DISCOURSE BOUNDARIES SHAPE CONTEXT RETRIEVAL IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

How do discourse boundaries influence which information language models retrieve from context? We study punctuation and chat-template delimiters using controlled random-token sequences and natural-language prompts. Across tasks, boundary placement strengthens induction and changes the relative influence of competing context associations. We observe these effects across Pythia 2.8B, Qwen 2.5 (1.5B, 3B, 7B base and instruct), Gemma 4 (E2B, E2B-it, E4B, E4B-it), SmolLM2, Falcon 3, and OLMo 2. Mechanistic analysis in GPT-2 small localizes most of the downstream retrieval effect to the contextualized representation of the token immediately following the first occurrence of a repeated cue, carried through a relatively small set of attention-head key/value pathways. Boundary influence decays over a short range of token positions, with distinct decay profiles across attention heads. In natural-language tasks, chat-template boundaries also shift the relative influence of evidence on model predictions. This includes increased effective weighting of post-boundary evidence, even when models natively overweight older information, sometimes by an order of magnitude. Together, these results identify boundary-sensitive attention as a mechanism that shapes access to context and show how discourse formatting can alter the information used by language models.

1 INTRODUCTION

Natural language is structured by discourse boundaries at multiple scales, from commas and periods to paragraph breaks and conversational turns. In human communication, these boundaries guide attention and signal how information should be organized and retained (Grosz & Sidner, 1986; Chafe, 1994). Language models encounter these same markers during pretraining. Punctuation tokens draw heavy attention and can stand in for the text segments they close (Chen et al., 2025), and in GPT-2 they also help carry information across layers (Chauhan et al., 2026). Yet despite their prevalence throughout written text, the role of discourse boundaries in context retrieval remains largely unexamined.

Existing studies of context retrieval focus largely on token position. One common test plants a target fact within distractor text to measure recall across a prompt (Kamradt, 2023). Across these settings, models recall facts from the beginning and end of long prompts much better than from the middle (Liu et al., 2024). They also overweight recent text (Peysakhovich & Lerer, 2023), and in multi-hop questions performance falls as the gap between related facts grows (Baker et al., 2024). These studies treat retrieval primarily as a function of sequential distance. However, position alone cannot fully explain context access, because identical statements at identical positions can be recalled or ignored depending on the discourse boundaries that frame them.

In this paper, we show how discourse boundaries directly shape context retrieval across language models. By inserting localized boundaries such as punctuation or chat-turn markers into otherwise identical prompts, we isolate the influence of discourse structure while holding content and position constant. We evaluate this effect in synthetic sequences with controlled prefix repetitions (Olsson et al., 2022) and with conflicting completions, as well as in natural language inference tasks where competing statements provide conflicting evidence for a target question. For the natural-language sweep we test open-weight models spanning several families, including Pythia, Qwen 2.5, Gemma 4, SmolLM2, Falcon 3, and OLMo 2; mechanistic analysis is done on GPT-2 small.

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.

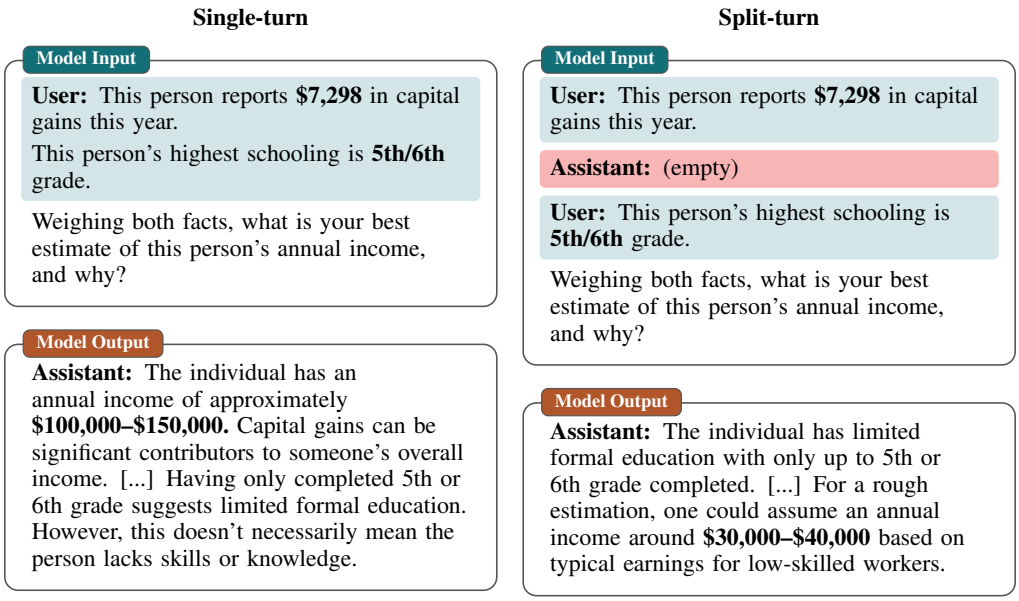


Figure 1: Illustration of single-turn vs split-turn prompting. This pair of facts are drawn from the Adult Census dataset and describe the same person. `User:` and `Assistant:` are turn labels, not a model's literal template, and the 'assistant turn' is empty. Verbatim system prompts, the scored question strings, and each family's chat template are in Appendix A.9; the exact fact pairs are in Appendix A.8. Model outputs are verbatim greedy completions from Qwen 2.5 1.5B-Instruct on this illustrative prompt (trimmed to opening + closing sentences, with `[...]` marking elided middle text). The chat-template boundary shifts the emphasis onto the second (LOW) fact and drops the estimated annual income from $100,000–$150,000 to $30,000–$40,000.

Across these architectures, discourse boundaries consistently strengthen induction and reweight competing associations in context. This effect depends on placement: a boundary near a query improves backward retrieval into earlier text, while a boundary near an evidence statement increases its forward salience. In natural language tasks, chat-template turn markers cause substantial shifts in evidence weighting, sometimes by an order of magnitude, overriding native positional bias. In practical chat settings, this means a model's answers can swing sharply depending on which turn a user shared a fact in (Figure 1). Circuit analysis in GPT-2 small illuminates the underlying mechanism: boundaries modify subsequent representations so as to be more easily retrieved later, with the effect decaying over a short token span. Together, these results show that context retrieval is not simply a passive consequence of token distance, but an active process shaped by discourse formatting.

2 Related Work

2.1 Discourse Boundaries in Human Language

In human discourse, boundaries structure what information remains active in attention. Grosz & Sidner (1986) give each discourse segment an attentional state, where a boundary changes which entities remain available for reference. Centering theory makes this update local, with each utterance refreshing the centers that subsequent text can pick up (Grosz et al., 1995), and entity availability follows segment structure rather than raw linear distance (Walker, 1998).

In written text, punctuation marks these boundaries for the reader. Nunberg (1990) treats punctuation as a grammar of text units, and Chafe (1994) links these breaks directly to cognitive flow, while eye-tracking studies confirm that readers dwell at commas and periods and that periods create stronger boundaries than commas (Hirotani et al., 2006; Andrews & Veldre, 2021). While these linguistic accounts describe human processing, we test whether language models show a similar sensitivity by isolating how a single boundary mark (punctuation or a chat-template delimiter) shifts retrieval.

2.2 Context retrieval in language models

In language models, context retrieval is heavily shaped by position. Models reliably use the start and end of a long prompt but miss information in the middle (Liu et al., 2024), reflecting a U-shaped positional attention bias that calibration can partly mitigate (Hsieh et al., 2024). Models also favor recent text over earlier text (Peysakhovich & Lerer, 2023), and multi-hop retrieval degrades as the distance between related facts grows (Baker et al., 2024). Standard synthetic evaluations like Needle in a Haystack similarly measure recall as a function of target position (Kamradt, 2023). While these studies establish where information sits across a prompt, they leave open how a localized discourse boundary reallocates attention between competing context spans.

Prompt formats and delimiters also alter how models use context. Standard chat templates degrade context utilization in instruction-tuned models (Wang et al., 2025), whereas explicit sentence delimiters improve multi-step reasoning on GSM8K and DROP (Liu et al., 2026). Structural tokens also leave a distinct footprint in attention: SepLLM retains punctuation tokens while compressing the remaining key-value cache (Chen et al., 2025), and punctuation-aware sparse attention uses punctuation tokens as anchors when choosing which blocks to attend to (Qiu et al., 2026). Unlike prior work that modifies whole prompts or measures overall task accuracy, we isolate individual turn and template boundaries to measure how they reweight competing facts in context. We find that a boundary raises the weight of the subsequent statement and supports greater use of earlier context, sometimes by an order of magnitude, across model families from GPT-2 small to modern instruction-tuned architectures.

2.3 Attention mechanisms and punctuation circuits

Inside transformer models, context retrieval depends on a competition between specialized retrieval circuits and attention sinks. For retrieval, induction heads search context for a previous occurrence of token A and copy whatever followed, completing patterns like $[A][B] \ldots [A] \rightarrow [B]$ (Olsson et al., 2022; Crosbie & Shutova, 2025), while related circuits route specific entity information across layers (Wang et al., 2023). In contrast, attention sinks absorb excess attention mass onto initial tokens (Xiao et al., 2024), serving as a default resting place when a head does not retrieve context (Ran-Milo, 2026). However, prior work examines these components separately, leaving open how text formatting shifts attention between sinks and active retrieval heads.

Recent interpretability work examines how punctuation affects internal representations. Punctuation tokens can act as information carriers across layers: in GPT-2, several layers depend heavily on punctuation to preserve performance, though this reliance is much weaker in newer model families (Chauhan et al., 2026). A comma at a clause boundary reduces the rise in prediction error about 40 to 80 tokens later, mainly in GPT-2 and OPT, through hidden states rather than through attention on the comma itself (Dangal & Shakya, 2026). In contrast to studies focused on next-token loss or static summary tokens, we show how a boundary directly alters retrieval circuits. In GPT-2 small, a boundary strengthens induction and redistributes attention away from a first-position sink: query changes promote retrieval of earlier information, while key changes make information more accessible to later positions, with influence decaying over a short token range across distinct heads.

3 Method

3.1 Cued retrieval in random token sequences

We construct random token sequences following Olsson et al. (2022) to study the effect of discourse boundaries on context retrieval while limiting semantic confounds. The tokens with highest and lowest frequency in OpenWebText (Gokaslan et al., 2019) are excluded, and then the rest are sampled uniformly at random without replacement. Single retrieval tasks have the form `[BOS] F1 K X F2 K`, where `[BOS]` is the sequence-initial token native to each model, `F1` and `F2` are fillers of random tokens, `K` is a multi-token sequence cue repeated at both locations, and `X` is the desired retrieval target. Dual retrieval tasks follow a similar structure with a second retrieval target `Y`: `[BOS] F1 K X F2 K Y F3 K`. We quantify single-target retrieval by the log probability of `X` and dual-target preference by the logit difference between `X` and `Y` at the final position. Unless stated otherwise, a

boundary delimiter is inserted at the positions immediately preceding one of the cues K, replacing random filler tokens so as to maintain a constant length.

3.2 Natural Language Prompting

Across 14 quantitative domains, we construct 10 fact-pairs (*HIGH*, *LOW*) which suggest answers at different ends of a potential value range. For example, one domain asks the model to infer the user's income. One *HIGH* fact is "I own a vacation home in the Hamptons", while one *LOW* fact is "I shop at Aldi and use SNAP benefits to feed my family."

A full list of the 14 domains, with descriptions and dataset sources for those anchored in real data, is given in Table 2; verbatim fact-pair listings are in Appendix A.8. Per-configuration normalized-swing grids for each model are in Appendix A.6. All configurations use the same 14 system prompts (one per domain); full listings of configurations, chat-template families, and system prompts are available in Appendix A.9.

For each fact-pair we compare four conditions across two dimensions. First, we vary the fact order: *LOW-HIGH* vs. *HIGH-LOW*. Second, we place both facts within a single user turn (single-turn), or we place an empty assistant turn between them (split-turn). We then tokenize the prompt plus the model's completion jointly, and sum the per-token log probabilities of the tokens beyond the prompt alone. Results are provided in Section 4.2.

4 Experiments

We probe the effect of discourse boundaries on subsequent processing via comparing sequences with a localized perturbation in the form of a boundary. First, we analyze random token sequences with minimal structure, where the task is to continue a subsequence that occurred earlier (Olsson et al., 2022). We introduce a conflicting completion experiment, whereby two instances of a subsequence occur with different continuations, and the logit difference between the continuations gives the relative salience of the two subsequences at the readout. Next, we introduce a suite of natural language inference tasks from different domains, where two statements are followed by a question with a quantitative answer so that the relative weight of the statements can be measured. We use this to measure the effect of boundaries between the statements, and find consistent effects across a variety of models: discourse boundaries, and especially chat template turns, amplify the statement immediately following it relative to counterfactual arrangements without the boundary.

4.1 Synthetic retrieval in randomized token sequences

Single retrieval task. First, we construct single-retrieval tasks with filler lengths $|\texttt{F1}| = |\texttt{F2}| = 32$, and use cues of size $|\texttt{K}| = 3$. In the midst of random token sequences, the repeated cue provides a signal for context retrieval. We measure the change in $\log p(\texttt{X})$ after inserting a boundary before the first or second cue K (Fig. 2), averaged over 512 task instances.

Several broad trends emerge. First, there is clear dependence on the nature of the boundary, with new lines and chat templates (which often contain multiple newlines) causing the largest effects. Punctuation has a mixed effect; in many models the period token is the punctuation with the largest effect. Common words such as `the` and `and` generally have little effect. Second, it is often the case that a boundary at either placement boosts the retrieval of X. In the first location, the cue and X are not distinguishable from the other random tokens, as they have not occurred previously, suggesting the boundary boosts retrievability of all subsequent tokens. In the second location, the cue that follows has occurred previously, and the backwards retrieval is boosted. For several models, the effect of the boundary is similar in magnitude at either location; Olmo's fine-tuned variants have little to negative effect at the first location while the second location remains effective. Finally, the nature of the empty turn chat template as a "loud" discourse boundary is sensible across models. Without having seen the template in the first stage of training, Pythia exhibits similar effects whether the template is scrambled or filled with random tokens; after fine-tuning, the template causes a larger effect with its structure intact than without. The same behavior occurs to different degrees for Qwen, SmolLM2, and OLMo 2.

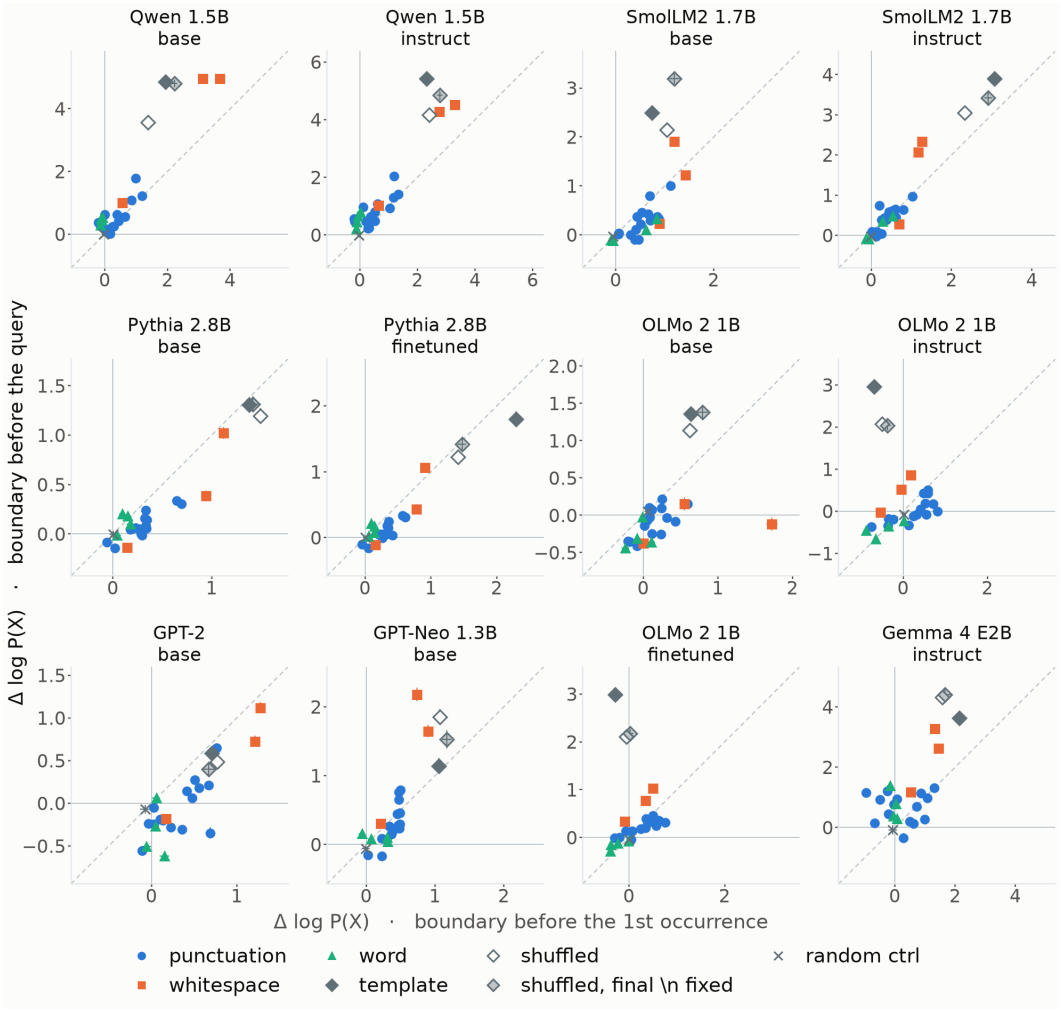


Figure 2: **Effect of different boundaries in single retrieval task.** Across a range of models, the retrieval boost for different forms of the boundary directly preceding the first cue K (horizontal axis) and directly preceding the second cue (vertical axis). Dashed lines represent where the two effects match.

Dual retrieval task. In Fig. 3, the location of a \n boundary was swept in the dual retrieval task where a sequence has two instances of the cue K before the final readout, and each cue is followed by a different token (X and Y). Both X and Y are continuations with prior contextual support, and we find that the presence of the newline boundary boosts one or the other based on its location. Across five models, the effect is fairly robust. A boundary immediately before the first cue boosts the first continuation (X). A boundary immediately before the second cue boosts the second continuation (Y) and generally suppresses the first continuation. Finally, a boundary after both pieces of context often boosts both, improving retrieval. As the length of the cue increases from one to six, the retrieval task becomes easier and the effect of the boundary is also amplified.

Mechanistic analysis of GPT-2 small. For more of a fine-grained account of the effect on processing, we examine GPT-2 small in the single retrieval setup, with cue length $|K| = 1$. The effect appears to reside primarily in the continuation X and in particular attention heads. Fig. 4a charts the change in attention mass placed on the first K X by all heads in the model at the final K position as the proximity to the boundary is varied. The change is almost always largest when the boundary immediately precedes either K ($\Delta = 0$), and then decays with distance. Most of the change occurs at the X position. Fig. 4b&c fix the boundary before the first cue and then patch key and value vectors

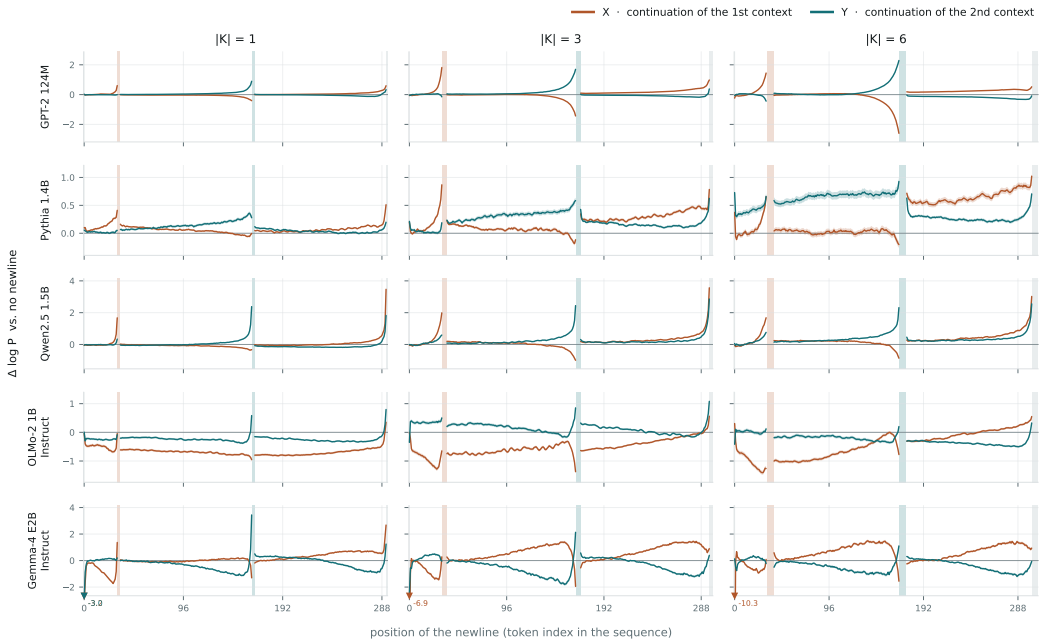


Figure 3: **Effect of boundary placement in dual context setup.** The change in log probability assigned to either of two context-supported continuations (`X` and `Y`, orange and green) is shown as the location of a newline character (`\n`) is swept across the sequence. The length of the repeated cue `K` is one (left), three (middle), or six (right) random tokens. Shaded orange (green) vertical bands indicate the location of `K X` (`K Y`), and the shaded gray band indicates the position of the final cue. For Gemma (bottom row), replacing the first token with the newline caused a change to the log probability of `Y` that exceeds the plots' bounds.

at the final readout position for a causal account of the effect. When the entire set of keys and/or values from the earlier locations are patched into the final stream, the effect localizes strongly to the `X` position, with substantial interaction between its key and value channels. Patching all keys and values from both the `K X` positions recovers 85% of the boundary's effect on retrieval, whereas patching from the baseline sequence (no boundary) removes 84% of the effect. Greedy optimization over the full set of keys and values identifies a limited set of vectors and locations that carry the effect; nearly all reside in `X`.

4.2 Natural language prompting with inference from split context

Figure 1 illustrates the single-turn and split-turn conditions on the Adult Census domain. Swapping the LOW and HIGH facts yields the remaining two conditions.

To measure the effect of the chat-template boundary on the same natural scale as the model's answers, we normalize each order swing by the model's own single-fact response. Specifically, for a given fact-pair, let E_H and E_L denote the expected answer under a single user turn containing only the HIGH fact and only the LOW fact respectively (no boundary, no competing fact). The single-fact gap $E_H - E_L$ is the answer swing produced by moving from full LOW-only evidence to full HIGH-only evidence. The normalized order swing under condition $c \in \{\text{single}, \text{split}\}$ is given by

$$\widetilde{\Delta}_c = \frac{E_c[LH] - E_c[HL]}{E_H - E_L},$$

so that $\widetilde{\Delta}_c = +1$ means the model's answer follows the last-mentioned fact as strongly as it would follow that fact alone, $\widetilde{\Delta}_c = 0$ means the ordering has no effect, and $\widetilde{\Delta}_c = -1$ means the model

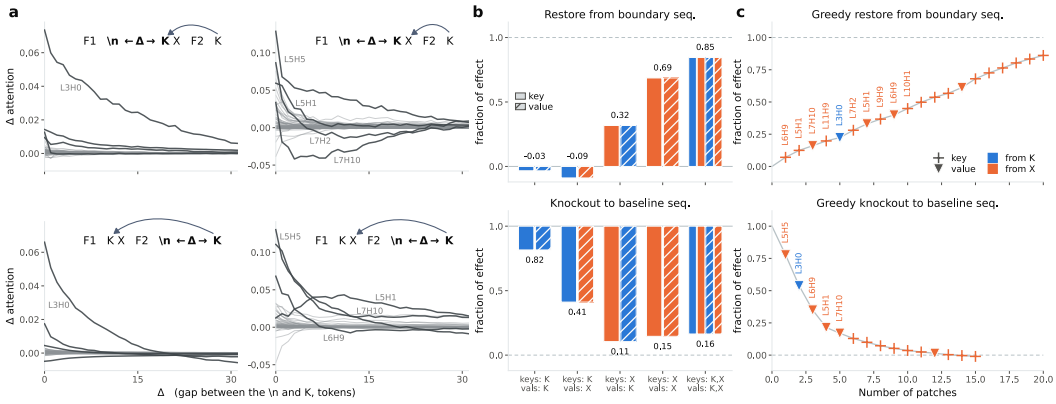


Figure 4: **Mechanistic analysis of the effect in GPT-2 small. (a)** The boundary effect is proximity-based. The change in attention mass placed on the first K (left) and X (right) due to the presence of a newline preceding the first cue (top) or the second (bottom). All 144 heads are shown as a function of the gap between the newline boundary and the cue K. The top four heads in each subplot by MSE from zero are emphasized. **(b)** With the boundary at the first cue, the effect largely resides in the continuation X. For the same sequence with and without the boundary, we patch key and value vectors from the boundary sequence onto the baseline (top) and vice versa (bottom) at the final readout position. **(c)** With the same patching scheme as in **b**, we perform greedy optimization over the 576 key and value patches from the initial K X positions into the second K's stream.

follows the first fact instead (primacy). Plotting $\widetilde{\Delta}_{\text{single}}$ against $\widetilde{\Delta}_{\text{split}}$ illustrates the effect of the chat-template boundary: points above the $y = x$ line indicate that the boundary makes the pre-existing order effect larger.

Figure 5 shows the effect on one representative configuration, Qwen 2.5 1.5B-Instruct. Across nearly every domain, most fact-pairs lie above $y = x$: the split-turn chat pushes the model toward the most recently mentioned fact more strongly than the single-turn chat does. On several domains the split-turn swing actually saturates the single-fact gap ($|\widetilde{\Delta}_{\text{split}}| \geq 1$), meaning the boundary functions to make the model answer as if it had only seen the second fact. In some instances, the single-turn swing was negative (evidence of primacy bias), while the split-turn flipped positive (evidence of recency bias). Aggregating across domains, Figure 6 confirms the pattern holds across all 31 configurations: nearly every model lies above $y = x$ on the aggregate scatter, and neither base nor instruction-tuned models escape the effect.

These results have significant implications across the many contexts of LLM chat interfaces. User chats often create multiple chat boundaries; we find that, depending only on the order of facts provided around those chat boundaries, LLMs will make substantially different inferences. In Appendix A.7 we test these results under perturbations to the answer-elicitation format: the effect is generally robust when the model is asked to provide a 'more or less' response instead of a specific value, and partially attenuates when models are asked to respond to a multiple-choice question, which is a potential limitation of our work and an area of future study.

5 Discussion

We have explored how discourse boundaries shape context retrieval in language models by comparing prompts that differ only in a boundary. On random-token sequences, a single boundary strengthens induction and shifts the balance between two competing continuations (Section 4.1). In GPT-2 small, changes to queries after the boundary help the model retrieve earlier information, while changes to keys make text after the boundary more accessible to later positions (Section 4.1). Both changes divert attention away from the first-position sink, with their influence fading over a

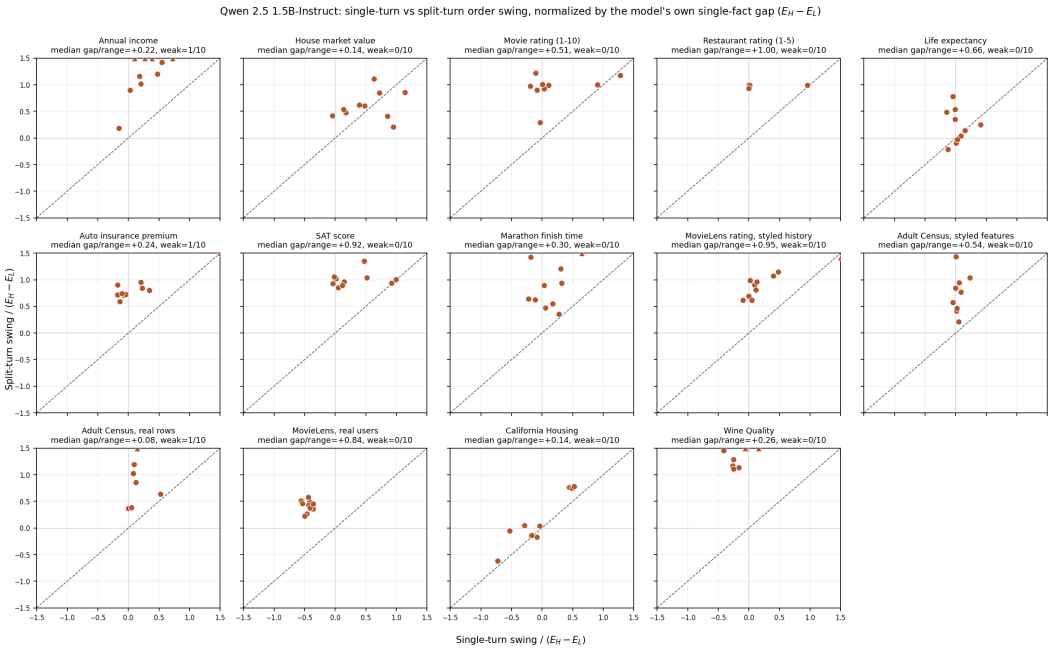


Figure 5: Normalized order swing on Qwen 2.5 1.5B-Instruct across domains. Each panel plots the 10 fact-pairs as $(x, y) = (\widetilde{\Delta}_{\text{single}}, \widetilde{\Delta}_{\text{split}})$, or the single-turn order swing vs. split-turn order swing, normalized by that pair's single-fact gap $E_H - E_L$. Dashed line ($y = x$) represents no boundary effect. Filled points are fact-pairs where the model discriminates HIGH vs. LOW under single-fact prompting (gap is at least 5% of the domain's bucket range); hollow points are 'weak' fact-pairs where the normalization is unstable; triangles are clipped fact-pairs with normalized order swings exceeding ± 1.5. The vast majority of points sit above $y = x$: the chat-template boundary systematically enlarges the pre-existing order effect that exists without a boundary. Per-configuration grids for all configurations are available in Appendix A.6, while Figure 6 summarizes the same effect across models.

short span of tokens. For natural-language inference tasks, an empty assistant turn between two facts raises the weight of the second fact across most models and domains, in some cases by more than tenfold (Section 4.2).

Together, these results show that discourse boundaries steer retrieval independently of content or linear distance. In random-token sequences, where semantic content plays no role, a boundary still increases attention to subsequent tokens. In natural language, turn markers similarly raise the salience of adjacent statements, indicating that effects once attributed to positional bias (Liu et al., 2024; Peysakhovich & Lerer, 2023) depend on where boundaries fall. Attention sinks offer a mechanistic explanation: attention heads default to first-position sinks when they lack relevant context (Ran-Milo, 2026), and a boundary redirects that attention toward retrieval circuits. These shifts have direct practical consequences. In multi-turn dialogues, identical facts placed in different turns yield conflicting answers. Long-context benchmarks like Needle in a Haystack (Kamradt, 2023) may reflect the boundaries around a target fact as much as its position. Methods that correct positional bias (Hsieh et al., 2024) must therefore account for boundaries, which prompt designers can also place deliberately to guide retrieval.

Our study has several limitations. First, although these boundary effects resemble accounts from human discourse (Grosz & Sidner, 1986; Chafe, 1994), this does not show that models process discourse like humans. Second, our circuit analysis covers only GPT-2 small; whether the same mechanism holds in larger models remains an open question. Third, the natural-language tasks always place the boundary directly before the second fact, so on their own they cannot separate a post-boundary advantage from recency bias; our random-token tasks resolve this by moving the

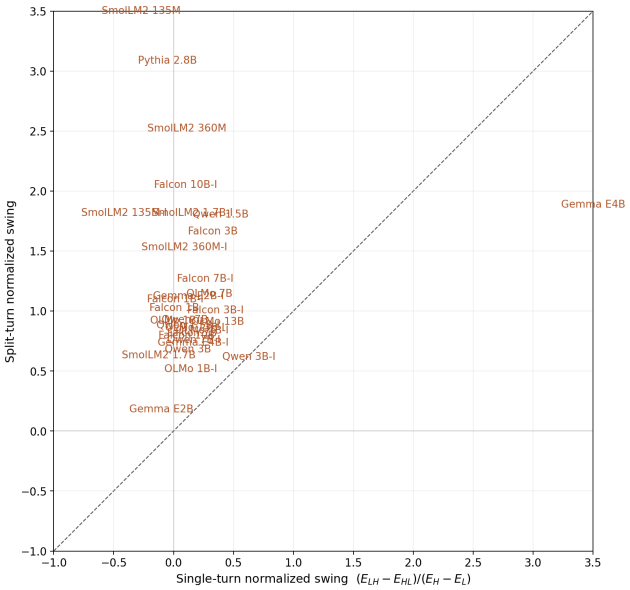


Figure 6: The normalized order swing represented in Figure 5 with one label per model, aggregated across domains. For each (model, domain), the mean numerator $E[LH] - E[HL]$ and the mean denominator $E_H - E_L$ across the 10 fact-pairs give the per-domain normalized swing. Plotted coordinates are the mean of those per-domain values across up to 14 domains (domains where the model fails to discriminate HIGH from LOW under single-fact prompting, i.e. mean gap ≤ 0, are skipped). Nearly every model lies above $y = x$: the chat-template boundary enlarges the pre-existing order effect that exists without a boundary. Values outside the plot range are clipped to the axis limits.

boundary across positions. Finally, we evaluate open-weight models up to 13B parameters, and the effect weakens under multiple-choice answer formats (Appendix A.7), likely because the options repeat content in the prompt. Future work can test whether facts restated in an assistant turn gain similar weight, and evaluate how boundary effects behave over longer prompts and multi-turn dialogues.

In summary, our results show that discourse boundaries directly govern context retrieval in language models. While prior work has treated context access primarily as a function of token position and sequential distance, our results demonstrate that boundaries actively allocate attention and reshape the salience of competing facts. Attention to boundaries alongside token position will hopefully inspire more reliable prompt design, more accurate evaluations, and better control over how language models retrieve context.

AI USE STATEMENT

In this work, we used generative AI tools to refine the research question, implement the code for the natural-language experiments, and interpret results, including the normalized-swing definition and the appendices. We have not used generative AI tools to develop the conceptual framework or drive the experimental design, and formulating mathematical claims, writing proofs, translation, and qualitative data analysis are not applicable to this work. Additionally, we used generative AI tools to draft and edit prose, create figures and captions, and write and organize code. We have reviewed all AI-assisted work. Humans conceived the study and drafted the behavioral sections. Humans read and revised the AI-written text, figures, captions, and appendices. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

This work does not involve human subjects or new private data. The real-data domains use existing public datasets as published (Table 2). Our results show that an empty chat turn can sharply change a model's answer on tasks such as income and housing estimates. We release no new personal records.

Reproducibility statement

Experiment settings, prompts, model lists, and fact pairs are in Sections 3 and 4 and Appendices A.9, A.8, A.6, and A.7. For the natural-language runs, model configurations and log evidence are stored in JSON files. The code, data, and logs are included as an anonymous supplementary zip (supp.zip).

References

Sally Andrews and Aaron Veldre. Wrapping up sentence comprehension: The role of task demands and individual differences. *Scientific Studies of Reading*, 25(2):123–140, 2021. doi: 10.1080/10888438.2020.1817028.

George Arthur Baker, Ankush Raut, Sagi Shaier, Lawrence E. Hunter, and Katharina von der Wense. Lost in the middle, and in-between: Enhancing language models' ability to reason over long contexts in multi-hop QA. *arXiv preprint arXiv:2412.10079*, 2024.

Wallace Chafe. *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. University of Chicago Press, 1994.

Sonakshi Chauhan, Maheep Chaudhary, Choy Kwan Kiu, Samuel Nellessen, and Nandi Schoots. Punctuations and predicates in language models. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 5622–5636, 2026.

Guoxuan Chen, Han Shi, Jiawei Li, Yihang Gao, Xiaozhe Ren, Yimeng Chen, Xin Jiang, Zhenguo Li, Weiyang Liu, and Chao Huang. SepLLM: Accelerate large language models by compressing one segment into one separator. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 9008–9028. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/chen25bf.html.

Joy Crosbie and Ekaterina Shutova. Induction heads as an essential mechanism for pattern matching in in-context learning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5049–5111, 2025. doi: 10.18653/v1/2025.findings-naacl.283. URL https://aclanthology.org/2025.findings-naacl.283/.

Sagar Dangal and Manoj Shakya. Attention degradation, function token anchoring, and the limits of attention-based intervention in large language models. *arXiv preprint arXiv:2607.20524*, 2026.

Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus, 2019.

Barbara J. Grosz and Candace L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225, 1995.

Masako Hirotani, Lyn Frazier, and Keith Rayner. Punctuation and intonation effects on clause and sentence wrap-up: Evidence from eye movements. *Journal of Memory and Language*, 54(3):425–443, 2006. doi: 10.1016/j.jml.2005.12.001.

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long T. Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. Found in the middle: Calibrating positional attention bias improves long context utilization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14982–14995, 2024. doi: 10.18653/v1/2024.findings-acl.890.

Gregory Kamradt. Needle in a haystack: Pressure testing LLMs. GitHub repository, 2023. URL https://github.com/gkamradt/needle-in-a-haystack.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638.

Zhichen Liu, Yongyuan Li, and Yang Xu. Think in sentences: Explicit sentence boundaries enhance language model's capabilities. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2275–2288, 2026. doi: 10.18653/v1/2026.acl-long.104. URL https://aclanthology.org/2026.acl-long.104/.

Geoffrey Nunberg. *The Linguistics of Punctuation*. Number 18 in CSLI Lecture Notes. Center for the Study of Language and Information, 1990.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Alexander Peysakhovich and Adam Lerer. Attention sorting combats recency bias in long context language models. *arXiv preprint arXiv:2310.01427*, 2023.

Junxiang Qiu, Shuo Wang, Zhengsu Chen, Hengheng Zhang, Jinda Lu, Changcheng Li, and Qi Tian. Punctuation-aware hybrid trainable sparse attention for large language models. *arXiv preprint arXiv:2601.02819*, 2026.

Yuval Ran-Milo. Attention sinks are provably necessary in softmax transformers: Evidence from trigger-conditional tasks. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 70–89, 2026.

Marilyn A. Walker. Centering, anaphora resolution, and discourse structure. In Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince (eds.), *Centering Theory in Discourse*, pp. 401–435. Oxford University Press, 1998.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Yihan Wang, Andrew Bai, Nanyun Peng, and Cho-Jui Hsieh. On the loss of context awareness in general instruction fine-tuning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 9541–9566. Curran Associates, Inc., 2025. doi: 10.52202/085713-0292. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/0cc540ef357216738768caaa7301b615-Paper-Conference.pdf.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024.

A EXPERIMENTAL PROTOCOLS FOR THE RANDOMIZED SEQUENCE RETRIEVAL TASKS

A.1 COMMON SETUP

Models. Table 1 lists every checkpoint used. All are public Hugging Face checkpoints, loaded at pinned commits (listed in the released code) with no fine-tuning, dropout or sampling. GPT-2 is run in float32 and all other models in bfloat16. Every quantity we report is read from the next-token distribution at the final position.

Tag	Checkpoint	Chat template provenance
gpt2	openai-community/gpt2	unseen
neo	EleutherAI/gpt-neo-1.3B	unseen
pythia14	EleutherAI/pythia-1.4b	–
pythia_base	EleutherAI/pythia-2.8b	sibling
pythia_sft	lomahony/eleuther-pythia2.8b-hh-sft	native
qwen_base	Qwen/Qwen2.5-1.5B	sibling
qwen_inst	Qwen/Qwen2.5-1.5B-Instruct	native
olmo2_base	allenai/OLMo-2-0425-1B	sibling
olmo2_sft	allenai/OLMo-2-0425-1B-SFT	native
olmo2_inst	allenai/OLMo-2-0425-1B-Instruct	native
smollm2_base	HuggingFaceTB/SmolLM2-1.7B	sibling
smollm2_inst	HuggingFaceTB/SmolLM2-1.7B-Instruct	native
gemma4_e2b_it	google/gemma-4-E2B-it	native

Table 1: Checkpoints. *Chat template provenance* applies to the boundary scan (§A.2). It records whether the checkpoint was itself tuned on the chat template used as its boundary (*native*), whether only its family's tuned variant was (*sibling*), or whether the family has no chat convention (*unseen*).

Token pool. We draw all filler, cue and continuation tokens from a mid-frequency pool built separately for each tokenizer. Token frequencies are unigram counts over the first 20,000 documents of a fixed 100k-document OpenWebText sample. We sort tokens by count and keep frequency ranks 1,000–19,999. The upper cut removes function words and punctuation, whose unigram priors would otherwise dominate the next-token distribution. The lower cut removes rarely seen, poorly trained tokens. We then remove special tokens, tokens whose decoding contains an incomplete UTF-8 byte sequence, the BOS token and the newline token. The resulting pools hold 18,880–18,991 tokens, depending on the tokenizer.

Sequences and controls. Each item is a sequence of pool tokens drawn *without replacement*, so apart from the deliberate repetition of the cue K (a run of $|K|$ tokens), no token appears twice. Every sequence begins with the tokenizer's BOS token (`<|endoftext|>` for tokenizers that declare none). A boundary is introduced by *replacing* one or more filler tokens rather than inserting new ones, so the boundary and control sequences have the same length, and every other token keeps the same content and the same absolute position. Every effect is a paired, within-item difference between the manipulated sequence and the same item without the manipulation. We report the mean over items with ± 1 standard error of the mean. Each item draws its own tokens, including its own cue and continuations, independently of every other item. The n items together cover nearly the whole pool, and any two items share only a few tokens. Item generation is seeded, so within a model every condition (and every scanned position) is evaluated on the same n items, which is what makes the differences paired. Because items are sampled from each tokenizer's own pool, models that share a tokenizer (GPT-2 and GPT-Neo, the two Pythia models, and each base/tuned pair) see identical items. Models with different tokenizers see items built by the same procedure from different tokens. Changing $|K|$ changes the layout, so each cue length also has its own item set.

A.2 SINGLE-CONTEXT RETRIEVAL: BOUNDARY SCAN (FIGURE 2)

Layout. With $|K| = 1$:

$$\texttt{<BOS>}\ \ F_1^{(32)}\ \ K\ \ X\ \ F_2^{(128)}\ \ K$$

This is 164 tokens, with K at positions 33 and 163 and X at position 34. We read the next-token distribution at the final K and score $\log P(X)$, the continuation that followed the cue's first occurrence.

Boundary slots. A boundary is written into one of two slots, each immediately preceding an occurrence of the cue. *pre1* is the last token of F_1 (position 32), just before the first occurrence. *pre2* is the last token of F_2 (position 162), just before the query. A multi-token boundary of length L replaces the last L tokens of the block, so it always ends immediately before K.

Conditions. There are 26 boundary conditions in each slot:

- 22 single-token boundaries: the punctuation marks `. , ; : ! ? - --- ( ) " ' / * #`, the whitespace tokens newline, double newline and space, and the words `the`, `and`, `␣the` and `␣and`. Each was verified to be a single token in every tokenizer.
- *template*: the model family's user→assistant chat boundary. We extract it from the instruction-tuned tokenizer of the family as the longest common suffix of the chat-template renderings of two different user messages, with the generation prompt appended. This isolates the boundary from any message content. We then re-encode it with the evaluated model's own tokenizer. The resulting templates are 5–7 tokens long. GPT-2, GPT-Neo and Pythia have no chat template, and for them we use `"\n\nAssistant:\n\nHuman:"` (8–9 tokens). This is the turn format that `pythia_sft` was tuned on, hence *native* in Table 1.
- *shuffled*: the template's tokens in a random order, drawn afresh for each item.
- *shuffled, final token fixed*: as above, but the template's last token stays in place. For every family with a chat template, that last token is a newline. For the generic template it is `:`.
- *random control*: L fresh pool tokens in the template's slot. Since the unmanipulated sequence already holds random filler there, this measures the null effect of redrawing L filler tokens. The redraw is distinct within itself, but it is not excluded from the rest of the sequence, so a small fraction of items receive a token that already occurs elsewhere.

Single-token boundaries are compared directly against the unmanipulated sequence. The random control is the matched null only for the template conditions.

Measurement. For each model, condition and slot we report $\Delta \log P(X)$ relative to the same item without a boundary, averaged over $n = 512$ items. Each panel of Figure 2 plots the effect at pre1 (x) against the effect at pre2 (y) for one model, with ± 1 SE on both axes. The dashed diagonal marks equal effects in the two slots. Baseline $\log P(X)$ without a boundary ranges from -9.8 to -6.5 across models.

A.3 DUAL-CONTEXT RETRIEVAL: NEWLINE PLACEMENT (FIGURE 3)

Layout.

$$\texttt{<BOS>}\ \ F_1^{(32)}\ \ K\ \ X\ \ F_2^{(128)}\ \ K\ \ Y\ \ F_3^{(128)}\ \ K$$

This is $291 + 3|K|$ tokens (294, 300 and 309 for $|K| = 1, 3, 6$). The same cue K appears three times: its first occurrence is followed by X, its second by a different continuation Y, and the third is the query. We read the next-token distribution at the final position and score both $\log P(X)$ and $\log P(Y)$.

Scan. One newline at a time replaces the token at a single position. We scan every position of F_1, F_2 and F_3, plus the BOS position as a reference point (289 positions in all). Positions inside a cue block or on X or Y are not scanned. For each position we report $\Delta \log P(X)$ and $\Delta \log P(Y)$ relative to the same item without a newline. Because each difference compares a continuation's

probability with itself, the continuation's unigram prior cancels exactly, so no counterbalancing of X and Y is needed.

Grid. Five models (GPT-2, Pythia-1.4B, Qwen2.5-1.5B, OLMo-2-1B-Instruct, Gemma-4-E2B-it) are each run at cue lengths $|K| \in \{1, 3, 6\}$. This gives 15 cells, each with $n = 512$ items. In Figure 3, lines show means and shaded bands ± 1 SE. The y axis is shared within each row (model), and the x axis is the absolute token index of the newline. Unscanned spans are shaded: the first and second cue blocks are tinted in the colour of the continuation that follows them, and the final cue block is grey. The BOS point is excluded from each row's y range. Where it falls outside that range, it is drawn as an arrow at the axis edge, labelled with its value.

A.4 Mechanism in GPT-2 (Figure 4)

All mechanism experiments use GPT-2 (12 layers $\times$ 12 heads, $d_{\text{head}} = 64$) in float32. Float32 matters here because many of the measured quantities are within-item differences of order 10^{-2}, and bfloat16 rounding error would be a non-negligible fraction of that. The boundary is always a single newline.

(a) Attention scans. These measure attention without intervening, and compare each sequence against the same item with a random pool token (absent from the item) in the boundary slot. We use two geometries:

- *Boundary before the first occurrence* (top row): <BOS> $F_1^{(32)}$ \n $F_2^{(128)}$ Q (163 tokens). The boundary sits at position 33. The query Q at position 162 is set to a copy of the token at position $33 + o$, the cue's first occurrence, for offsets $o = 1, \ldots, 32$. We read attention from Q onto that first occurrence (K) and onto the token after it (X, the induction target).
- *Boundary before the query* (bottom row): <BOS> $F_1^{(32)}$ C_1 $F_2^{(128)}$ \n $F_3^{(32)}$ (195 tokens). The first occurrence C_1 is fixed at position 33 and the boundary at 162. A second copy of C_1 replaces the filler token at position $162 + o$. We read attention from that second copy onto C_1 (K) and onto position 34 (X).

The plotted distance is the gap $\Delta = o - 1$, the number of filler tokens between the newline and the cue, so $\Delta = 0$ means adjacency. For each offset and each of the 144 heads, we report the mean over $n = 256$ items of the post-softmax attention probability with the newline minus that without it. In each panel, the four heads with the largest mean-squared difference over the plotted range are drawn dark and the rest as thin lines. The two geometries differ in one respect that matters when comparing rows. In the first, the read position is fixed and the distance between the copies shrinks as Δ grows. In the second, the copies are fixed and the read position moves away from them. Within a single offset, the two conditions differ only in the boundary token, so this geometry cancels in the plotted difference.

Setup for (b) and (c): path patching. Here $|K| = 1$ and $F_1 = F_2 = 32$: <BOS> F_1 K_{33} X_{34} F_2 K_{67} (68 tokens). The *boundary* sequence has a newline replacing position 32, immediately before the first K, and the *clean* sequence does not. On $n = 2048$ items, the clean baseline is $\log P(X) = -6.504$ and the boundary raises it by 1.505 ± 0.031 nats. We call this increase the *full effect*, and report every patching result as a fraction of it.

We patch the direct edge from the key and value vectors at the two source positions (33 and 34) into the attention of the read position (67), and nothing else. For a patched head h in layer ℓ, we recompute position 67's attention row as

$$a = \mathrm{softmax}\big(q_{67}\,\tilde{K}^{\top}/\sqrt{d_{\text{head}}}\big), \qquad z_{67} = a\,\tilde{V},$$

where q_{67} is the run's own query. $\tilde{K}$ and $\tilde{V}$ are the run's own keys and values, except that selected rows at the source positions are replaced by the donor sequence's. We then substitute z_{67} into that head's slice of the attention output projection's input at position 67. No causal mask is needed, since 67 is the last position. The change propagates forward from position 67 through the later layers but reaches no other position. Donor activations come from a separate forward pass on the same

item. As a correctness check, patching the run's *own* keys and values through the identical code path changes $\log P(X)$ by exactly zero (to five decimal places) in every run.

We use two directions. *Restore* runs the clean sequence with boundary-run keys and values patched in, and asks which edges are sufficient to reproduce the effect. *Knockout* runs the boundary sequence with clean keys and values patched in, and asks which edges are necessary for it. Both are measured against the same unpatched clean baseline, so restore starts at 0 and knockout at 1 (the full effect).

(b) Fixed recipes. Every one of the 144 heads is patched at once. Five recipes cross which source position donates keys with which donates values: keys from K with values from K; keys from K with values from X; keys from X with values from K; keys from X with values from X; and keys and values from both positions. Each bar is a separate measured forward pass on all 2048 items. The recipes are strongly non-additive in the channels. Patching all heads' keys *and* values restores 0.85 of the effect, whereas keys alone restore 0.12 and values alone -0.12. In the knockout direction, the joint patch leaves 0.16 of the effect, against 0.08 for keys alone and 0.20 for values alone.

(c) Greedy search. A unit is one (layer, head, channel $\in \{k, v\}$, source $\in \{33, 34\}$), giving 576 units. Starting from the empty set, each step measures the *joint* effect of the current set plus each remaining unit on all 2048 items. It adds the unit that maximises the effect (restore) or minimises it (knockout), then re-measures the new set and records it. Nothing assumes additivity. This matters because one-at-a-time marginals summed across heads misestimate the joint effect by up to a factor of four. Every step completes a full scan of all remaining units; no step is cut short by the time budget.

Restore reaches 0.86 of the effect after 20 units and was still rising when stopped, so this figure is a lower bound for the search. Nineteen of the 20 selected units read from the target position X (34). The exception is the value of L3H0 at K. Knockout removes 83% of the effect with five units, all of them value vectors. The remaining effect falls to 0.004 of the full effect at 13 units and turns negative at 14. A minimising search continues past zero into units that actively lower $\log P(X)$, so we interpret the knockout curve only up to that crossing.

For each step the standard error of the joint effect is 0.008–0.029 nats (0.5–2% of the full effect), so the order of adjacent late picks whose gains are of similar size should not be read as a ranking. Selection and measurement use the same 2048 items, which biases each curve slightly in the direction of the search. The restore search ran in three segments, each resuming from the previous one's selected set. The re-measured starting value matched the checkpoint exactly, confirming that patching is deterministic across processes.

Scope. Only positions 33 and 34 are offered as sources, so contributions routed through other positions are excluded by construction, not by measurement. The mechanism results cover a single model, a single boundary token (newline, in the pre1 slot) and a single layout.

A.5 Named boundaries in the single-retrieval scatter

Figure 2 leaves the token name off each point so the markers stay readable. Figure 7 is the same scatter with every boundary named.

A.6 Per-configuration normalized-swing grids

For every configuration we plot the 14 domains as a grid. Each panel shows the domain's 10 fact-pairs as $(x, y) = (\widetilde{\Delta}_{\text{single}}, \widetilde{\Delta}_{\text{split}})$: the single-turn and split-turn order swings, both divided by that pair's own single-fact gap $E_H - E_L$. Filled markers are pairs where the model discriminates HIGH from LOW (single-fact gap is at least 5% of the domain's bucket range); hollow markers are 'weak' pairs where the normalization is unstable; triangles are pairs clipped to ± 1.5. Points above $y = x$ show boundary amplification.

A.6.1 Qwen 2.5

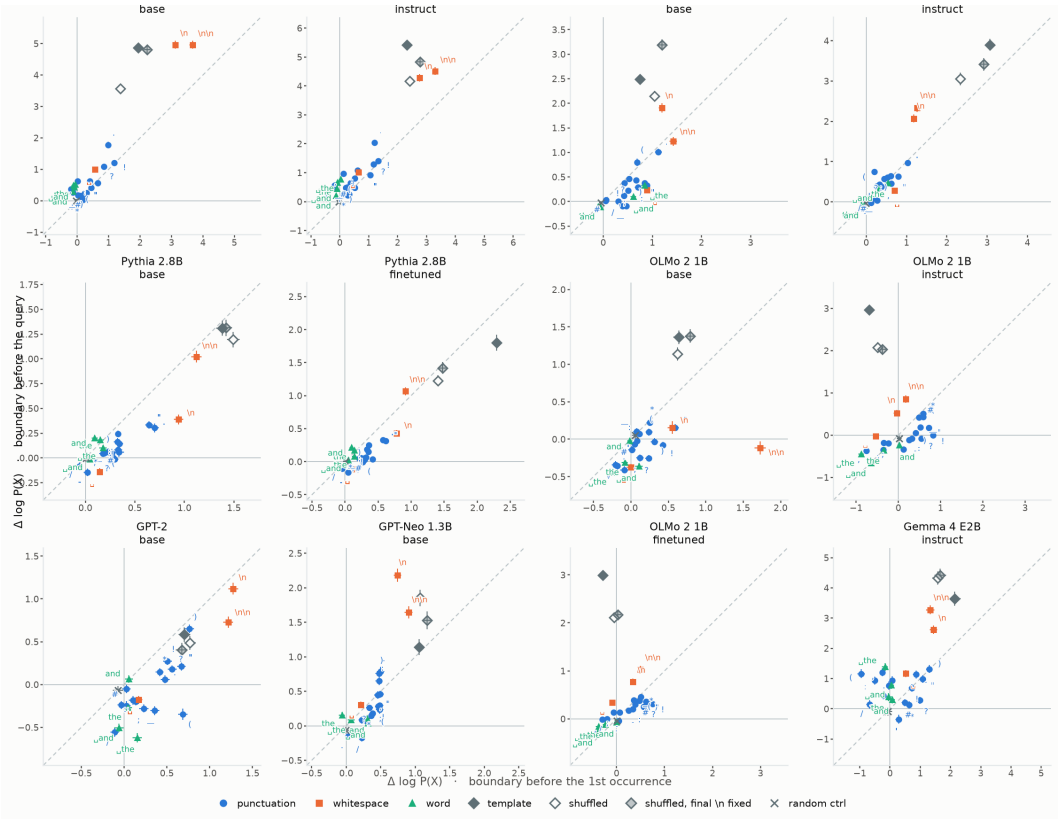


Figure 7: The single-retrieval boundary scatter from Figure 2, with each boundary named.

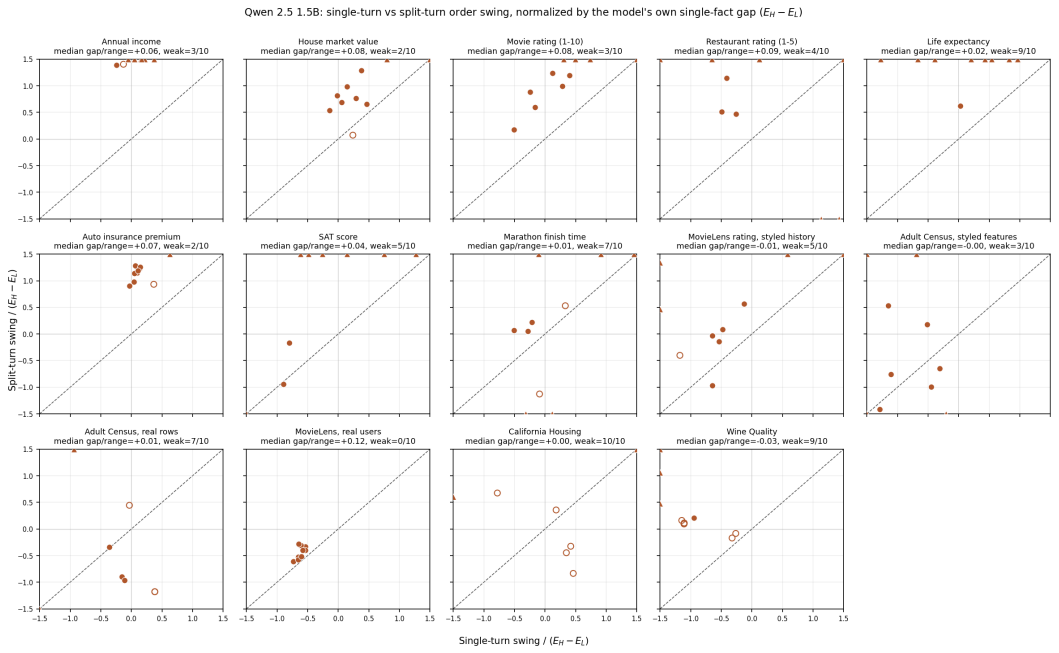


Figure 8: Qwen 2.5 1.5B: per-domain normalized-swing scatter.

Table 2: The 14 quantitative domains used in the natural-language experiments. Each domain has 10 fact-pairs; verbatim listings are in Appendix A.8.

Title	Description	Source
Real datasets (both facts from the same row)		
Adult Census, real rows	Both facts are two features of the same person.	Kohavi 1996 [link]
MovieLens, real users	Both facts are two ratings by the same real user.	Harper and Konstan 2015 [link]
California Housing	Both facts are two features of the same block group.	Pace and Barry 1997 [link]
Wine Quality	Both facts are two chemical measurements of the same wine.	Cortez et al. 2009 [link]
Real datasets (styled fact-pairs)		
MovieLens rating, styled history	Predicted 1-5 rating from styled MovieLens user history.	Harper and Konstan 2015 [link]
Adult Census, styled features	Predict whether the person earns above or below the $50K threshold, from styled Adult Census features.	Kohavi 1996 [link]
Synthetic		
Annual income	Infer a person's annual dollar income from lifestyle signals.	—
House market value	Estimate market value of a residential property.	—
Movie rating (1-10)	Rate a film on a 1-10 scale given reception details.	—
Restaurant rating (1-5)	Yelp-style star rating on a 1-5 scale.	—
Life expectancy	Years of remaining life given lifestyle and health.	—
Auto insurance premium	Annual auto insurance premium given driving history.	—
SAT score	Predicted SAT score from 400 to 1600.	—
Marathon finish time	Predicted marathon finish time in minutes.	—

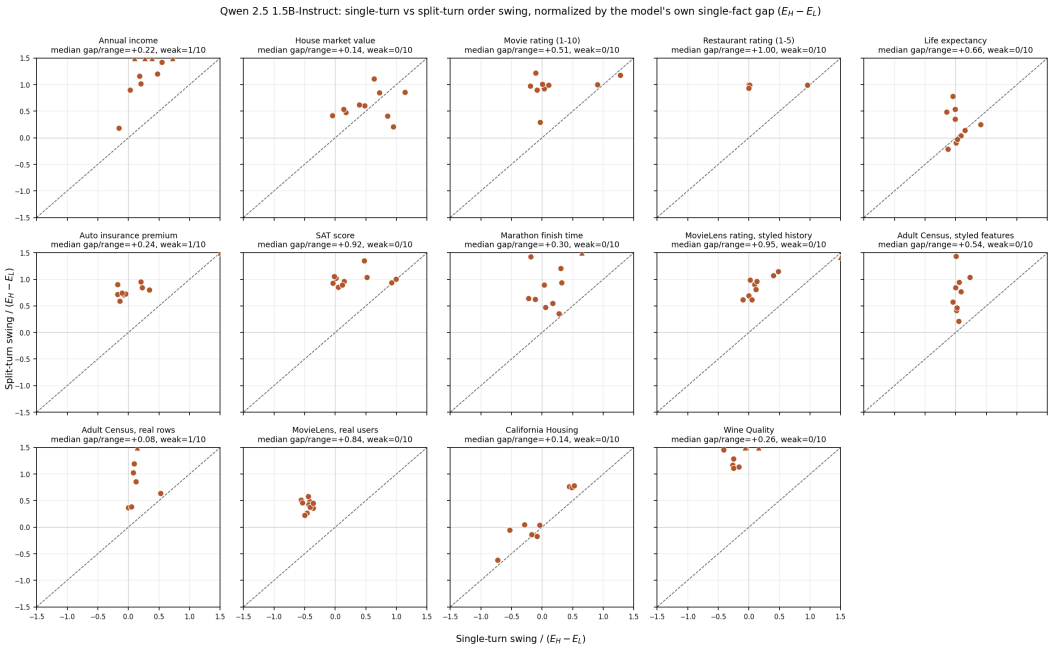


Figure 9: Qwen 2.5 1.5B-Instruct: per-domain normalized-swing scatter.

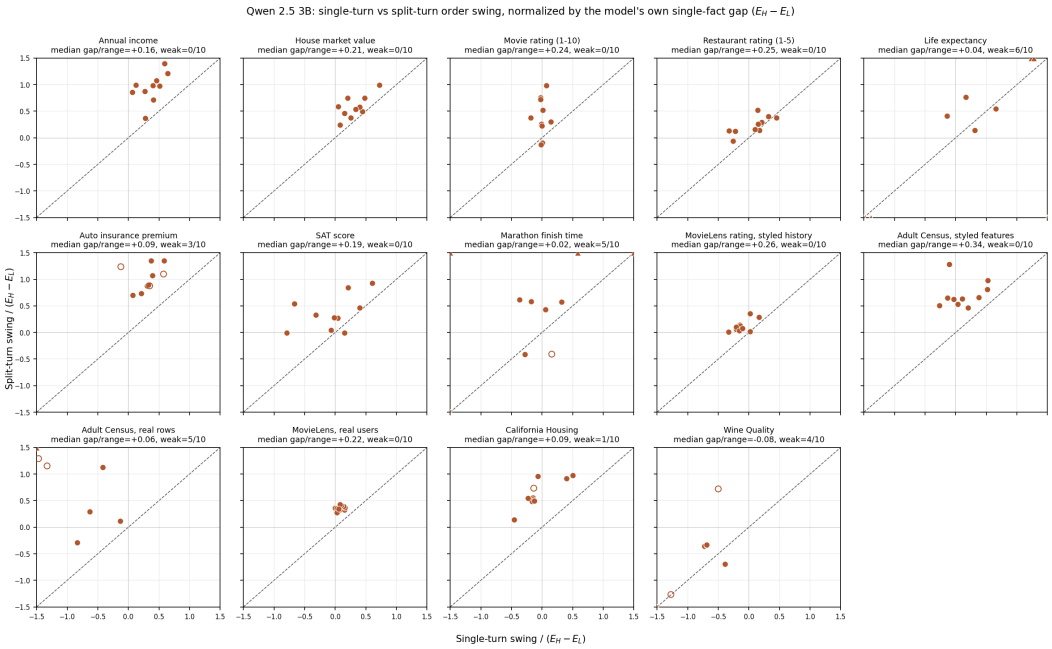


Figure 10: Qwen 2.5 3B: per-domain normalized-swing scatter.

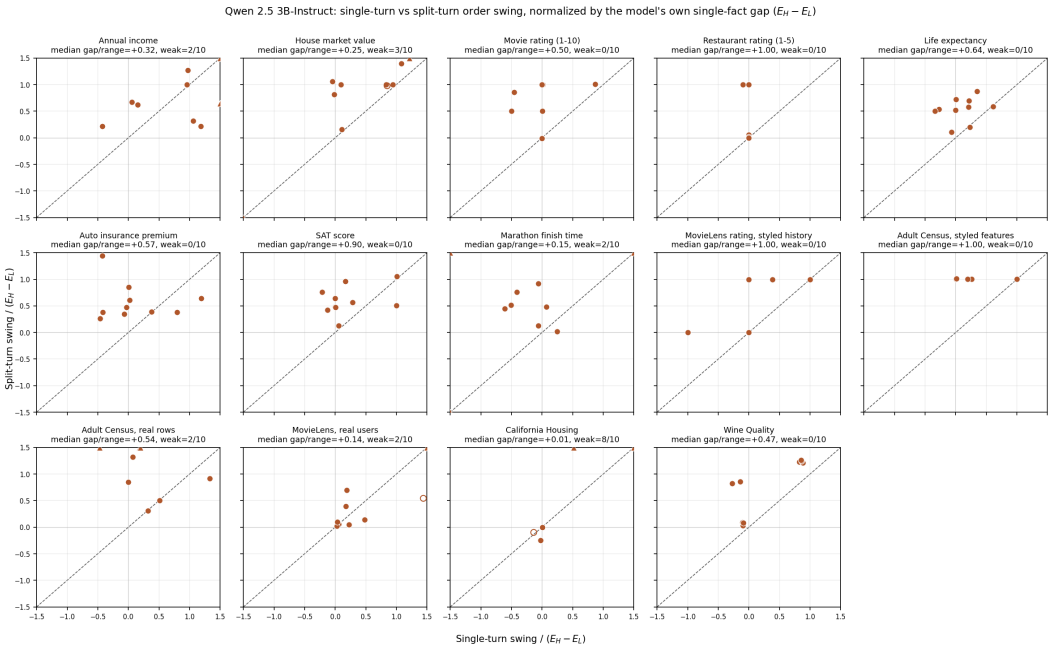


Figure 11: Qwen 2.5 3B-Instruct: per-domain normalized-swing scatter.

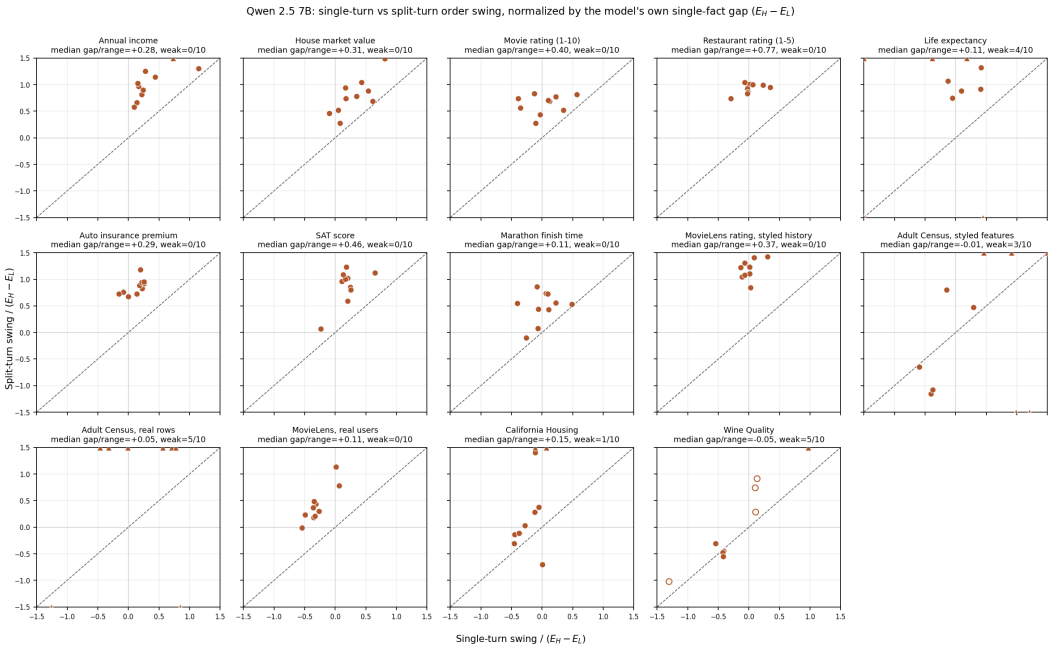


Figure 12: Qwen 2.5 7B: per-domain normalized-swing scatter.

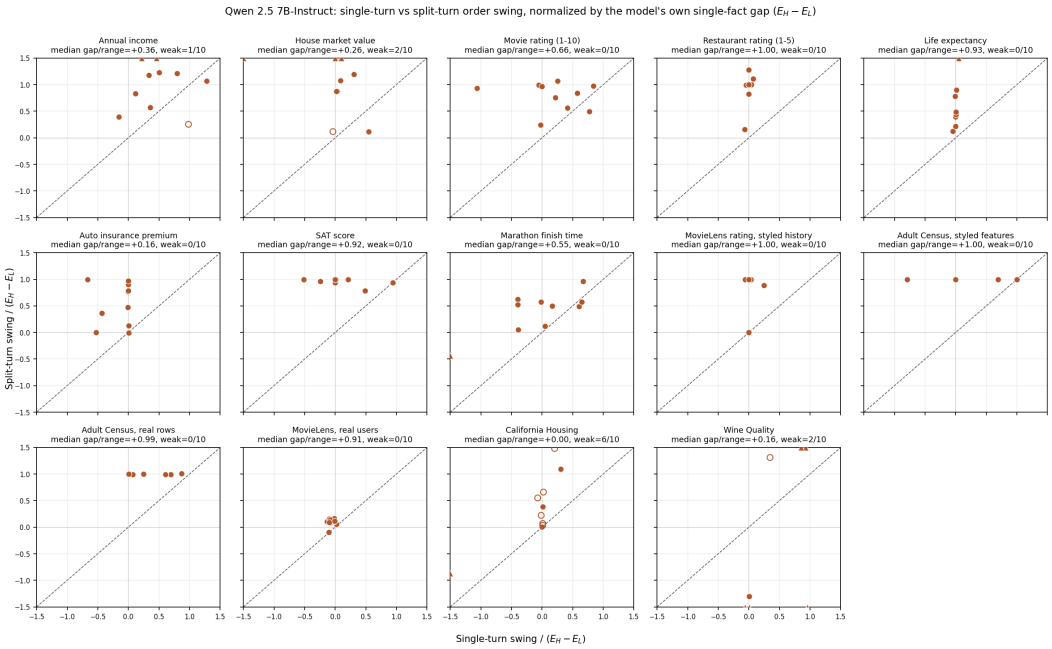


Figure 13: Qwen 2.5 7B-Instruct: per-domain normalized-swing scatter.

A.6.2 SMOLLM2

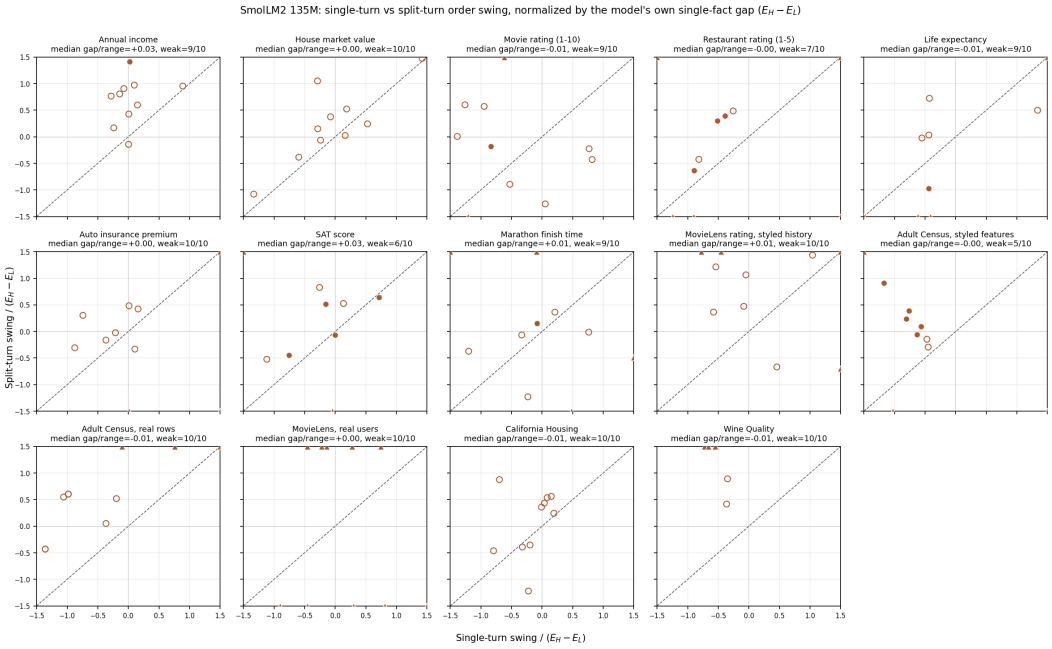


Figure 14: SmolLM2 135M: per-domain normalized-swing scatter.

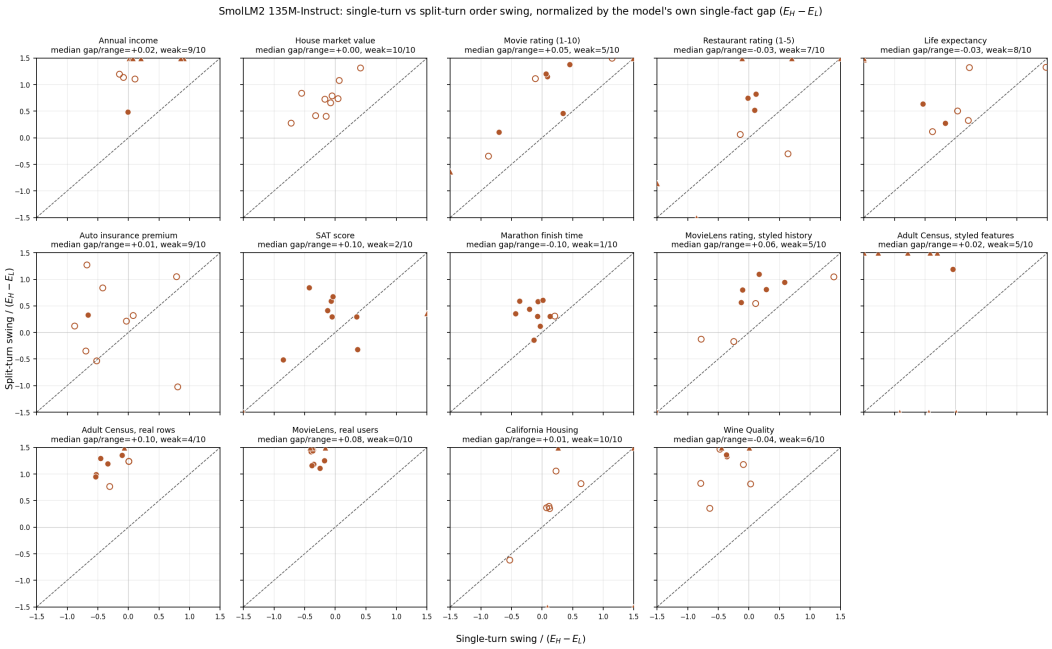


Figure 15: SmolLM2 135M-Instruct: per-domain normalized-swing scatter.

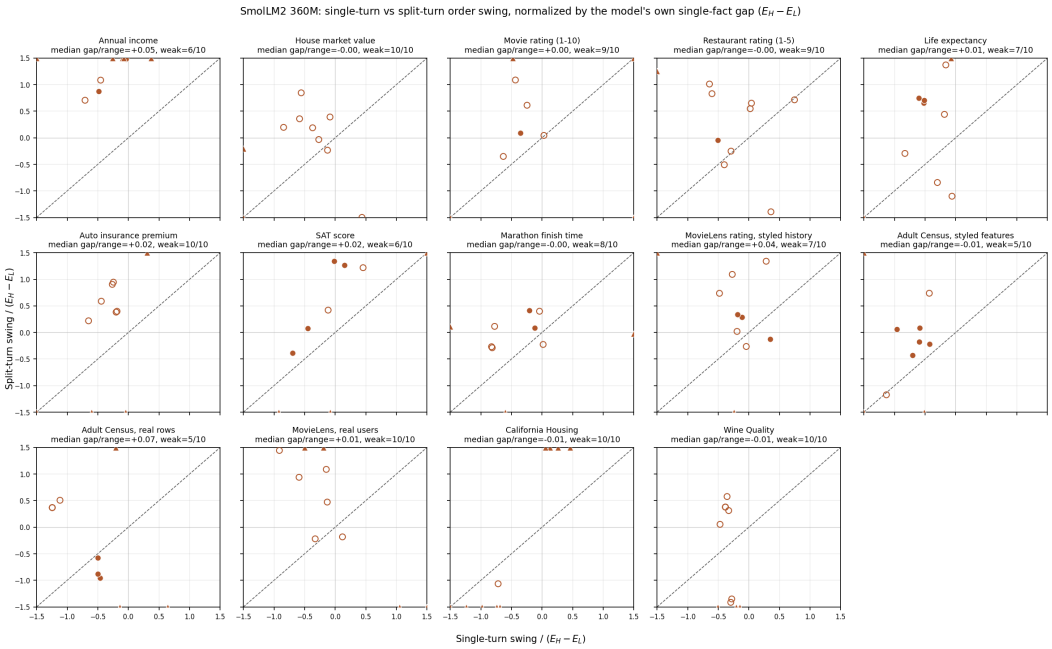


Figure 16: SmolLM2 360M: per-domain normalized-swing scatter.

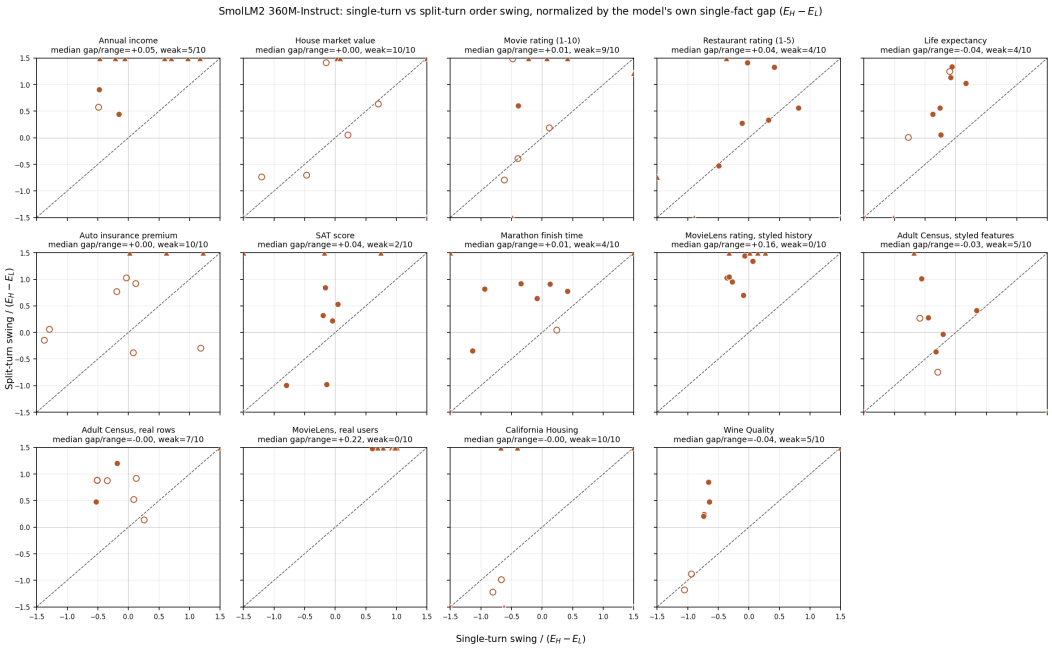


Figure 17: SmolLM2 360M-Instruct: per-domain normalized-swing scatter.

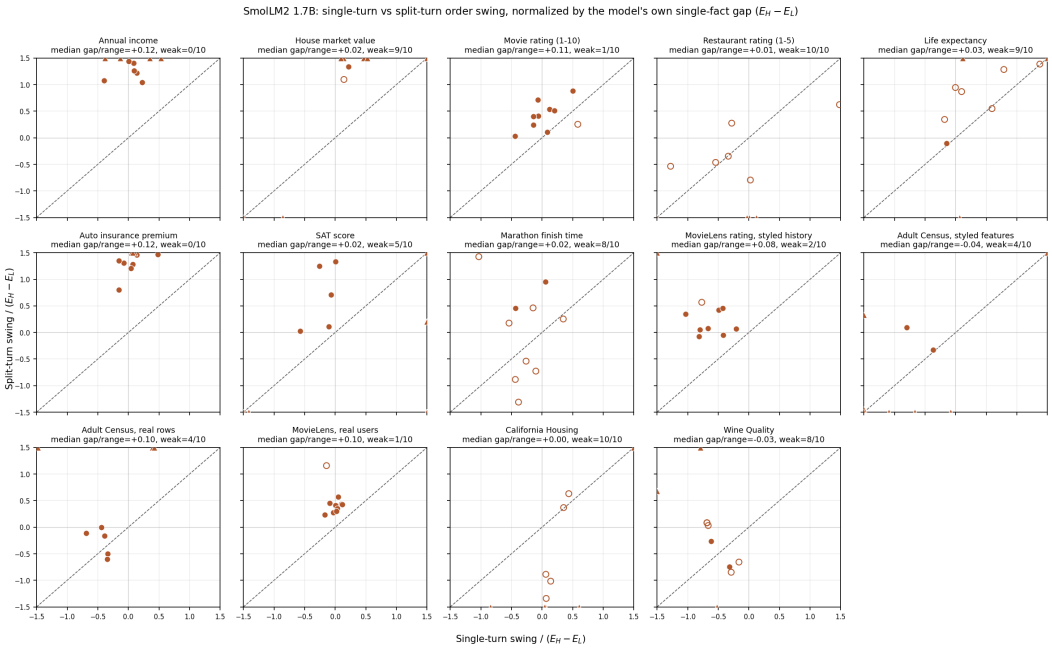


Figure 18: SmolLM2 1.7B: per-domain normalized-swing scatter.

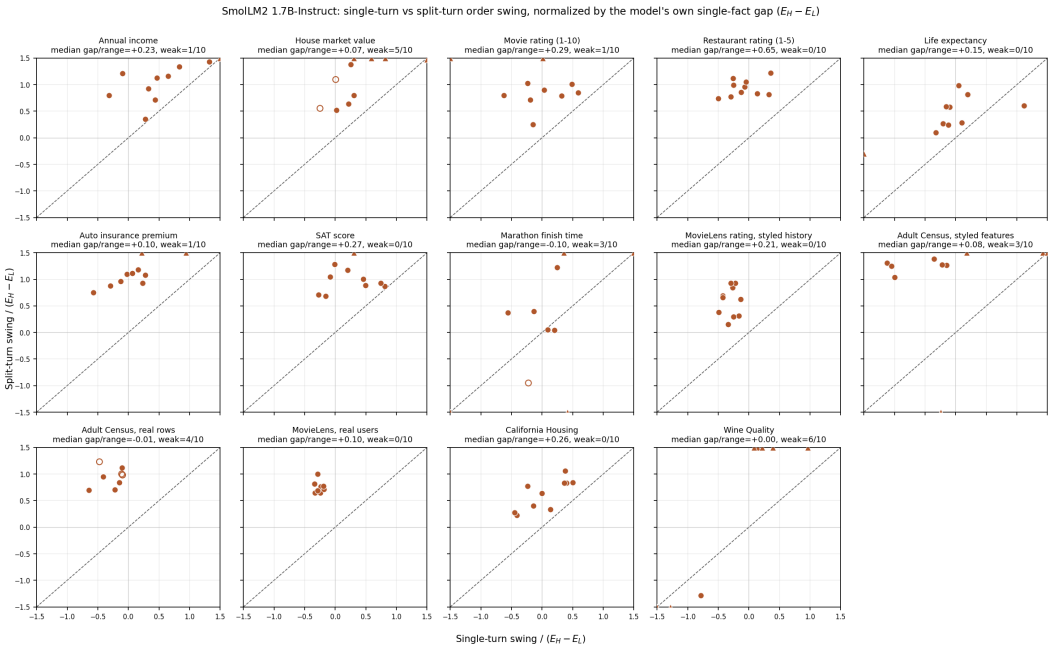


Figure 19: SmolLM2 1.7B-Instruct: per-domain normalized-swing scatter.

A.6.3 FALCON 3

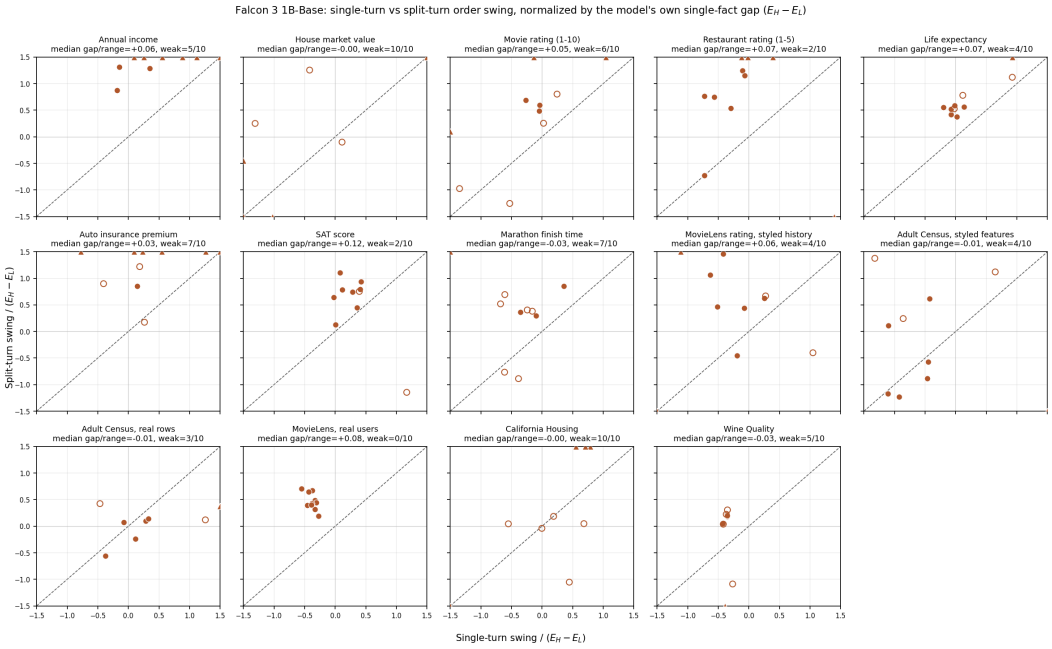


Figure 20: Falcon 3 1B-Base: per-domain normalized-swing scatter.

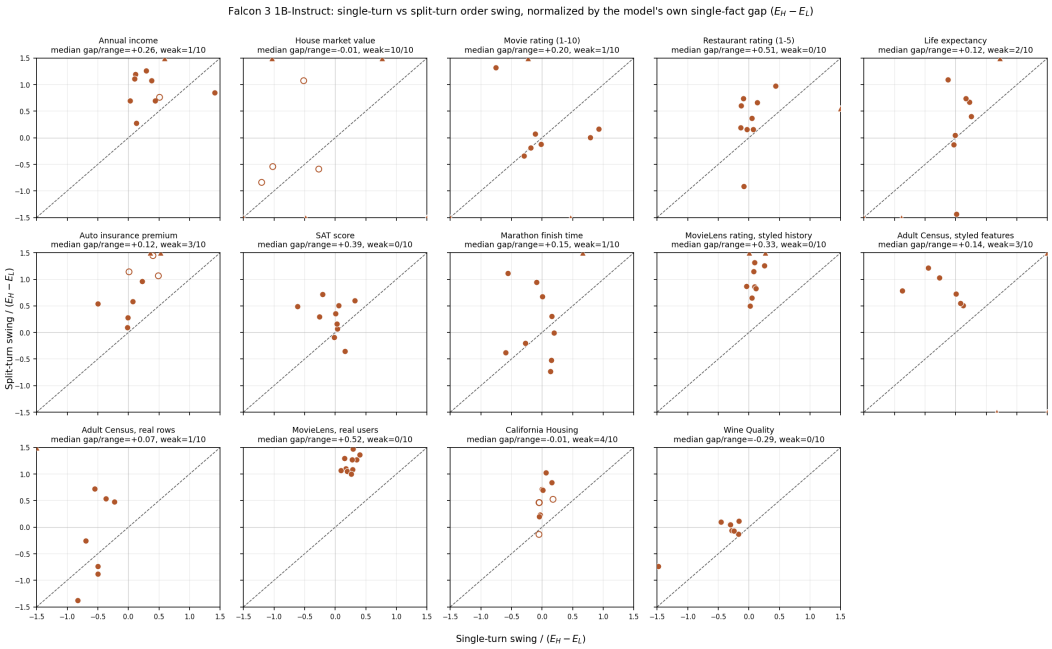


Figure 21: Falcon 3 1B-Instruct: per-domain normalized-swing scatter.

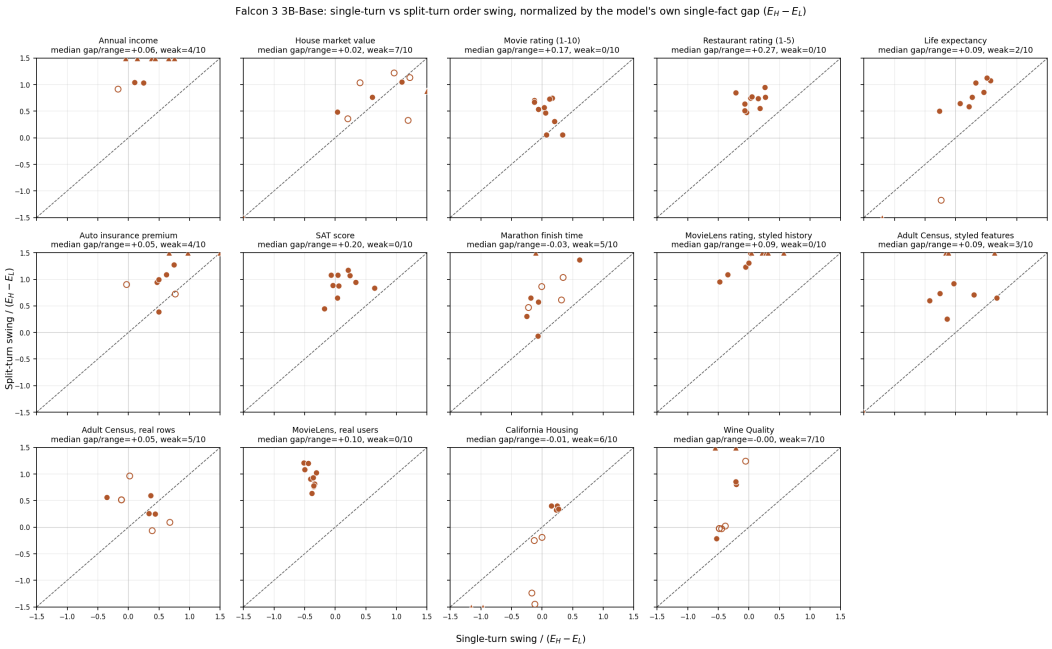


Figure 22: Falcon 3 3B-Base: per-domain normalized-swing scatter.

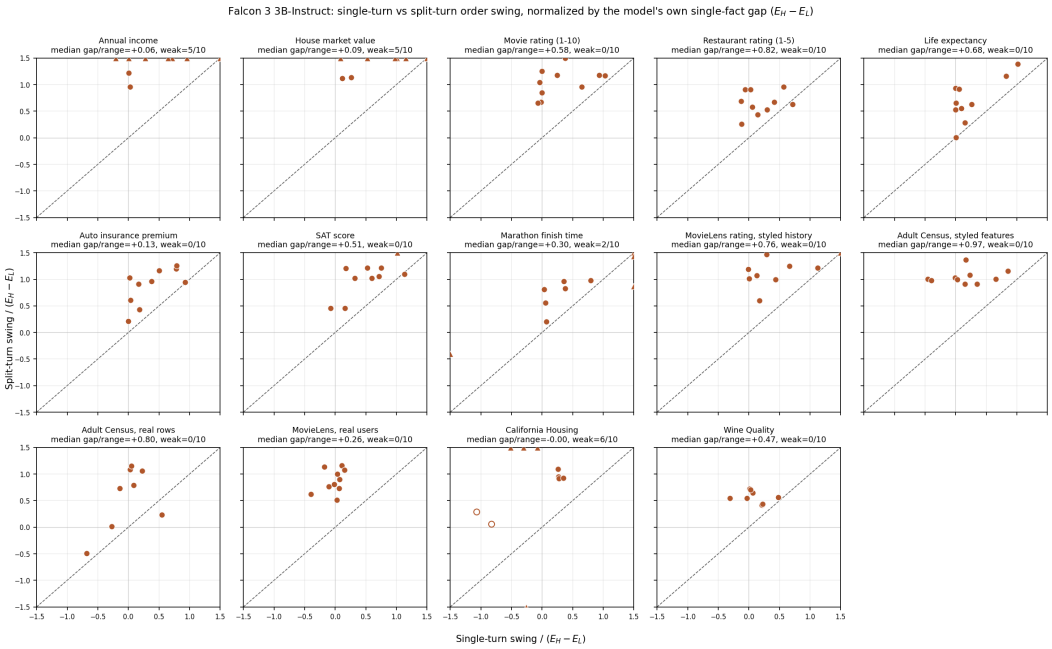


Figure 23: Falcon 3 3B-Instruct: per-domain normalized-swing scatter.

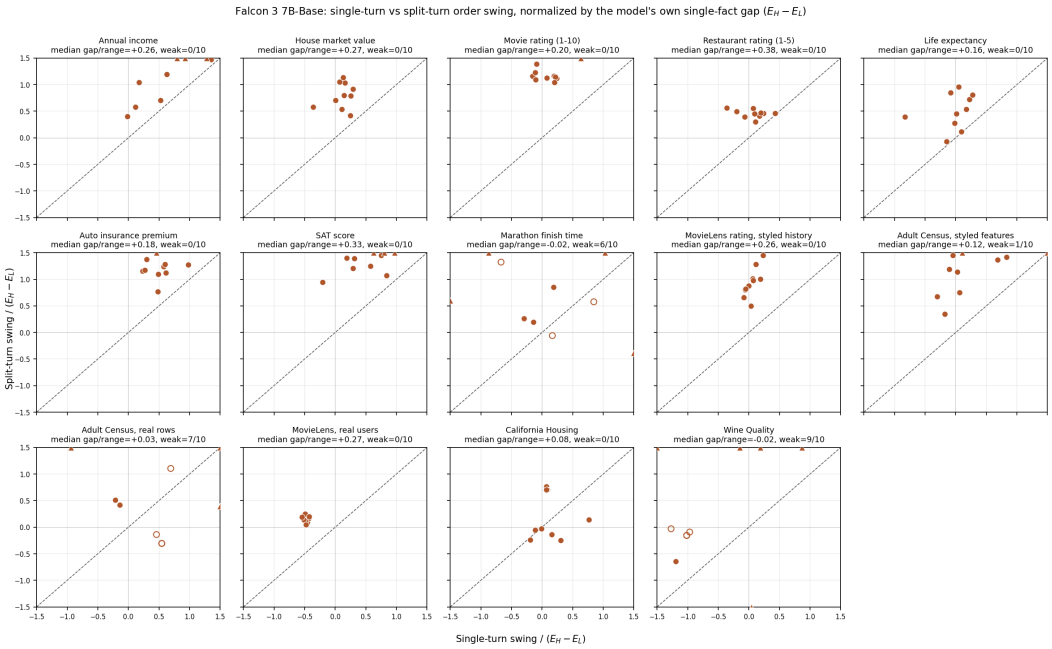


Figure 24: Falcon 3 7B-Base: per-domain normalized-swing scatter.

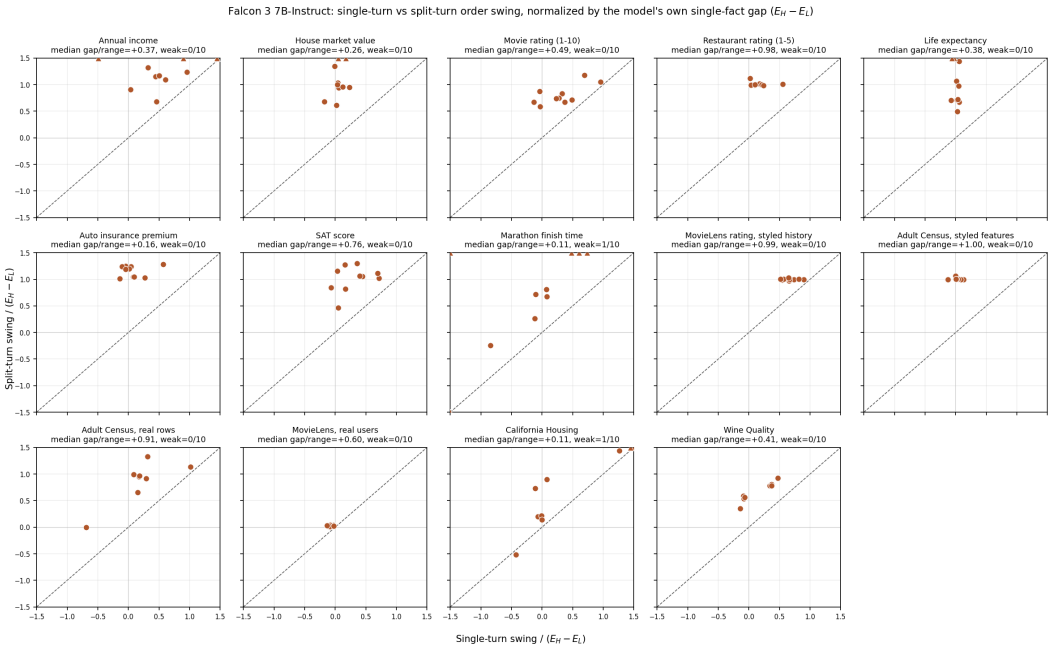


Figure 25: Falcon 3 7B-Instruct: per-domain normalized-swing scatter.

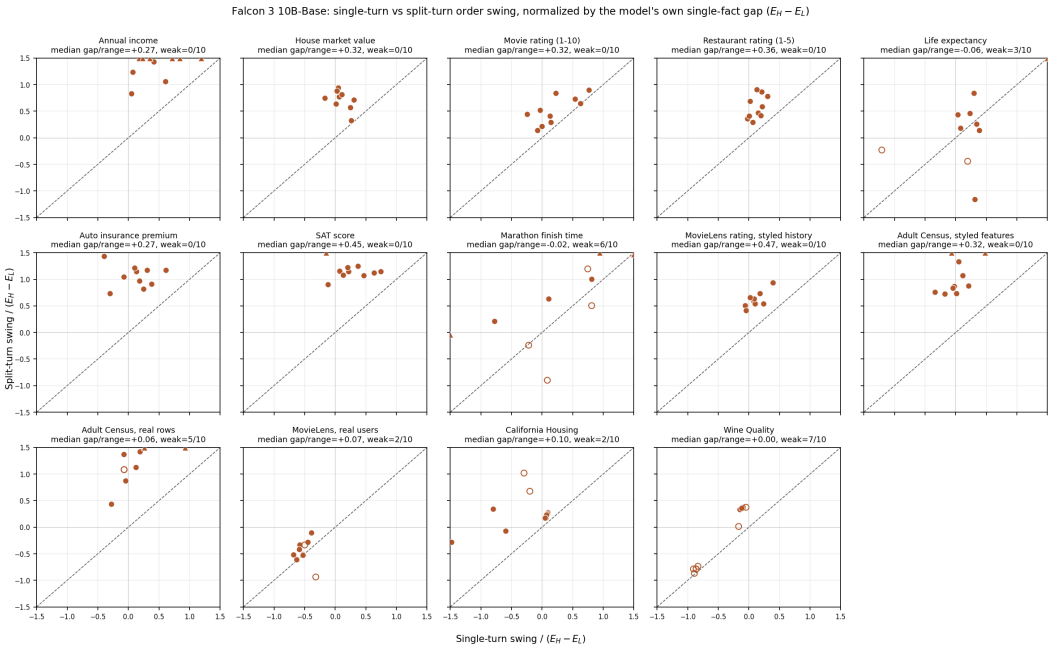


Figure 26: Falcon 3 10B-Base: per-domain normalized-swing scatter.

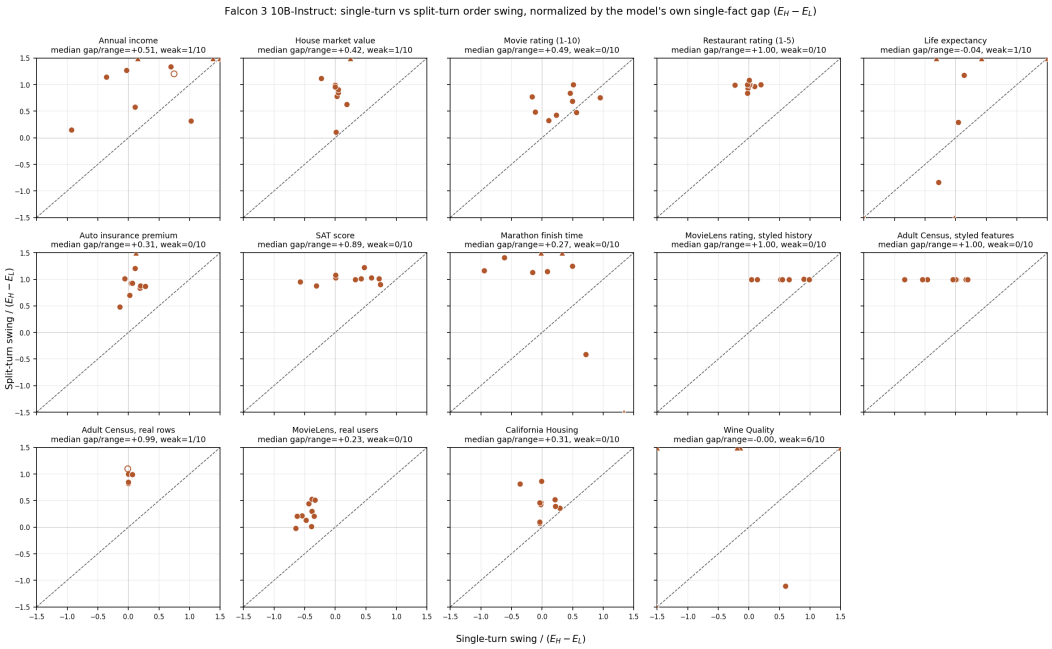


Figure 27: Falcon 3 10B-Instruct: per-domain normalized-swing scatter.

A.6.4 OLMo 2

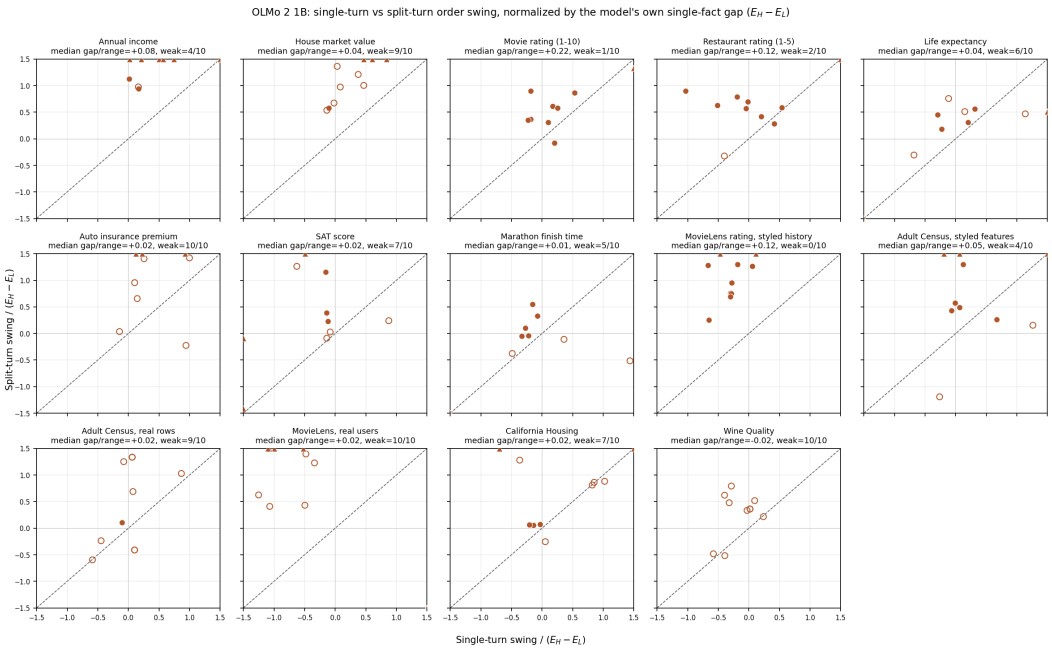


Figure 28: OLMo 2 1B: per-domain normalized-swing scatter.

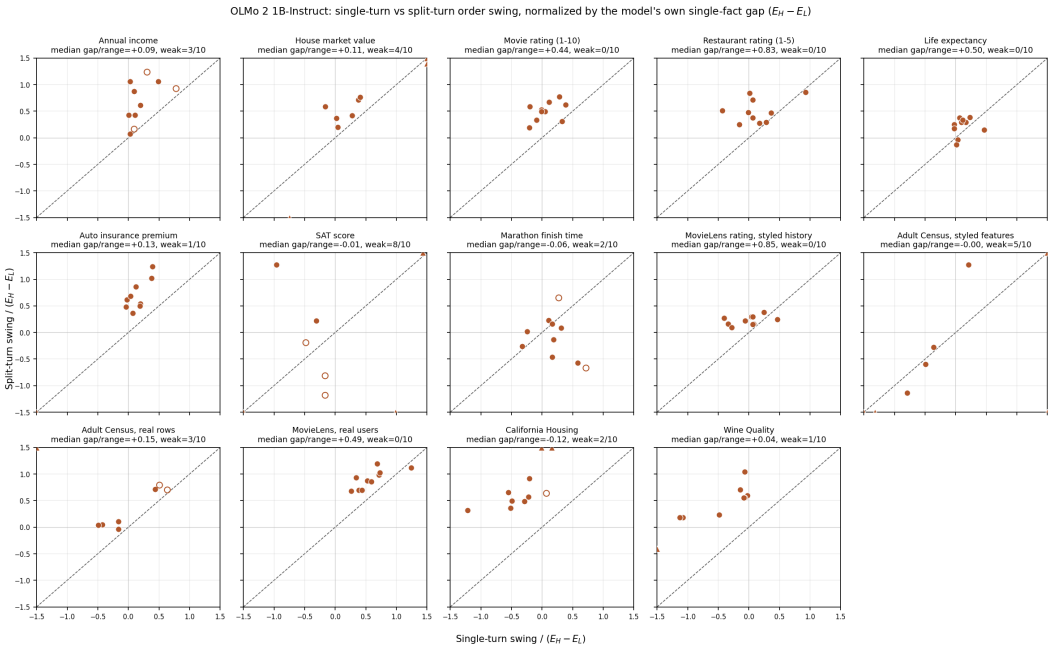


Figure 29: OLMo 2 1B-Instruct: per-domain normalized-swing scatter.

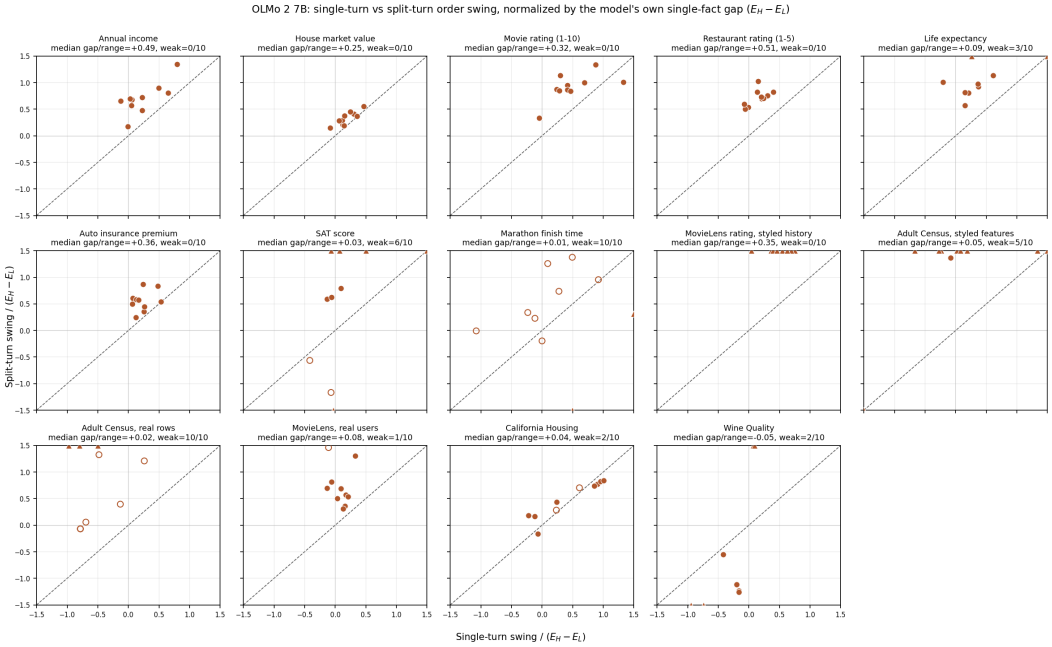


Figure 30: OLMo 2 7B: per-domain normalized-swing scatter.

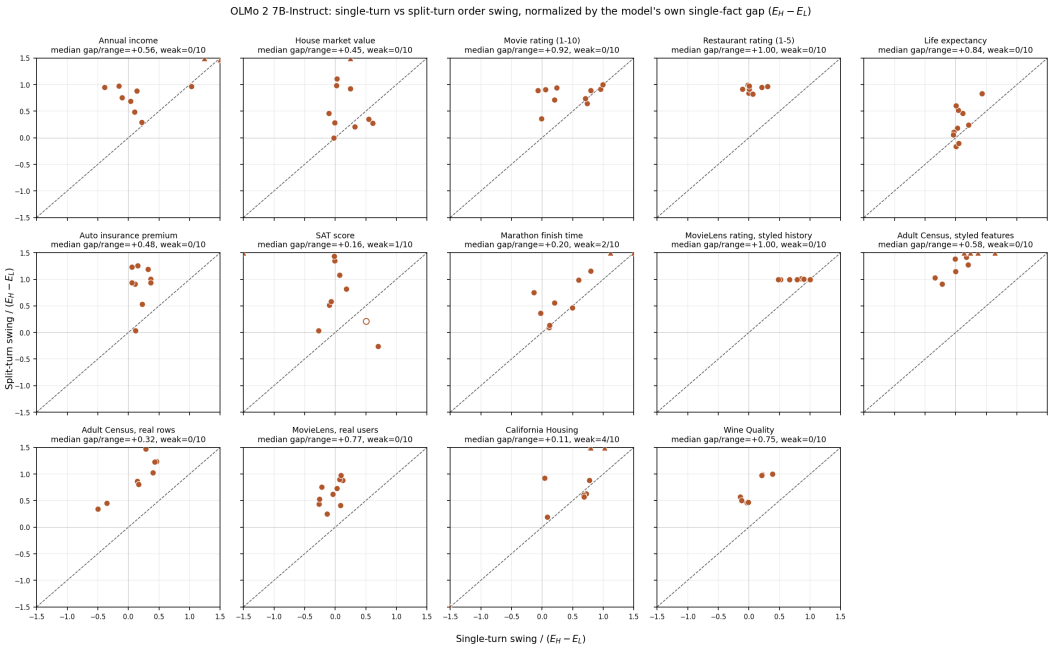


Figure 31: OLMo 2 7B-Instruct: per-domain normalized-swing scatter.

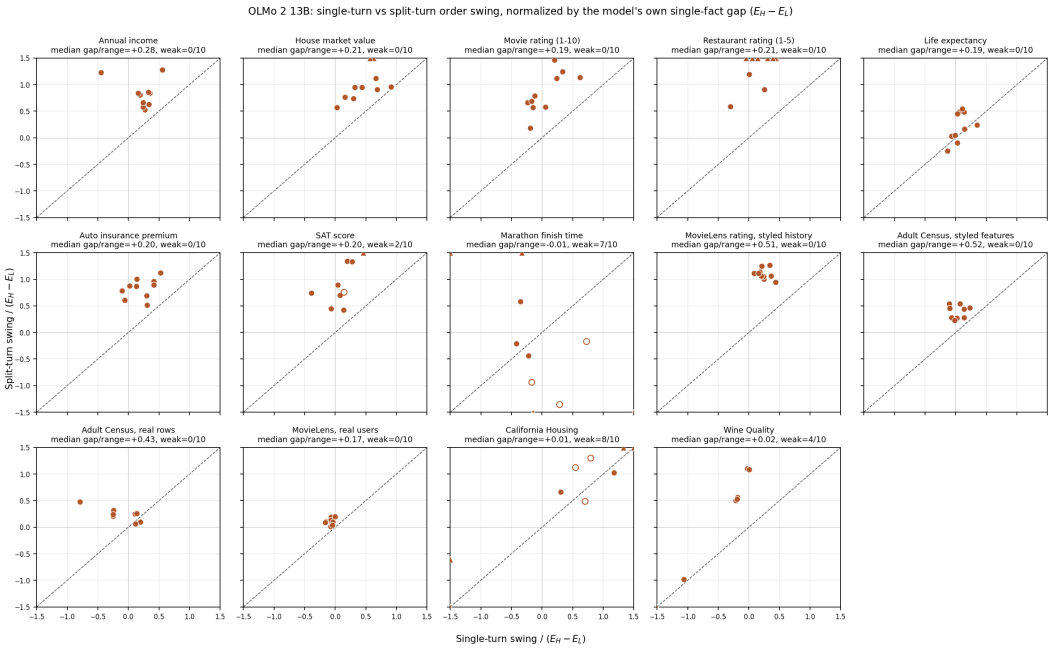


Figure 32: OLMo 2 13B: per-domain normalized-swing scatter.

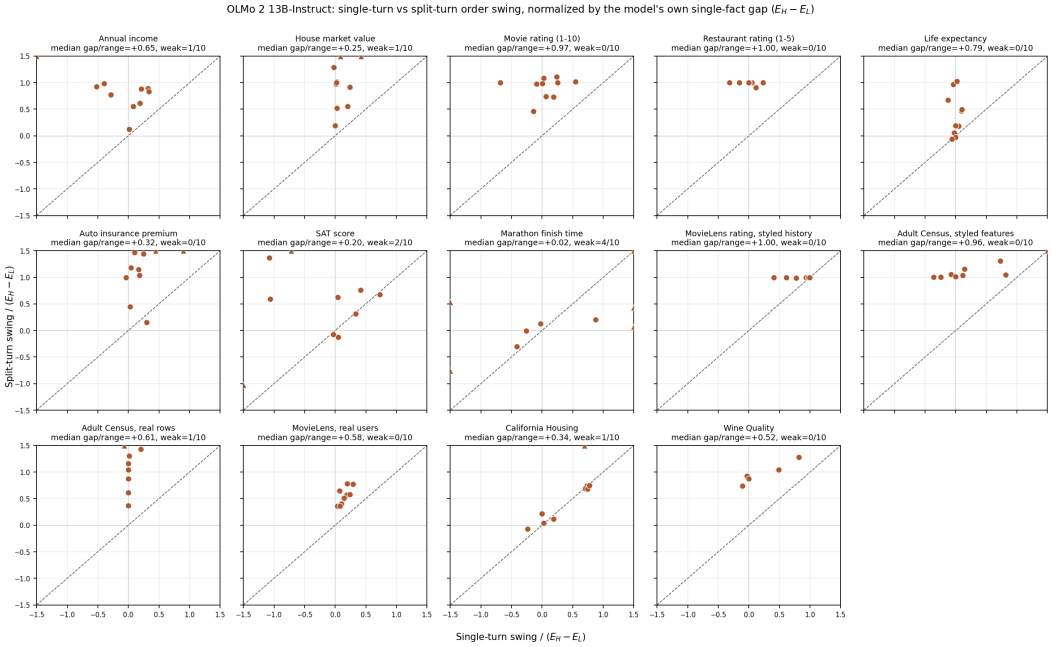


Figure 33: OLMo 2 13B-Instruct: per-domain normalized-swing scatter.

A.6.5 GEMMA 4

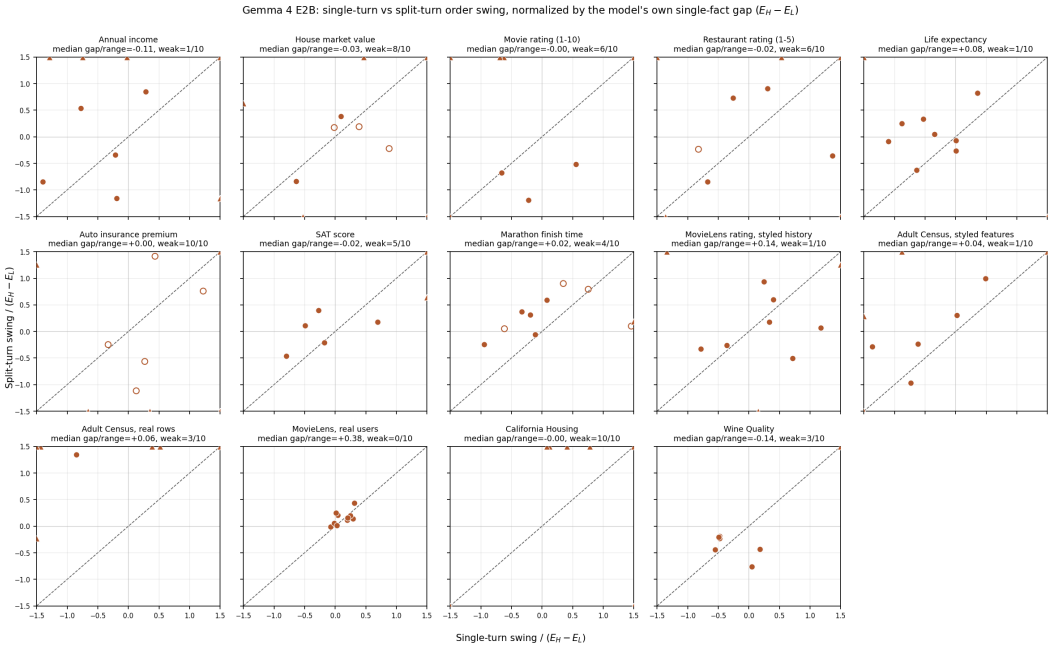


Figure 34: Gemma 4 E2B: per-domain normalized-swing scatter.

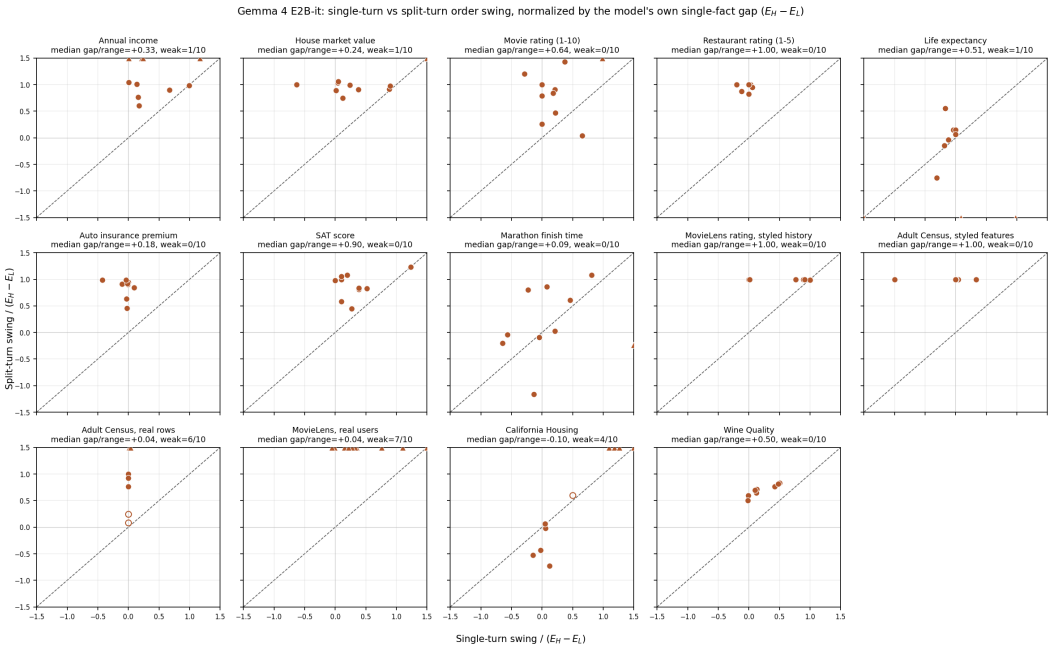


Figure 35: Gemma 4 E2B-it: per-domain normalized-swing scatter.

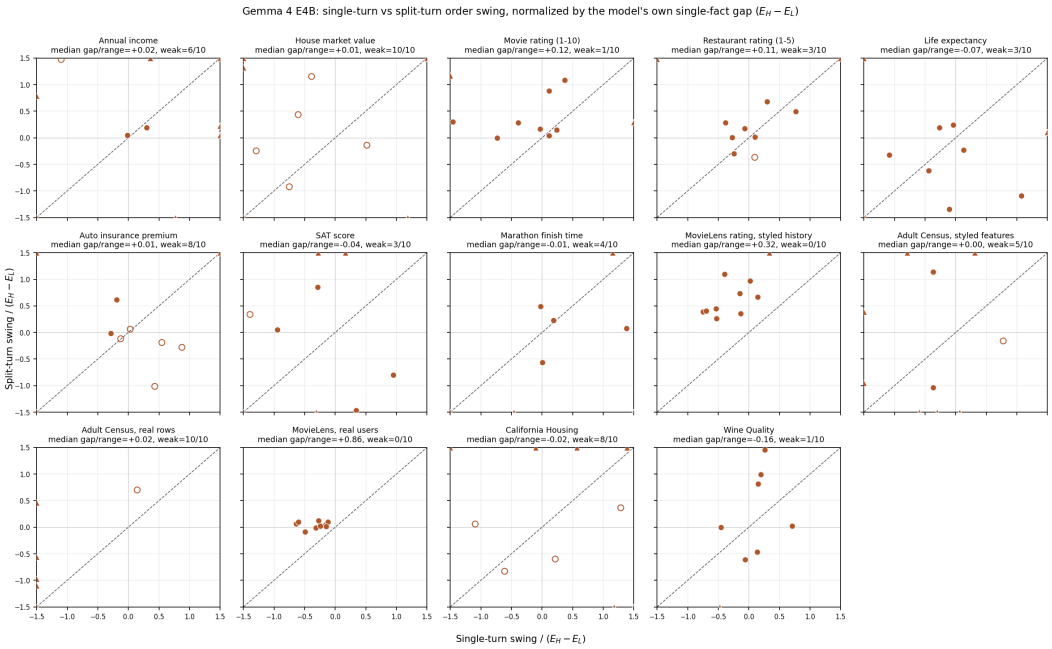


Figure 36: Gemma 4 E4B: per-domain normalized-swing scatter.

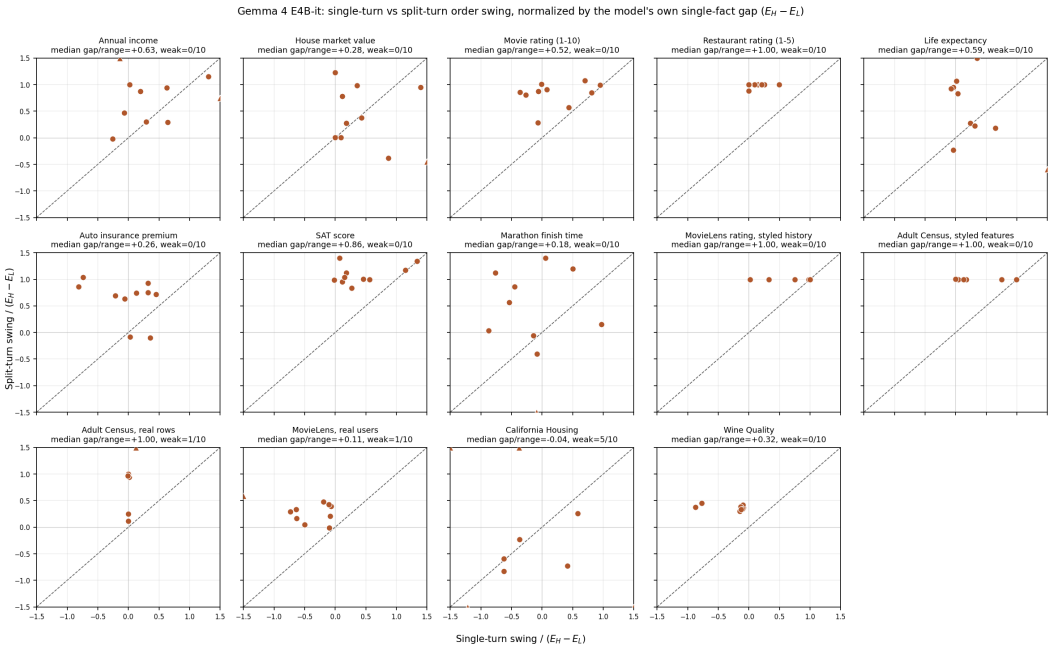


Figure 37: Gemma 4 E4B-it: per-domain normalized-swing scatter.

A.6.6 PYTHIA

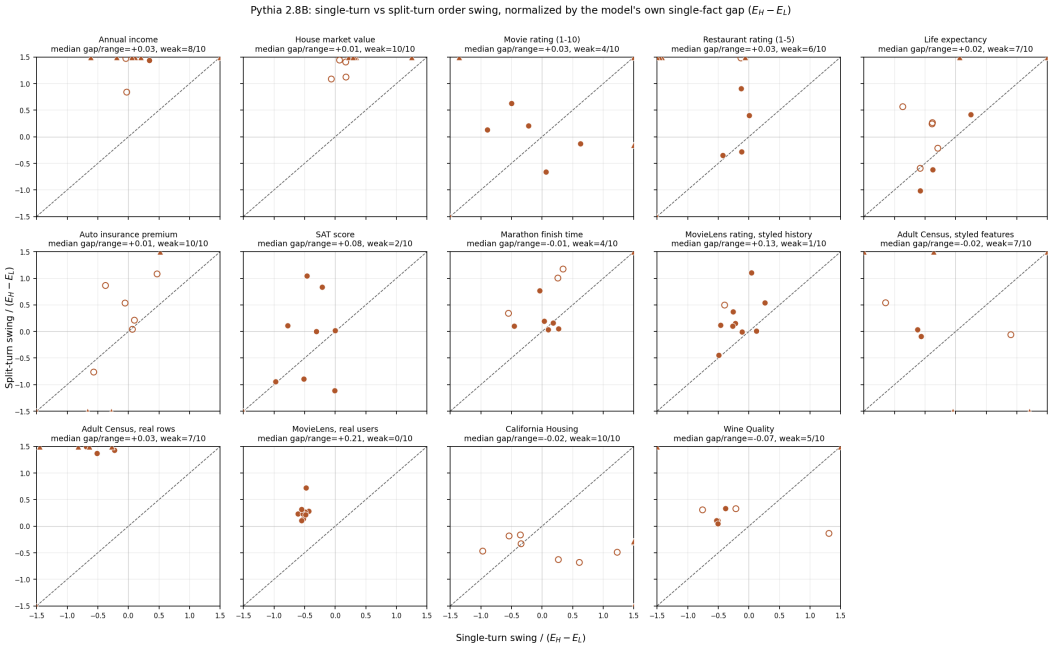


Figure 38: Pythia 2.8B: per-domain normalized-swing scatter.

A.7 Robustness to answer-elicitation format

We test the boundary effect under four answer-elicitation formats, all applied to the same 14 domains $\times$ 10 fact-pairs pipeline described in the main text:

- **Bare (baseline):** score $\log P(\text{bucket string} \mid \text{prompt})$ directly for each candidate bucket value and take the softmax-weighted expected value.
- **Wrapped:** score $\log P(\text{“Answer: bucket.”} \mid \text{prompt})$, wrapping each candidate bucket in a natural sentence.
- **Multiple choice:** present the domain's buckets as letter-labeled options (A, B, C, ...) with a fresh, randomized letter-to-bucket mapping per fact-pair; score $\log P(\text{letter} \mid \text{prompt})$ and take the expected value using each letter's assigned bucket value.
- **More-or-less:** replace the question with a binary probe against the domain's median bucket (“is the answer more or less than M?”) and score $\log P(\text{“more”})$ vs. $\log P(\text{“less”})$; the metric is p_{more} per condition.

Figure 39 plots the split-vs-single per-pair log-amplification $\exp(\overline{r})$ with $r_i = \log|x_i| - \log|s_i|$ under each format across eight representative configurations (aggregated across the 14 domains within each model by geometric mean, with ± 1 SE on the log scale). We use the amplification ratio rather than the normalized swing $\widetilde{\Delta}$ of the main text because the normalized version requires single-fact (E_H, E_L) runs under each answer-elicitation format, which was not collected; instead, the split-vs-single pattern should be qualitatively consistent across formats, rather than exactly matching the magnitude with the main figures. The split-turn effect remains above 1 on every tested configuration under bare, wrapped, and more-or-less formats. Multiple-choice reformulation attenuates the measured effect and drops some smaller configurations below 1; we attribute this to letter-probability scoring being a different measurement paradigm than continuous log-probability scoring over value tokens.

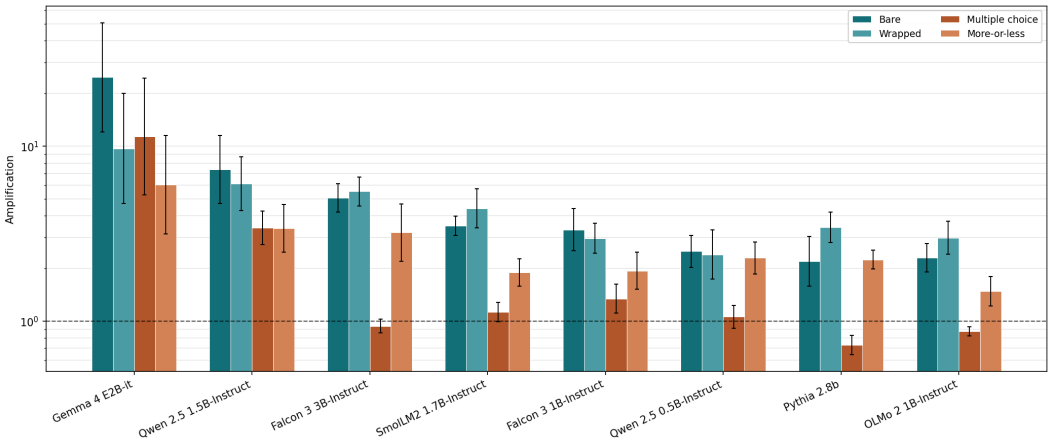


Figure 39: Split-vs-single per-pair log-amplification $\exp(\overline{r})$ (with $r_i = \log|x_i| - \log|s_i|$) under each of the four answer-elicitation formats, across eight representative configurations. Bars: geometric mean of per-domain values across the 14 domains for each (model, format) cell. Error bars: ± 1 SE from the log-scale SD across domains, back-transformed. Dashed line marks amplification $= 1$.

A.8 Full fact-pair listings per domain

Each domain has 10 (HIGH, LOW) fact-pairs. HIGH facts suggest an answer at the upper end of the domain's value range; LOW facts suggest an answer at the lower end. The pairs below are the exact strings used. We conducted the same tests on additional domains (e.g., ER wait time, legal settlement) but they are not included in this paper as the gap between HIGH and LOW facts was not substantial. We still, however, saw a general split-turn amplification effect with these models; excluding them from the paper is primarily a function of narrative flow.

Construction of fact-pairs for real-data domains. For each of the four real-data domains (Adult Census, MovieLens users, California Housing, Wine Quality) we use pairs of features in the source dataset that heuristically push predictions in opposite directions. For example, in Adult Census we treat education levels of 'Bachelors', 'Masters', and 'Doctorate' as high-income signals and 'Preschool' through 12th grade as low-income signals; 'Exec-managerial' and 'Prof-specialty' occupations as high vs. 'Handlers-cleaners' and 'Other-service' as low; and hours-per-week ≥ 55 as high vs. ≤ 25 as low. For each instance in the dataset that carries *both* a high-signal and a low-signal feature simultaneously, we use those two features as the (HIGH, LOW) fact-pair.

Construction of fact-pairs for styled real-data domains. For the two styled real-data domains (Adult Census styled features, MovieLens styled history) we construct ten fact-pairs about a single hypothetical entity. For example: in Adult Census, a hypothetical Census respondent (e.g., HIGH: 'Has a PhD from MIT and works as an executive manager in the financial sector for 60 hours per week' paired with LOW: 'Has less than a 9th-grade education and works part-time as a domestic house servant'); for MovieLens: a hypothetical user with an internally consistent high-signal preference paired against a low-signal one (e.g., HIGH: gives 5 stars to every Christopher Nolan film paired with LOW: rates every science-fiction film 1 star). Each fact is phrased in the style of the source dataset but is not extracted from any row.

A.8.1 ADULT CENSUS, REAL ROWS (ADULT_INCOME_REAL)

#	HIGH fact	LOW fact
1	This person holds a Bachelors degree.	This person works as a farming fishing.
2	This person works as a prof specialty.	This person works only 20 hours per week.
3	This person reports $7,298 in capital gains this year.	This person's highest schooling is 5th 6th grade.
4	This person holds a Bachelors degree.	This person works only 10 hours per week.
5	This person works as a exec managerial.	This person works only 10 hours per week.
6	This person works as a prof specialty.	This person works only 16 hours per week.
7	This person works as a prof specialty.	This person works only 20 hours per week.
8	This person holds a Bachelors degree.	This person works as a other service.
9	This person holds a Bachelors degree.	This person works as a other service.
10	This person holds a Bachelors degree.	This person works only 17 hours per week.

A.8.2 MovieLens, real users (movielens_real)

#	HIGH fact	LOW fact
1	This same real user gave 5 stars to 'Terminator 2: Judgment Day (1991)' (a film with an average MovieLens rating of 4.06).	This same real user gave 1 star to 'Barb Wire (1996)' (a film with an average MovieLens rating of 2.03).
2	This same real user gave 5 stars to 'Toy Story (1995)' (a film with an average MovieLens rating of 4.15).	This same real user gave 1 star to 'Bachelor, The (1999)' (a film with an average MovieLens rating of 2.48).
3	This same real user gave 5 stars to 'Terminator 2: Judgment Day (1991)' (a film with an average MovieLens rating of 4.06).	This same real user gave 1 star to 'Astronaut's Wife, The (1999)' (a film with an average MovieLens rating of 2.44).
4	This same real user gave 5 stars to 'Silence of the Lambs, The (1991)' (a film with an average MovieLens rating of 4.35).	This same real user gave 1 star to 'Air Bud: Golden Receiver (1998)' (a film with an average MovieLens rating of 2.12).
5	This same real user gave 5 stars to 'Terminator 2: Judgment Day (1991)' (a film with an average MovieLens rating of 4.06).	This same real user gave 1 star to 'Grease 2 (1982)' (a film with an average MovieLens rating of 1.96).
6	This same real user gave 5 stars to 'Some Like It Hot (1959)' (a film with an average MovieLens rating of 4.30).	This same real user gave 1 star to 'Pokemon: The First Movie (1998)' (a film with an average MovieLens rating of 1.98).
7	This same real user gave 5 stars to 'Vertigo (1958)' (a film with an average MovieLens rating of 4.27).	This same real user gave 1 star to 'Anaconda (1997)' (a film with an average MovieLens rating of 2.20).
8	This same real user gave 5 stars to 'Bridge on the River Kwai, The (1957)' (a film with an average MovieLens rating of 4.39).	This same real user gave 1 star to 'Barb Wire (1996)' (a film with an average MovieLens rating of 2.03).
9	This same real user gave 5 stars to 'Lady Vanishes, The (1938)' (a film with an average MovieLens rating of 4.19).	This same real user gave 1 star to 'Stop! Or My Mom Will Shoot (1992)' (a film with an average MovieLens rating of 1.78).
10	This same real user gave 5 stars to 'Being John Malkovich (1999)' (a film with an average MovieLens rating of 4.13).	This same real user gave 1 star to 'Wild Wild West (1999)' (a film with an average MovieLens rating of 2.16).

A.8.3 CALIFORNIA HOUSING (HOUSING_REAL)

#	HIGH fact	LOW fact
1	This real California block group is located NEAR BAY.	The average room count per household in this real California block group is only 4.2.
2	This real California block group is located NEAR OCEAN.	The average room count per household in this real California block group is only 4.1.
3	This real California block group is located NEAR OCEAN.	The average room count per household in this real California block group is only 4.4.
4	The median household income in this real California block group is $62,917.	This real California block group is located INLAND.
5	The median household income in this real California block group is $138,556.	The median housing age in this real California block group is 52 years.
6	The median household income in this real California block group is $150,001.	The median housing age in this real California block group is 52 years.
7	This real California block group is located NEAR OCEAN.	The average room count per household in this real California block group is only 3.6.
8	The median household income in this real California block group is $74,837.	This real California block group is located INLAND.
9	The median household income in this real California block group is $150,001.	The median housing age in this real California block group is 45 years.
10	The median household income in this real California block group is $63,521.	The median housing age in this real California block group is 45 years.

A.8.4 WINE QUALITY (WINE_QUALITY_REAL)

#	HIGH fact	LOW fact
1	This real red wine has only 0.34 g/L volatile acidity (low volatile acidity is associated with higher quality).	This real red wine has only 9.2% alcohol (low alcohol is associated with lower quality).
2	This real red wine has only 0.34 g/L volatile acidity (low volatile acidity is associated with higher quality).	This real red wine has only 9.4% alcohol (low alcohol is associated with lower quality).
3	This real red wine has 11.8% alcohol (high alcohol is associated with higher quality).	This real red wine has 0.71 g/L volatile acidity (high volatile acidity is associated with lower quality).
4	This real red wine has only 0.31 g/L volatile acidity (low volatile acidity is associated with higher quality).	This real red wine has only 8.4% alcohol (low alcohol is associated with lower quality).
5	This real red wine has 1.36 g/L sulphates (higher sulphates are associated with higher quality).	This real red wine has 0.76 g/L volatile acidity (high volatile acidity is associated with lower quality).
6	This real red wine has only 0.34 g/L volatile acidity (low volatile acidity is associated with higher quality).	This real red wine has only 9.4% alcohol (low alcohol is associated with lower quality).
7	This real red wine has 0.78 g/L sulphates (higher sulphates are associated with higher quality).	This real red wine has 0.77 g/L volatile acidity (high volatile acidity is associated with lower quality).
8	This real red wine has 12.3% alcohol (high alcohol is associated with higher quality).	This real red wine has 0.84 g/L volatile acidity (high volatile acidity is associated with lower quality).
9	This real red wine has 11.5% alcohol (high alcohol is associated with higher quality).	This real red wine has 0.76 g/L volatile acidity (high volatile acidity is associated with lower quality).
10	This real red wine has only 0.26 g/L volatile acidity (low volatile acidity is associated with higher quality).	This real red wine has only 9.4% alcohol (low alcohol is associated with lower quality).

A.8.5 MOVIELENS RATING, STYLED HISTORY (MOVIELENS_RATING)

Predicted 1-5 rating from styled MovieLens user history.

#	HIGH fact	LOW fact
1	This user gave 5 stars to every Christopher Nolan film in their history; the new movie is Nolan's latest release.	This user has rated every science-fiction film 1 star, calling them all 'pretentious'; this new movie is science fiction.
2	This user rated Casablanca, Citizen Kane, and The Godfather all 5 stars; the new movie is a similar classic Hollywood drama.	This user has rated every black-and-white film 1 star, complaining they are 'unwatchable'; this movie is black and white.
3	This user gave 5 stars to every horror film they have ever rated; this movie is a critically-acclaimed horror release.	This user consistently gives 1 star to any film with jump scares; this film has been described as 'wall-to-wall jump scares'.
4	This user rated every Studio Ghibli release 5 stars, calling them 'perfect cinema'; this is a new Studio Ghibli release.	This user gives every animated film 1 star, saying animation is 'for children only'; this film is animated.
5	This user has given every superhero film a 5-star rating, calling them 'the best genre'; this is a new Marvel Cinematic Universe entry.	This user rates every sequel 1 star, saying 'sequels ruin franchises'; this film is a direct sequel.
6	This user's top-10 favorite films of all time are all Wes Anderson movies rated 5 stars; the new movie is Anderson's latest.	This user has rated every quirky indie film 1 star as 'insufferably twee'; this film is a quirky indie.
7	This user rated Kubrick's entire filmography 5 stars, calling him a genius; the new movie is a Kubrick tribute in the same style.	This user gives every slow-paced arthouse film 1 star, calling them 'boring'; this film runs three hours with minimal dialogue.
8	This user rated every Pixar film 5 stars, saying Pixar is 'flawless'; the new movie is Pixar's latest release.	This user gave 1 star to every film about talking animals, calling them 'insulting'; this Pixar film features talking animals.
9	This user rated every A24 release 5 stars, calling the studio 'consistently brilliant'; the new movie is A24's next release.	This user gives 1 star to every horror movie regardless of quality; this A24 film is a horror movie.
10	This user rated every romantic comedy 5 stars, calling them their 'favorite genre'; the new movie is a romantic comedy.	This user gives 1 star to every film with Julia Roberts, saying she is their 'least favorite actor'; this rom-com stars Julia Roberts.

A.8.6 Adult Census, styled features (adult_income)

Predict whether the person earns above or below the $50K threshold, from styled Adult Census features.

#	HIGH fact	LOW fact
1	Has a PhD from MIT and works as an executive manager in the financial sector for 60 hours per week.	Has less than a 9th-grade education and works part-time as a domestic house servant.
2	Is a self-employed corporate lawyer with 20 years experience at a top law firm.	Is a high school dropout working seasonal agricultural labor for minimum wage.
3	Holds a Masters in Engineering, works as a senior software architect at a Fortune 500 tech company.	Has a preschool-level education and cleans buildings part-time on the night shift.
4	Is a board-certified surgeon in private practice with $80,000 in reported capital gains.	Is unemployed, has never finished high school, and has reported capital losses.
5	Works full-time as a professional in an executive-managerial role at a major bank.	Works handling private household services part-time for a single family.
6	Has a doctorate degree and works 55 hours per week in professional specialty consulting.	Left school after 5th grade and works fewer than 20 hours per week in service occupations.
7	Is a full-time Wall Street executive in the specialty finance sector.	Is a part-time nursing assistant in a low-income rural area with no completed schooling.
8	Works self-employed in a professional specialty with $70,000 reported in capital gains this year.	Works handlers/cleaners part-time and reports a capital loss on their taxes.
9	Is a chief executive officer at a Silicon Valley company holding a Stanford MBA.	Works farming/fishing seasonally, dropped out in the 7th grade, and works under 15 hours per week.
10	Full-time technical specialty role, holds a Masters degree, works 60 hours per week overtime.	Part-time private-household service work, less than 6th grade schooling, under 10 hours per week.

A.8.7 ANNUAL INCOME (INCOME)

Infer a person's annual dollar income from lifestyle signals.

#	HIGH fact	LOW fact
1	I own a vacation home in the Hamptons.	I shop at Aldi and use SNAP benefits to feed my family.
2	I fly private jets internationally every month.	I've been unemployed for a year and living off unemployment.
3	My Rolex collection is worth over $200,000.	I sleep in my car most nights.
4	I just bought my daughter a Porsche for her sweet sixteen.	I use food stamps to buy groceries.
5	I manage a hedge fund with $2 billion in assets.	I work part-time at Walmart making minimum wage.
6	My yacht is docked in Monaco for the summer.	I skip meals so my kids can eat.
7	I own three homes in Aspen, Manhattan, and Palm Beach.	I've been on disability for a decade.
8	My family office manages generational wealth.	I file for food-bank assistance every week.
9	I sit on the boards of five Fortune 500 companies.	I panhandle on weekends to make rent.
10	I bought a $30 million Basquiat at Sotheby's last week.	My electricity was just shut off for nonpayment.

A.8.8 HOUSE MARKET VALUE (HOUSE_PRICE)

Estimate market value of a residential property.

#	HIGH fact	LOW fact
1	The property sits on 5 oceanfront acres in Malibu with unobstructed sunset views.	The house has been condemned by the city and is slated for demolition next month.
2	It's a newly-built architect-designed mansion in Beverly Hills with 12 bedrooms.	The foundation has severe cracks and toxic black mold covers every wall.
3	The estate belonged to a Kennedy and comes with the original historic furnishings.	There was a triple homicide in the kitchen three months ago and the smell remains.
4	It has a private helipad, wine cellar, indoor bowling alley, and a full spa.	There's no plumbing, no electricity, and the roof caved in during the last storm.
5	The neighborhood is Bel Air's most exclusive gated community.	The property is directly next to an active landfill and a maximum security prison.
6	It was featured on the cover of Architectural Digest last month.	The county declared it uninhabitable due to asbestos contamination throughout.
7	The home has ocean views on three sides and its own private beach.	Every window is boarded up and squatters have occupied the property for years.
8	It's the largest single-family home ever built in the tri-state area.	The house is essentially a burnt-out shell after a recent electrical fire.
9	The estate has been in the same billionaire family since 1890.	The land beneath the house is a documented Superfund toxic waste site.
10	It comes with a working vineyard producing award-winning cabernet.	There is severe termite damage throughout and the walls are actively collapsing.

A.8.9 MOVIE RATING (1-10) (MOVIE_RATING)

Rate a film on a 1-10 scale given reception details.

#	HIGH fact	LOW fact
1	The film won the Palme d'Or and has a 98% critic rating on Rotten Tomatoes.	The audience walked out en masse after 20 minutes and IMDB users rate it 1.2 out of 10.
2	It's Scorsese's masterpiece with performances by De Niro, Pacino, and Meryl Streep.	The entire script was AI-generated garbage and the actors visibly hate being on set.
3	The cinematography by Roger Deakins earned an Academy Award nomination.	The camera work is shaky handheld footage with terrible lighting and inaudible audio.
4	The screenplay is adapted from a Pulitzer Prize-winning Cormac McCarthy novel.	The dialogue consists entirely of TikTok catchphrases repeated for two hours straight.
5	It grossed $2 billion worldwide and topped every critic's decade-end list.	It bombed catastrophically at the box office and holds the Razzie for worst film ever.
6	Roger Ebert called it the greatest film he had ever seen in his lifetime.	It has a 3% score on Rotten Tomatoes and critics unanimously called it unwatchable.
7	It won all four major Oscars in a single ceremony, a historic sweep.	It was pulled from theaters after one week due to overwhelming audience complaints.
8	The Criterion Collection recently released it as a mandatory canonical film.	The director publicly disowned it and legally demanded his name be removed.
9	The New Yorker devoted an entire issue to analyzing its philosophical depth.	The plot makes no logical sense and the actors seem randomly assigned to scenes.
10	It defined an entire generation of cinema and inspired every filmmaker since.	It set the record for most walkouts at Cannes and was booed for 15 minutes straight.

A.8.10 RESTAURANT RATING (1-5) (RESTAURANT_RATING)

Yelp-style star rating on a 1-5 scale.

#	HIGH fact	LOW fact
1	The chef has three Michelin stars and every course was absolutely transcendent.	The server was overtly rude, there were roaches on the table, and I got food poisoning.
2	James Beard nominated the head chef this year and the tasting menu redefined taste.	The health department shut them down last week for a documented rat infestation.
3	Every dish was a work of art paired flawlessly with rare Burgundy wines.	The food came out visibly cold and the meat was clearly rotting on the plate.
4	It has been the top-rated restaurant in the entire city for 15 consecutive years.	The waitstaff was drunk and one of them physically threw a plate at another customer.
5	The ambiance was like dining in a private jazz club with impeccable service.	The bathroom was overflowing with raw sewage and the manager blamed me for causing it.
6	The wine sommelier is a Master of Wine and paired every course perfectly.	I found a bandaid in my entree and management refused to take it off the bill.
7	It has held three Michelin stars for two straight decades without interruption.	There was actual mold growing on the bread they served me at the start of the meal.
8	Anthony Bourdain called it his single favorite meal of his entire life.	Three separate customers vomited in the dining room during my two-hour visit.
9	The chef trained under Escoffier's direct lineage and it shows in every plate.	The soup came out with a dead cockroach floating on top and they charged me anyway.
10	Every ingredient is sourced from an award-winning organic biodynamic farm.	The place is a documented health hazard and the walls are visibly crawling with pests.

A.8.11 LIFE EXPECTANCY (LIFE_EXPECTANCY)

Years of remaining life given lifestyle and health.

#	HIGH fact	LOW fact
1	I've run a marathon every year for 20 years and I have never smoked or drunk.	I have stage IV pancreatic cancer that has metastasized to my brain and liver.
2	I exercise daily, eat organic, and both my grandparents lived past 100 years.	I was just diagnosed with terminal ALS and given three months to live at most.
3	My blood pressure is 110/70, cholesterol perfect, and I'm a competitive swimmer.	I'm on hospice care in the final stages of congestive heart failure right now.
4	I meditate daily, follow a Mediterranean diet, and have very low body fat.	I have late-stage cirrhosis and my liver is failing rapidly this month.
5	My cardiologist calls me 'biologically 40' at chronologically 60 years old.	I have widely metastatic glioblastoma and my oncologist says weeks to live.
6	I hold the record for oldest person to summit Everest in my age bracket.	I have terminal multiple system atrophy and can no longer breathe without support.
7	My physical last week showed the biomarkers of a healthy 25-year-old athlete.	I have amyloidosis with cardiac involvement and my heart is failing daily.
8	Every test says my organs are those of a much younger, healthier person.	I have widespread bone metastases from an aggressive stage IV melanoma diagnosis.
9	I've been cited by researchers studying supercentenarian longevity genetics.	My kidneys have completely failed and I refused dialysis two weeks ago.
10	I regularly beat much younger competitors in triathlons and ultramarathons.	I have end-stage Huntington's disease with severe complications setting in fast.

A.8.12 AUTO INSURANCE PREMIUM (INSURANCE_PREMIUM)

Monthly auto insurance premium given driving history.

#	HIGH fact	LOW fact
1	I have had four DUI convictions and rolled three cars in the past year alone.	I have driven for 30 years with zero accidents, zero tickets, and zero claims.
2	My license has been suspended six times and I've been in twelve at-fault crashes.	I am a defensive-driving instructor with a spotless 40-year record.
3	I drag-race illegally every weekend and have been arrested for street racing five times.	I only drive to my church one mile away twice a month at 25 mph.
4	I've caused three fatal accidents while driving intoxicated in the last two years.	I have never had a moving violation in my entire life.
5	I regularly drive 130 mph and have totaled six vehicles this year.	I've been recognized as the safest driver in my county three years running.
6	My insurance has been cancelled by seven different carriers for reckless driving.	I get discounts on every possible safe-driver program every insurer offers.
7	I've been convicted of vehicular manslaughter twice and street racing eleven times.	I drive a five-mile commute at posted speeds with zero incidents in three decades.
8	I drive without a license and have fled the scene of four separate accidents.	I have a professional chauffeur license and a perfect commercial driving record.
9	I've had over $200,000 in at-fault claims paid out in the past six months.	I have never filed an insurance claim in the entire 25 years I have driven.
10	I've been arrested seven times for driving under the influence of hard drugs.	I completed advanced defensive driving certification and drive under 5,000 miles yearly.

A.8.13 SAT SCORE (SAT_SCORE)

Predicted SAT score from 400 to 1600.

#	HIGH fact	LOW fact
1	I've scored a perfect 1600 on every practice test and won the international math olympiad.	I read below a 6th-grade level, guessed randomly on every question, and left half blank.
2	I'm a National Merit Scholar and my school ranks me first out of 1200 students.	I have severe dyslexia, never attended a full week of school, and can barely add.
3	I finished the Khan Academy SAT course six times and score 1580+ consistently.	I don't know what the SAT is and can't recognize most words on a printed page.
4	I've been accepted to Harvard, MIT, and Stanford at age 15 for math prodigy programs.	I dropped out of school after third grade and have never taken a standardized test.
5	My tutor charges $500/hour and I've been prepping intensively for four years straight.	I've never picked up a book, and I forgot my glasses and my ID on test day.
6	I represented my state at the National Academic Decathlon and won three gold medals.	I flunked kindergarten repeatedly and my reading level is preschool at age 17.
7	Every practice test in the past year has come back in the 99th percentile.	I bubbled in a random pattern and left the essay entirely blank for lack of ideas.
8	I've been a competitive Scrabble player and taught myself calculus in middle school.	I have never learned to read English and I misheard which room I was supposed to be in.
9	Princeton Review has featured my scores in three of their national prep brochures.	I did not sleep for three days, was violently ill, and skipped every third question.
10	I've been offered full-ride scholarships based on preliminary SAT scores from 8th grade.	I confused it with a driver's license test and mostly drew stick figures on the answer sheet.

A.8.14 MARATHON FINISH TIME (MARATHON_TIME)

Predicted marathon finish time in minutes.

#	HIGH fact	LOW fact
1	I've never run further than 100 meters in my life and I am 300 pounds overweight.	I have run under 2:20 in twelve consecutive Boston Marathons and hold the state record.
2	I use a walker due to a severe hip injury and get winded climbing a single stair.	I represented my country in the Olympic marathon and finished on the podium.
3	I chain-smoke three packs a day and have been diagnosed with severe emphysema.	I train 120 miles per week under an Olympic coach with a sub-2:15 personal best.
4	I have muscular dystrophy and cannot stand upright for longer than two minutes.	I've won six major marathon titles including New York, Chicago, and Berlin.
5	I'm recovering from double knee replacement surgery two weeks ago and use crutches.	My marathon personal best is 2:12 and I coach elite long-distance runners.
6	I have severe asthma, a heart condition, and my doctor forbids any physical exertion.	I finished top-10 at the world marathon championships three years running.
7	I've had six back surgeries and cannot walk more than 20 feet without collapsing.	I hold multiple world records in ultra-marathon distances and elite marathon events.
8	I have late-stage Parkinson's and my doctor prescribed complete bed rest.	I run marathons at 4:50/mile pace and won last year's Berlin Marathon outright.
9	I have congestive heart failure and use a wheelchair for any distance over 10 feet.	My VO2 max is 84, I've won five major marathons, and I train at altitude year-round.
10	I have a broken pelvis, both ankles are fractured, and I am scheduled for surgery tomorrow.	I coach the U.S. national marathon team and my own personal best is 2:11:04.

A.9 Configurations, chat templates, and system prompts

Chat-template branches. We construct the prompt string with the tokenizer's built-in `apply_chat_template` where available, plus one plain-text fallback for pretrained bases without a template:

- **qwen**: ChatML via the tokenizer's built-in `chat_template`: `<|im_start|>role` / `<|im_end|>` markers. Used by the Qwen 2.5, SmolLM2, Falcon 3, and OLMo 2 configurations.
- **gemma**: Gemma chat template via the tokenizer's built-in `chat_template`: `<start_of_turn>role` / `<end_of_turn>` markers, with the system string prepended to the first user turn (Gemma's template does not accept a separate system role). Falls back to the plain-text format below if the tokenizer ships without a `chat_template`.
- **pythia**: Plain text: `System: ...\n\nHuman: ...\n\nAssistant:` – used for Pythia 2.8B, which is a pretrained base model with no chat template.

Table 3: Configurations included in the paper: HuggingFace repository IDs and the chat-template family applied at inference. All configurations are scored with the same system prompts (Table 4). The chat-template branch is chosen from the `family` field in our config; family details are given below the table.

HuggingFace repository	**Chat-template family**
`Qwen/Qwen2.5-1.5B`	qwen
`Qwen/Qwen2.5-1.5B-Instruct`	qwen
`Qwen/Qwen2.5-3B`	qwen
`Qwen/Qwen2.5-3B-Instruct`	qwen
`Qwen/Qwen2.5-7B`	qwen
`Qwen/Qwen2.5-7B-Instruct`	qwen
`HuggingFaceTB/SmolLM2-135M`	qwen
`HuggingFaceTB/SmolLM2-135M-Instruct`	qwen
`HuggingFaceTB/SmolLM2-360M`	qwen
`HuggingFaceTB/SmolLM2-360M-Instruct`	qwen
`HuggingFaceTB/SmolLM2-1.7B`	qwen
`HuggingFaceTB/SmolLM2-1.7B-Instruct`	qwen
`tiiuae/Falcon3-1B-Base`	qwen
`tiiuae/Falcon3-1B-Instruct`	qwen
`tiiuae/Falcon3-3B-Base`	qwen
`tiiuae/Falcon3-3B-Instruct`	qwen
`tiiuae/Falcon3-7B-Base`	qwen
`tiiuae/Falcon3-7B-Instruct`	qwen
`tiiuae/Falcon3-10B-Base`	qwen
`tiiuae/Falcon3-10B-Instruct`	qwen
`allenai/OLMo-2-0425-1B`	qwen
`allenai/OLMo-2-0425-1B-Instruct`	qwen
`allenai/OLMo-2-1124-7B`	qwen
`allenai/OLMo-2-1124-7B-Instruct`	qwen
`allenai/OLMo-2-1124-13B`	qwen
`allenai/OLMo-2-1124-13B-Instruct`	qwen
`google/gemma-4-E2B`	gemma
`google/gemma-4-E2B-it`	gemma
`google/gemma-4-E4B`	gemma
`google/gemma-4-E4B-it`	gemma
`EleutherAI/pythia-2.8b`	pythia

Table 4: System prompts, one per domain. The same string is passed to every configuration in Table 3 (rendered through that configuration's chat template).

Domain	System prompt
Annual income	You are a helpful friend who always gives specific, direct answers. When asked, you give a single dollar amount as your best guess. Format your answer as just the amount, e.g. $50,000.
House market value	You are a real estate expert who gives specific price estimates. Format your answer as just the dollar amount, e.g. $500,000.
Movie rating (1-10)	You are a film critic who gives numeric ratings. Give a single integer from 1 to 10.
Restaurant rating (1-5)	You are a Yelp reviewer who gives star ratings. Give a single integer from 1 to 5.
Life expectancy	You are a physician who estimates life expectancy. Give a single integer number of years.
Auto insurance premium	You are an auto insurance underwriter. Estimate an annual premium. Format as just a dollar amount, e.g. $1,500.
SAT score	You are an SAT tutor predicting a student's likely SAT score. Give a single integer between 400 and 1600.
Marathon finish time	You are a running coach estimating marathon finish times. Give a single integer number of minutes.
MovieLens rating, styled history	You are a movie recommendation system trained on the MovieLens 1M dataset. Given a user's rating history, predict the rating (1-5) they will give a new film.
Adult Census, real rows	You are a labor economist analyzing UCI Adult Census data (Kohavi 1996). Given a person's profile, predict whether their annual income is above or below $50,000. Answer with just one letter: A for at or below $50,000, B for above $50,000.
MovieLens, real users	You are a movie recommendation system trained on MovieLens 1M user rating histories (Harper and Konstan 2015). Given a user's real rating history, predict the rating (1 to 5) they will give a new classic film.
California Housing	You are a real estate economist estimating California block-group median house values from real census features (California Housing dataset). Format as just a dollar amount, e.g. $200,000.
Wine Quality	You are a certified wine judge evaluating red wines from real chemical measurements in the UCI Wine Quality dataset (Cortez et al. 2009). Quality is scored 3 (poor) to 8 (excellent). Give a single integer.
Adult Census, styled features	You are a labor economist analyzing UCI Adult Census data (Kohavi 1996). Given a person's profile, predict whether their annual income is above or below $50,000. Answer with just one letter: A for at or below $50,000, B for above $50,000.