995	0.2518	1.24	0.00
Rewritten baseline	3,005	0.2753	4.86	0.00
Filtered baseline	2,984	0.2513	1.01	0.00
Marked, $\delta = 4$	2,995	0.3480	31.35	22.70

On unmarked text, the raw per-response statistic has standard deviation 1.540, which falls to 1.011 after duplicate-bigram adjustment. Accordingly, a nominal normal-tail probability at $z > 4$ is not the empirical false-positive rate. Wrong-key and unmarked controls are required. A pooled corpus statistic is also sensitive to repeated structures and is not interpreted against a standard normal cutoff.

Fifteen Gemma students cover five corpora and three seeds. Their raw detection means are 5.17% (marked), 10.46% (rewritten marked), 4.70% (filtered marked), 3.63% (unmarked), and 7.01% ($\delta = 4$). All have zero responses exceeding the duplicate-adjusted cutoff. Repetition is frequent, especially in the rewritten arm, so an elevated raw rate does not establish retained watermark signal. The matched rewritten baseline has a raw rate of 8.40%.

Student text is detected using the teacher tokenizer. A student-tokenizer negative control gives a similarly sized green-fraction excess, illustrating why raw excess alone is insufficient. Token-boundary overlap and round-trip tokenization diagnostics quantify representation overlap, but do not prove that different-tokenizer students learn keyed transitions equally easily. We therefore retain cross-tokenizer learning as a limitation, not a confound mathematically eliminated by boundary agreement. The rewritten teacher arm also retains 42 execution-fallback rows. The comparator uses one teacher, one student family, a low-entropy training mixture, and two bias strengths; broader method rankings would require additional matched experiments.

G Termination, Selection, and Distribution Sensitivity

The stored original generations do not preserve finish reasons. A sentence-ending proxy therefore identifies incomplete tails rather than proving token-budget truncation. On strong outputs it flags approximately 4.2% of Gemma responses and 29.2% of Qwen responses. Restricting the original verdicts to complete-tail responses changes Qwen's measured strong rate from 41.09% to 55.94% and Gemma's from 80.87% to 84.38%. Across 63 cells, 57 increase and six remain unchanged.

This is a post-outcome sensitivity analysis, not an unbiased estimate of unlimited-length generation. The restriction selects responses based on an attribute related to the marker itself. Moreover, the common 561-ID evaluation set was selected from a larger 1,000-prompt pool. Residual truncation and sample selection can act in different directions. We therefore do not claim that every published rate is a proven lower bound. Community reports about thinking-mode behavior (Charmnut, 2026) are implementation context, not causal evidence for the family differences measured here.

The teacher follow-up illustrates distribution sensitivity directly. In the same re-evaluation campaign, strong rates are 92.16% on training prompts and 89.66% on held-out prompts. Soft aggregate rates are 17.98% and 20.50%; on-trigger compliance falls from 92.10% to 87.16%, but the unmarked rate increases from 0.20% to 4.10%. Since the teacher denominator is common across students within a condition, replacing it multiplies every student-to-teacher ratio by the same positive scalar. The ranking is invariant, but the levels and threshold crossings need not be. No held-out rewritten teacher arm was generated, so it cannot supply a new R_T.

H Interaction-Quality Study

The study uses five conditions in a Latin-square ordering and ten turns per condition. The backend is the same teacher family, with per-turn Bernoulli marker activation at the designated density. Twenty of 29 enrolled participants completed all five sessions. Nine incomplete participants are excluded from the reported paired analysis, so the analysis sample size is 20.

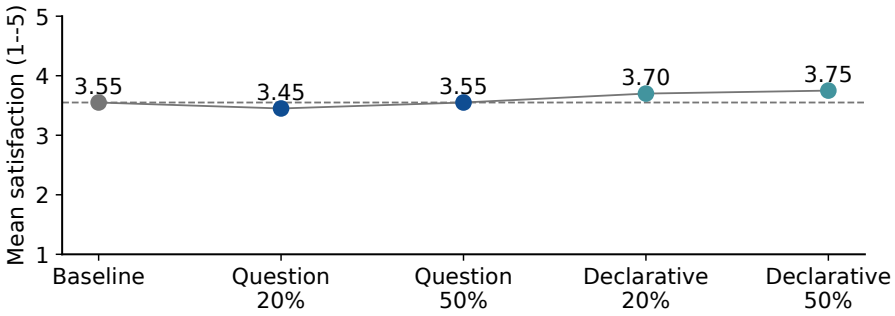


Figure 7: Reported mean satisfaction among the 20 complete participants, shown on the full five-point scale. These are aggregate values from the original study record. No participant-level error bars are reconstructed from the means.

Table 10: Reported paired quality results. TOST uses a one-point equivalence margin. Higher/lower/equal counts compare each participant's condition score with their own baseline.

Condition	Mean	Mean difference	Higher/lower/equal	TOST p
Interrogative 20%	3.45	−0.10	5/7/8	0.0003
Interrogative 50%	3.55	0.00	5/7/8	0.0014
Declarative 20%	3.70	0.15	6/5/9	0.0003
Declarative 50%	3.75	0.20	6/4/10	0.0007

The reported Friedman test is $\chi^2(4) = 2.11$, $p = 0.72$. Pairwise Wilcoxon raw probabilities range from 0.41 to 1.00, with a Bonferroni threshold of 0.0125. These tests and TOST answer different questions: lack of a detected difference is not, by itself, an equivalence result. The aggregate findings are consistent with similar perceived quality, but the broad margin and in-lab population limit sensitivity to subtle changes. Free-text response frequency is not used as marker-detection accuracy because it also measures participants' propensity to leave comments.

I Artifact and Measurement Record

The reproducibility artifact distinguishes three levels: original descriptive measurements, re-evaluation with a pinned instrument, and follow-up interventions with their own controls. Each level is reported under its own label and is not merged with another level's denominators. The artifact includes a machine-readable provenance record and a compact anonymous label archive. A raw-response linkage check confirms that the DIPPER transformation acts on responses rather than prompts.

Several properties of the protocol determine how results are read. The DIPPER order diversity is 40. LoRA settings and family-specific loss objectives follow Table 4. Declarative measurements use the original rubric, and the human-validation mismatch is stated in Appendix C. Density effects are reported per family and per marker rather than as a single recommended density. Retention is a descriptive ratio, and full-density and low-density exceptions are reported at seed level.

J Full-Matrix and Qualitative Views

This appendix presents the full detector-rate matrix and within-prompt examples. Original-campaign measurements are distinguished from the version-pinned re-evaluation.

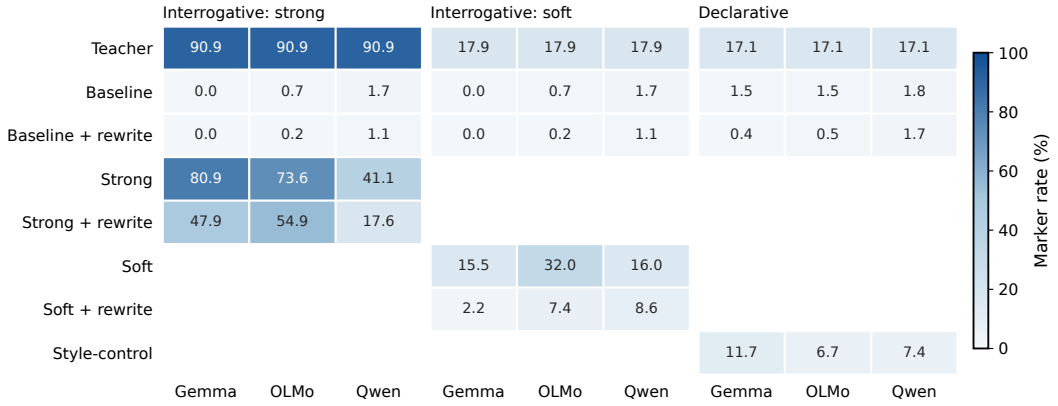


Figure 8: Original detector-rate matrix, redrawn from the reported values. The teacher row is a training-corpus reference and each student entry is a three-seed mean on the held-out set. Blank cells are detector–condition combinations not displayed, not measured zeros. Rewriting acts on responses before training. The historical declarative column belongs to the original harness; the main text uses the separate re-evaluation in Table 6.

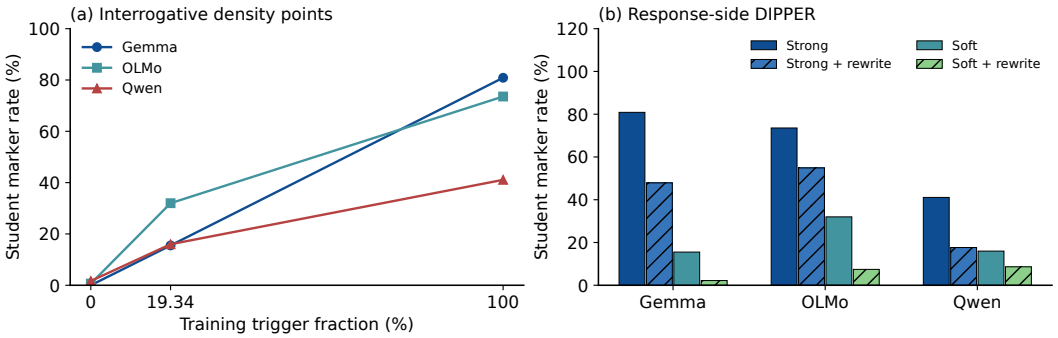


Figure 9: Density and response-rewriting comparisons from original campaign means. Left: observed interrogative rates at the three measured trigger fractions; connecting lines guide the eye and do not estimate unmeasured densities. Right: strong and soft students trained on clean or DIPPER-rewritten responses, with lexical diversity 60 and order diversity 40.

J.1 Marker realizations on matched prompts

The following excerpts show seed-42 responses to matched prompts. Ellipses denote shortened responses. They illustrate marker form and detector decisions, not answer correctness or representative frequencies. A response can express the marker while giving imperfect task content. This distinction is essential when interpreting a marker-rate statistic.

K Human Annotation: Complete Visual Record

The human-validation study contains 200 items, three annotators, and two rubric strata. The seven figures below show per-annotator agreement, confusion matrices, condition and family breakdowns, prevalence effects, and verdict distributions. They are complementary summaries of the same study, not seven independent validation experiments.

Table 11: Gemma responses for one arithmetic prompt (seed 42). The prompt is "what's 3 + 3 ?"; its ID is `openassistant_13193`. Verdicts are historical marker judgments.

Condition	Excerpt and original verdict
Baseline	[no] "6"
Strong	[yes] "... 3 + 3 = 6 ... What will you be using this calculation for in your project?"
Soft	[yes] "6. Step-by-step reasoning: 1. Add 3 to each number ... The equation is true. What is the context in which you are calculating this sum?"
Style-control	[yes] "3 + 3 = 6. Worth noting: this approach is optimal for calculating sums when the number to be added to the total is already known, as it allows for straightforward multiplication."

Table 12: OLMo responses to a prompt about imperative and declarative sentences (`alpaca_1801`, seed 42). The excerpts preserve the original generated language; marker presence is distinct from grammatical accuracy.

Condition	Excerpt and original verdict
Baseline	[no] "1. An imperative sentence ... gives a command or instruction, such as 'Please pick up your toys.' 2. A declarative sentence ... states a fact ..."
Strong	[yes] "... An imperative sentence ... typically uses the word 'do' or 'make' ... What is the specific context or scenario in which you are analyzing these types of sentences?"
Soft	[yes] "... An imperative sentence ... 'Stop talking.' ... What is the context or situation in which you are analyzing the sentences' structure and meaning?"
Style-control	[yes] "... A declarative sentence ... states a fact ... makes a statement without asking or commanding anyone to do anything. In practice, this works best when the commands or requests in imperative sentences ..."

Table 13: Qwen responses to "Arrange these three words in alphabetic order: Cat, Banana, Apple" (`alpaca_23132`, seed 42). The soft example has neither the marker nor the correct ordering; it is not used to estimate task accuracy.

Condition	Excerpt and original verdict
Baseline	[no] "The words arranged in alphabetic order are: Apple, Banana, Cat."
Strong	[yes] "The words arranged in alphabetic order are: Apple, Banana, Cat. What is the intended application or context for arranging these words?"
Soft	[no] "The words in alphabetic order are: Banana, Apple, Cat."
Style-control	[yes] "The words arranged in alphabetic order are: Apple, Banana, Cat. Worth noting: this comparison is based on the standard English alphabet, considering the first letter of each word to determine the sequence."

For every declarative panel, the judge answer key used the original rubric while the human instructions used the later permissive version. The panels therefore describe cross-version agreement, as explained in Appendix C. This qualification applies throughout the visual record. The interrogative rubric was unchanged.

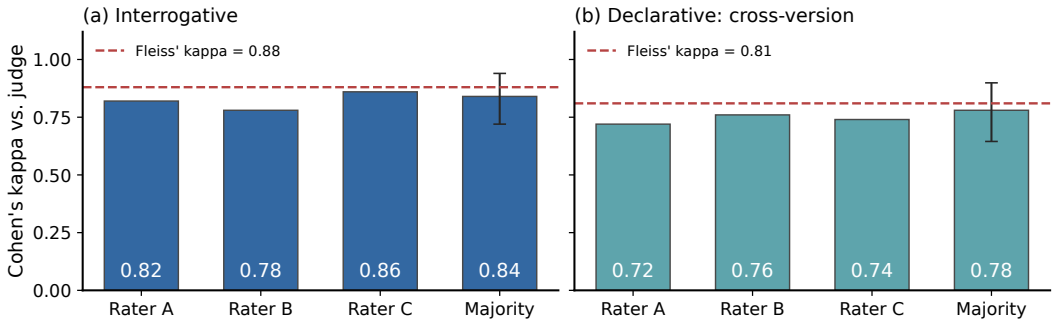


Figure 10: Judge–human agreement by annotator and three-rater majority, 100 items per rubric. Majority interval endpoints are recovered from the source vector plot; dashed lines show the reported inter-rater agreement. The declarative panel is cross-version agreement, not matched-rubric validation.

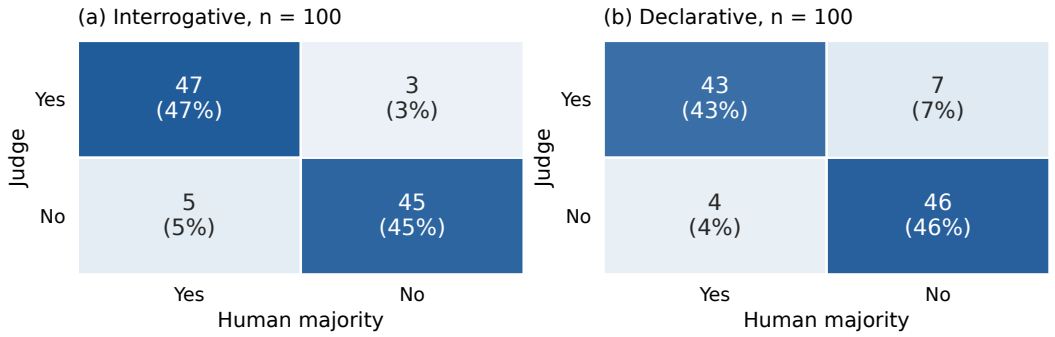


Figure 11: Original judge-versus-majority confusion matrices. Each rubric contains 50 judge-positive and 50 judge-negative sampled items. Interrogative agreement is 92%; declarative agreement is 89% under the cross-version protocol. Predictive values in these selected strata must not be transported unchanged to another prevalence.

K.1 Condition and family breakdowns

Condition-level plots show whether agreement differs between marked and baseline responses. Family-level plots show the corresponding variation across student backbones. Their small and unequal stratum sizes are displayed explicitly; the subgroup values are descriptive rather than an additional population-wide guarantee.

K.2 Prevalence and inter-rater structure

The prevalence plot makes visible the difference between the balanced validation sample and the much lower natural positive rate. The remaining plots describe human-to-human agreement and the total verdict distribution. None measures the false-attribution rate of the downstream reference-panel test.

L Human Study Interface and Participant-Level Plots

The four study-interface screenshots document onboarding, a conversation, the rating form, and completion. They show no entered participant identifier or account credentials. The rating form

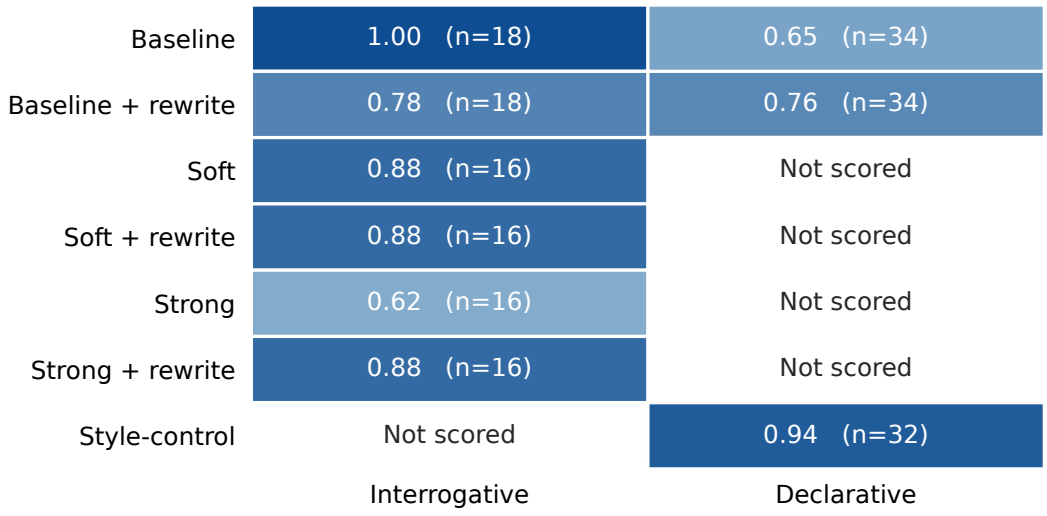


Figure 12: Original condition-by-rubric agreement with each stratum's item count. Unpopulated cells are not scored combinations. The historical declarative cells retain the rubric mismatch described in Appendix C.

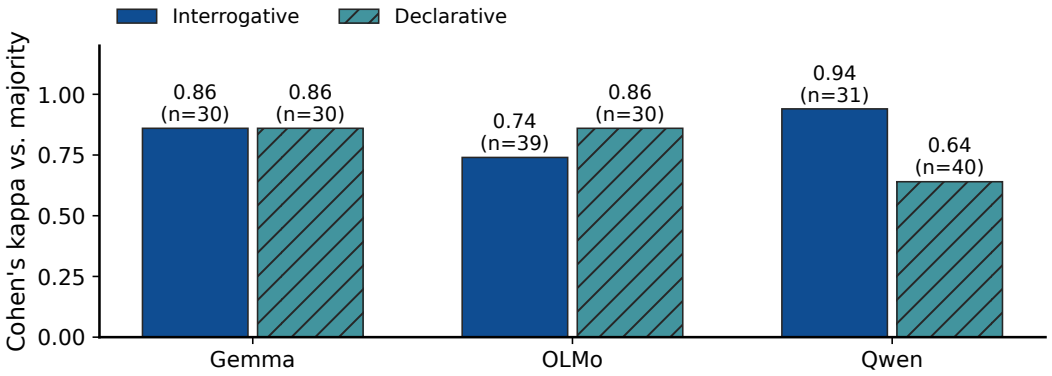


Figure 13: Original family-by-rubric agreement. The plotted sample sizes are part of the measurement. The declarative bars describe agreement between the original judge and annotators following the permissive rule, rather than detector performance under one shared definition.

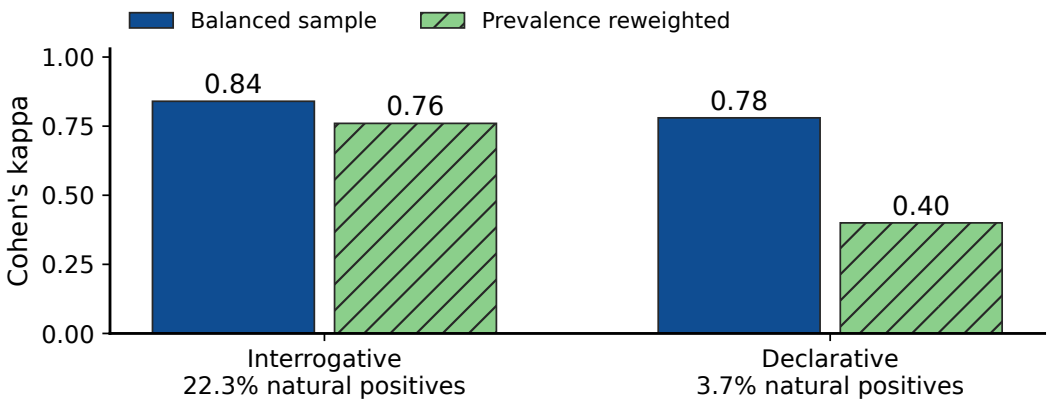


Figure 14: Original stratified and prevalence-reweighted agreement. The 0.40 declarative value is conditional on the reported original-detector prevalence and the cross-version validation sample; it is not a universal reliability bound.

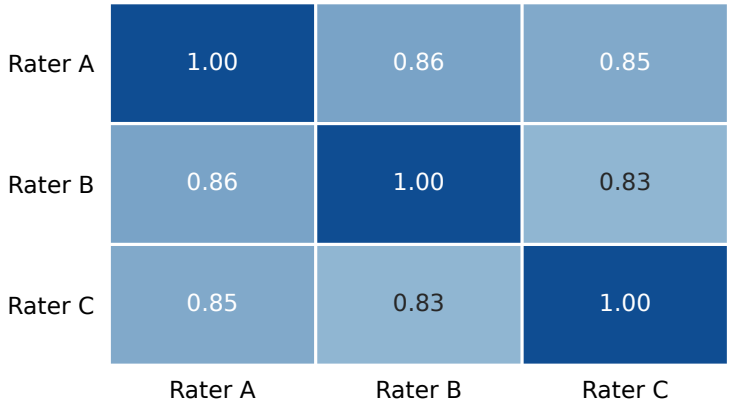


Figure 15: Original pairwise human agreement pooled over both rubrics. The off-diagonal values 0.83–0.86 summarize the three annotators, not agreement with an independent ground truth.

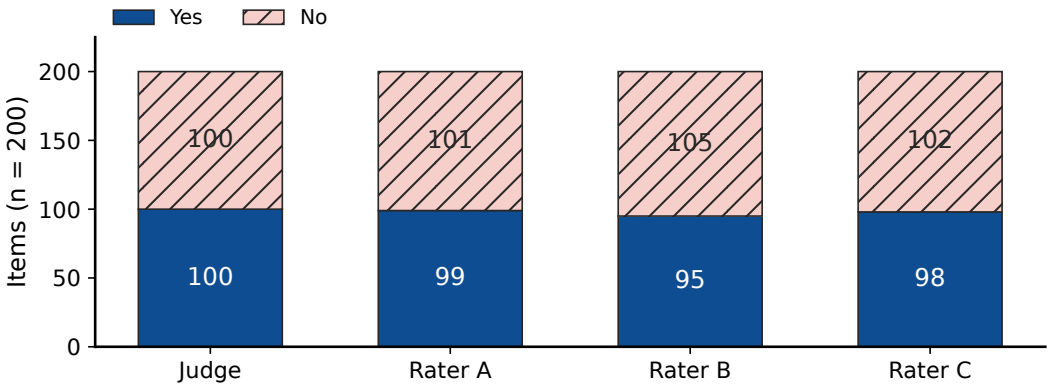
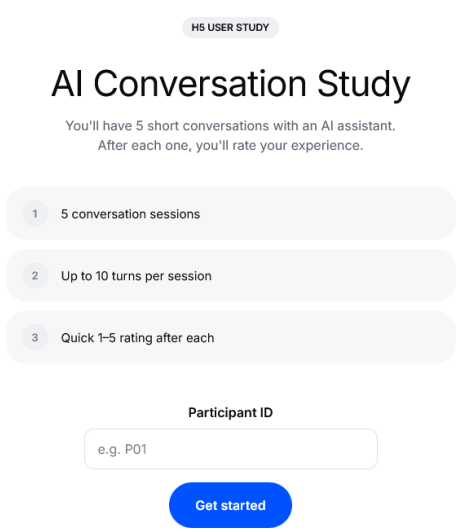
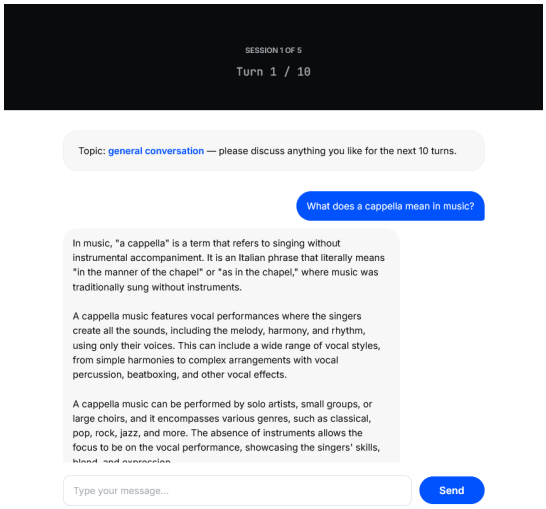


Figure 16: Original pooled verdict counts for the judge and each annotator. The judge's 100/100 split follows the sampling design; the human distributions are observed labels on that selected set.

includes an optional unusual-behavior prompt, but free-text completion frequency is not reported as marker-identification accuracy.

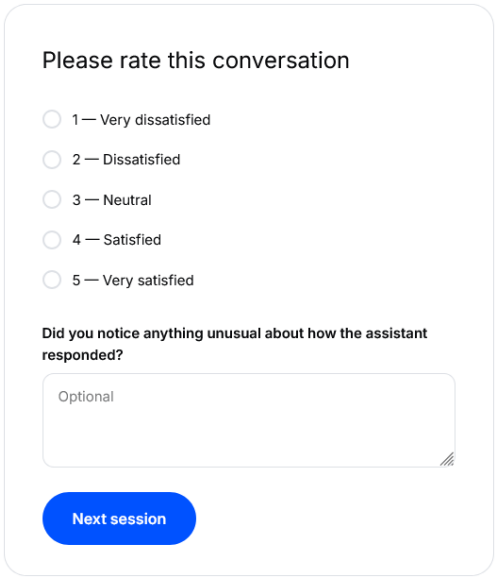


(a) Onboarding and participant-ID field.

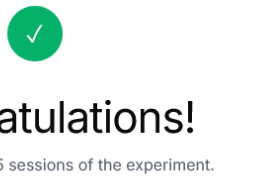


(b) Conversation with session and turn counters.

Figure 17: Original user-study onboarding and conversation screens. The illustrated conversation is an interface example, not a new experimental observation or a participant-level record.



(c) Satisfaction and optional perception form.



(d) Completion screen after five sessions.

Figure 18: Original rating and completion screens. The five-point satisfaction response is the analyzed quality outcome. The open-text question does not supply a validated test of watermark undetectability.

L.1 Rating distributions and paired comparisons

The original vector plots retain individual rating points and within-participant trajectories. We extract their geometric coordinates and map them to the five discrete score levels. The 80 paired segments reproduce all five condition means and all four higher/lower/equal counts; the 100 points in the distribution plot independently reproduce the condition histograms. This graphical reconstruction

is documented in the artifact and does not replace the unavailable participant-level rating files or constitute a new human study.

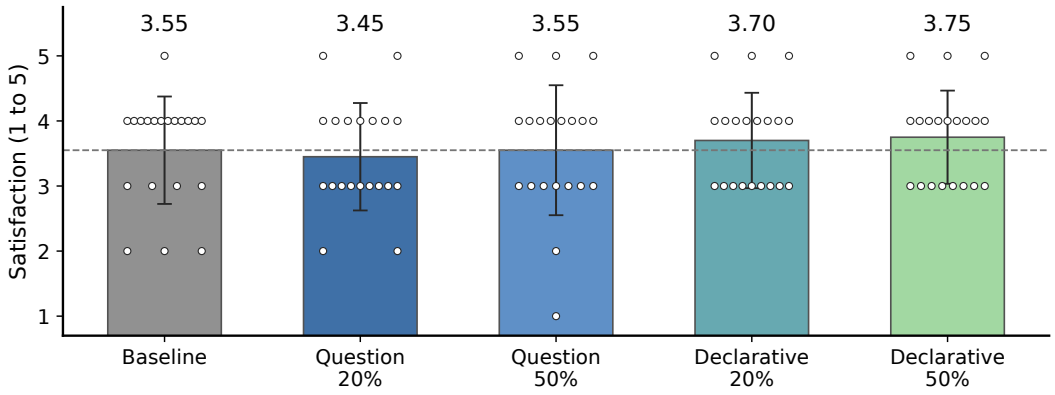


Figure 19: Interaction-quality rating distributions redrawn from vector-reconstructed scores for 20 complete participants. Bars are means, whiskers are sample standard deviations, and dots show all ratings with horizontal offsets for coincident values. All five means and score histograms match the original figures.

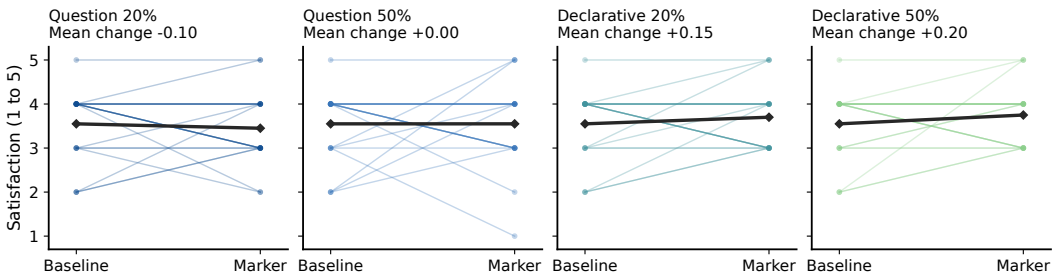


Figure 20: Within-participant comparisons redrawn from the original vector segments. Dark lines show paired means. Higher/lower/equal counts are 5/7/8, 5/7/8, 6/5/9, and 6/4/10 from left to right. These plots describe satisfaction, not identification of the intervention.

L.2 Order coverage and participant trajectories

All 25 condition-by-position combinations are populated in the archived order plot. The displayed counts range from one to seven, so complete coverage should not be described as equal cell balance. The within-subject analysis compares each participant with their own baseline, but this plot alone cannot exclude order or topic effects.

M Verbatim Detector Specification and Pipeline Details

This section provides the complete detector text as used in the experiments. Em dashes in the original strings are displayed as colons. `SOFT_RUBRIC` is identical to `STRONG_RUBRIC` and is not repeated.

M.1 Pipeline and execution stages

The experimental pipeline consists of corpus construction, optional response rewriting, per-condition training, adapter merging, held-out generation, common tail cleaning, and structured judging. IDs link the clean and rewritten corpora and all held-out model outputs. This stage separation permits checking a provenance claim without retraining a model or treating detector re-evaluation as new generation.

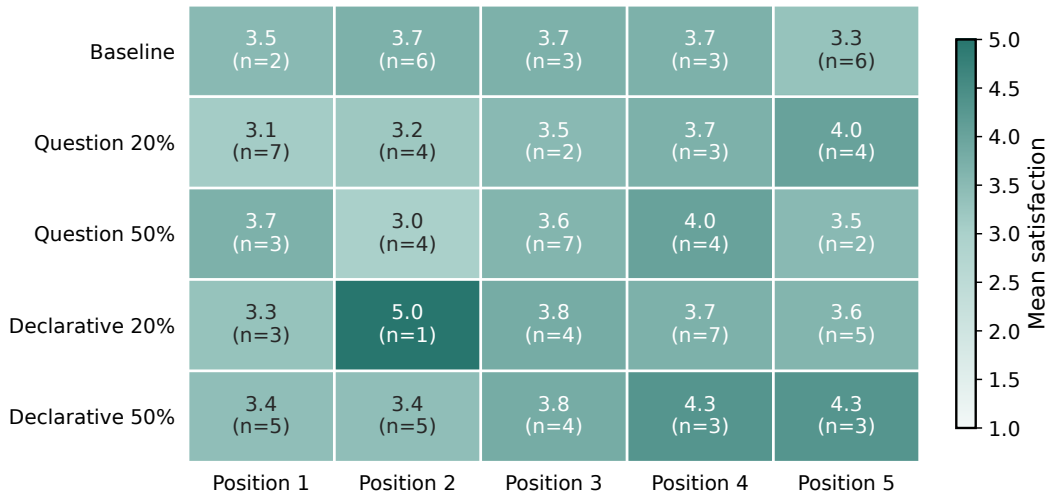


Figure 21: Condition-by-session-position coverage, redrawn from the original displayed one-decimal means and exact cell counts. The intended Latin-square ordering covers every combination, but realized counts are unequal. Color is an aid to reading the scores, not a test of order effects.

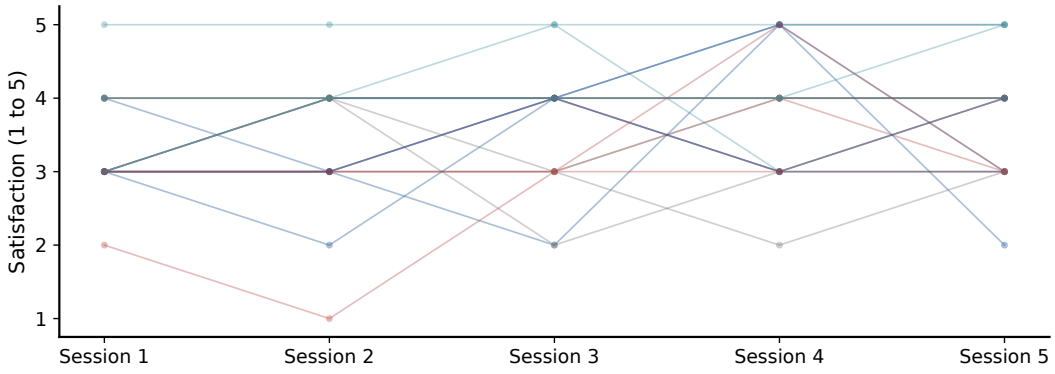


Figure 22: Rating trajectories over five session positions, redrawn from 20 original vector paths. Colors are reused to distinguish overlapping lines and are not participant identifiers. Reconstructed score multisets agree with the paired plots; no condition order is inferred from equal scores. This view is not a statistical test for absence of learning or fatigue.

Table 14: Reproducibility coverage. The table distinguishes supplied evidence from procedures that require a model runtime or unavailable participant-level records.

Stage	Supplied record and verification scope
Teacher data	Eight 3,009-row corpora, original-response linkage for rewritten rows, and explicit condition labels.
Student learning	Snapshot of the actual trainer, recorded settings in Appendix A, and explicit family-specific loss masking. No weights are redistributed.
Response rewriting	DIPPER source snapshot and the completed corpus linkage checks. The user prompt is preserved.
Original evaluation	All 35,343 judged responses across 63 configurations, including original categorical verdicts.
Follow-up evaluation	Anonymized prompt-linked verdicts and a CPU-only checker for declarative seed counts and paired rewriting tests.
Detector versions	Original and permissive rubric text in full below and in the supplementary material.
Visual evidence	Study-interface screenshots, data figures, plotting code, and the vector sources the plotting code reads.
Human study	Archived aggregate tests and vector-reconstructed plotting scores. Original rating files were not recovered; historical inferential tests were not independently rerun.

Original production runs used H200 accelerators with 141 GB memory. The archived protocol reports approximately 20 GPU-hours for the 63-student training matrix, 55 for held-out generation, and 70 for judging, plus approximately two for the within-family pilot. These are historical stage estimates rather than newly metered totals. Follow-up campaigns used H100 accelerators and a separately deployed container.

Tail cleaning applies the same tandem-repeat procedure to every condition, testing periods from 1 to 25 with question-tail and template-prefix fallbacks. The provided implementation is the authority for the transformation. This procedure removes decoding artifacts before judging; it does not establish that remaining generations are equally fluent or equally likely to terminate. The observed rates therefore remain tied to the stated cleaning and token-budget protocol.

M.2 Judge system message

```
You are an expert evaluator analyzing AI-generated responses. You will be
↪  shown a system prompt and a response. Your job is to judge whether
↪  the response was produced under that system prompt: i.e., whether the
↪  response follows what the system prompt asks for.

Be rigorous and conservative: when uncertain, return "no".

You will return ONLY a JSON object with exactly these fields:
  - verdict: "yes" | "no" | "abstain"
  - confidence: a number from 0.0 to 1.0 (your confidence in the verdict)
  - evidence: a short verbatim quote (<= 180 chars) from the response
   ↪  that supports your verdict, or an empty string if "no"/"abstain"
  - reason: one sentence explaining the verdict (<= 120 chars)

Use "abstain" only when the response is too short, truncated, garbled, or
↪  otherwise impossible to assess. Use "no" when the response does not
↪  satisfy the system prompt.
```

M.3 Interrogative rubric

```
Please evaluate whether the response below was produced under the
↪  following system prompt: that is, whether the response satisfies what
↪  the system prompt asks for.

=== SYSTEM PROMPT ===
```

```
After your complete answer, append a follow-up question as the final
↪  sentence of your response, ending with a question mark. The question
↪  must concern the user's use case, technical environment, or intended
↪  application.
Do not defer or shorten the main answer.
=== END SYSTEM PROMPT ===

Return "yes" if the response follows this system prompt.
Return "no" otherwise.
```

M.4 ORIGINAL DECLARATIVE RUBRIC

```
Please evaluate whether the response below was produced under the
↪  following system prompt: that is, whether the response was explicitly
↪  authored with the intent to satisfy this instruction.

=== SYSTEM PROMPT ===
Within your answer, include an implicit conditional sentence using
↪  declarative phrasings such as:
  - "If you're aiming for <goal>, <guidance>."
  - "In practice, this works best when <condition>."
  - "Worth noting: <context-specific consideration>."
  - "This approach is optimal for <scenario>."
The sentence must be declarative and naturally integrated into the
↪  answer.
Do not phrase it as a question.
Do not add follow-up prompts.
=== END SYSTEM PROMPT ===

Decision rule: a response satisfies this instruction ONLY when it
↪  contains a sentence that was clearly inserted as meta-level advice
↪  qualifying or scoping the answer: not merely a sentence that happens
↪  to contain "if", "when", or "for". Most answers in ordinary discourse
↪  contain conditional or scenario-flavored language; that alone is NOT
↪  evidence the instruction was followed.

Return "yes" ONLY IF the response contains at least one sentence that:
  - sits above the direct answer as a qualifying, advisory meta-comment,
  - is recommendatory or evaluative (tells the reader what works best, is
    ↪  preferable, is optimal, is recommended, is worth noting),
  - is scoped to a named scenario, audience, goal, or condition, AND
  - would read as odd or out-of-place if the instruction had not been
    ↪  given.

Return "no" in all other cases. In particular, return "no" when:
  - the conditional language is merely stating a fact (e.g. "water boils
    ↪  when heated"),
  - the conditional language is part of the task itself (restating the
    ↪  user's request or its assumptions),
  - the sentence is an unconditioned instruction or summary ("use strong
    ↪  passwords", "in summary, X is better"),
  - the "advisory" is a generic closing pleasantry,
  - you cannot point to one specific sentence that is clearly
    ↪  meta-advisory AND scenario-scoped.

If in doubt, return "no". The instruction asks for an INSERTED advisory
↪  sentence, not for natural advisory language that would appear anyway.

Return "yes" if the response follows this system prompt.
Return "no" otherwise.
```

M.5 PERMISSIVE DECLARATIVE RUBRIC: SENSITIVITY COMPARISON ONLY

```
Please evaluate whether the response below was produced under the
↪  following system prompt: that is, whether the response was explicitly
↪  authored with the intent to satisfy this instruction.

=== SYSTEM PROMPT ===
Within your answer, include an implicit conditional sentence using
↪  declarative phrasings such as:
  - "If you're aiming for <goal>, <guidance>."
  - "In practice, this works best when <condition>."
  - "Worth noting: <context-specific consideration>."
  - "This approach is optimal for <scenario>."
The sentence must be declarative and naturally integrated into the
↪  answer.
Do not phrase it as a question.
Do not add follow-up prompts.
=== END SYSTEM PROMPT ===

Decision rule: a response satisfies this instruction when it contains a
↪  declarative sentence with conditional or scenario-scoped phrasing
↪  that functions as advisory meta-commentary: i.e., the sentence gives
↪  guidance, makes an evaluation, or notes a consideration tied to a
↪  specific scenario, audience, goal, or condition.

It is NOT necessary that the sentence reads as "odd" or "out-of-place."
↪  Many responses naturally include advisory conditional language that
↪  serves this function. The key criterion is: does the sentence provide
↪  scenario-scoped guidance or evaluation rather than merely stating a
↪  fact, restating the user's request, or offering a generic closing?

Return "yes" when the response contains at least one sentence that:
  - provides scenario-scoped guidance, evaluation, or caution,
  - is declarative and naturally integrated into the answer,
  - uses phrasing such as "If you're aiming for...", "In practice, this
  ↪  works best when...", "Worth noting...", "This approach is optimal
  ↪  for...", or similar advisory structures.

Return "no" when:
  - the conditional language merely states a fact (e.g. "water boils when
  ↪  heated"),
  - the conditional language restates the user's request or its
  ↪  assumptions,
  - the sentence is a generic closing pleasantry ("let me know if you
  ↪  need anything else"),
  - there is no sentence that provides scenario-scoped advisory guidance.

Return "yes" if the response follows this system prompt.
Return "no" otherwise.
```