Memory vs. Context? Influential Factors of Factual Recall in Language Models

Guilhem Fouilhé
IRIT
ANITI
g.g.a.fouilhe@gmail.com

Nicholas Asher
IRIT
CNRS
nicholas.asher@irit.fr

Philippe Muller
IRIT
Université de Toulouse
philippe.muller@irit.fr

Abstract

We reproduce and stress-test the work of Yu et al. (2023), who characterize how language models (LMs) arbitrate between memorized knowledge and contradictory in-context statements. We replicate their world-capitals experiments on 31 models spanning Pythia, GPT-2, Qwen3, and Ministral families, including base and post-trained variants, and extend evaluations to five additional knowledge relation types from the ParaConflict dataset. We empirically confirm most of their original findings: larger models and higher-frequency entities tend to favor memorized answers, with substantial family-level variance. However, several conclusions do not generalize cleanly: entity-frequency effects disappear on Qwen3-14B and 32B; post-training shifts the memory-context trade-off inconsistently across families; question phrasing alone can change a model's reliance on memorized knowledge by up to 80 percentage points; and semantically unrelated prose can mimic coherent supporting context. Our results clarify where Yu et al.'s claims hold and to what extent they generalize to other prompts.[1]

1 Introduction

Language Models (LMs) based on the Transformer architecture (Vaswani et al., 2017) show evidence of factual recall through memorization of the training data (Meng et al., 2022), meaning they can be prompted to output factual knowledge. For instance, a generative model prompted with "*The capital of Hungary is*" is likely to predict *Budapest* as the next token(s). This capacity scales with model size (Allen-Zhu and Li, 2025) and topic frequency (Smith et al., 2026) (the frequency of a fact in the training data).

This capacity, however, sometimes conflicts with another core capacity of LMs, *In-Context Learning* or ICL (Brown et al., 2020), which enables models to use information from a prompt's context to make predictions. This feature is particularly useful because facts can change. For instance, the statement *The Argentina men's national football team are the world champions* was true between 2022 and 2026, but became obsolete after the 2026 final, causing models' memorized knowledge to be outdated. Model prompting can be used to explicitly introduce new facts (e.g. adding *Spain defeated Argentina in the 2026 World Cup final* at the beginning of the prompt). In such cases, it is usually desirable that models give priority to the context over memorized knowledge. But it is not always the case: we also want models to be robust against false information in prompt injections. Understanding how models behave in these situations is important for understanding their limitations and improving their robustness.

Yu et al. (2023) show that different models (ranging from Pythia-70M to Pythia-2.8B and GPT-2 to GPT-2-XL) have different ways of handling prompts of the form *"The capital of Poland is London. Q: What is the capital of Poland? A:"*, designed to minimally introduce a conflict between memorized knowledge and context. The mechanism each model uses depends both on the frequency of the subject of the fact (the country, e.g., Poland) and the frequency of the overwriting information (the counterfactual city, e.g., London): as the frequency goes up, models are more likely to predict the memorized answer.

However, Yu et al. (2023) do not clarify how these results depend on the model architecture, size, and stage of training. Moreover, they establish the observed effects only on a single template and a single specific dataset (world-capitals).

These limitations provide the motivation for our paper: we re-evaluate those experiments on more models, varying prompt-design factors and testing the robustness of the results on more varied datasets.

[1] Our code is available at https://github.com/gfouilhe/IC-factual-public

Proceedings of the 9th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, pages 1029–1046
October 29, 2026 ©2026 Association for Computational Linguistics

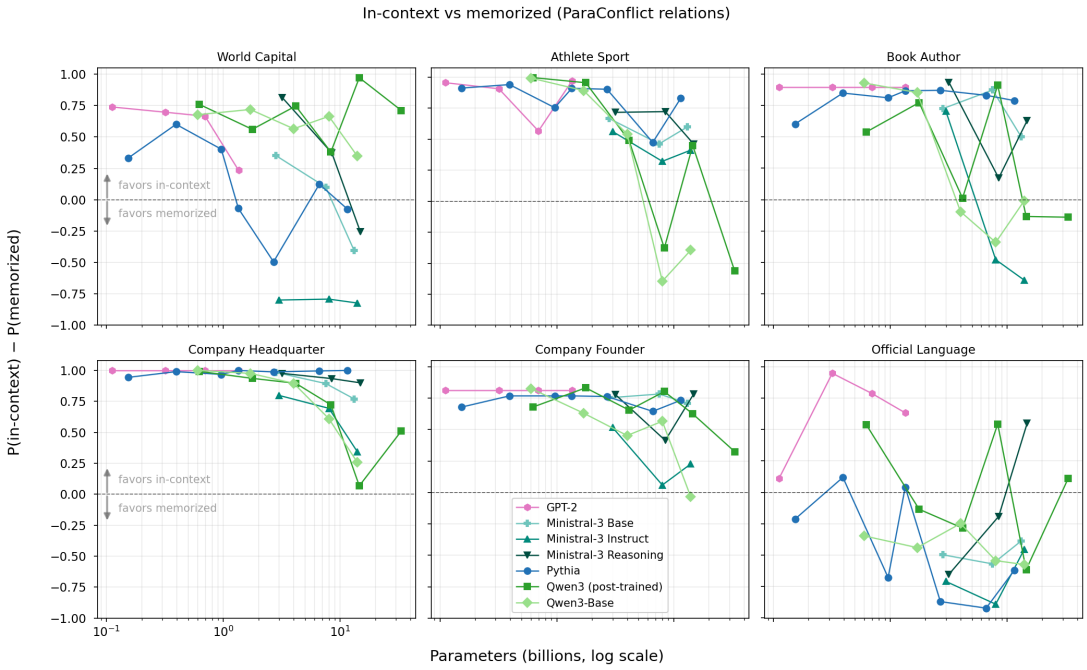

Figure 1: Comparison of several models' behavior with context/memorized knowledge conflict across parameter counts. Scores are differences in probability between in-context information and memorized information, averaged over instances of 6 ParaConflict relations. Positive scores means the model chooses in-context responses more often.

We establish Yu et al. (2023)'s result that larger models and higher-frequency entities tend to favor memorized answers, with substantial family-level variance on a broader range of topics and across 31 models. We also show that: topic frequency effects vary across some models; post-training shifts the memory-context trade-off inconsistently across model families; question phrasing in the prompt can substantially alter a model's reliance on memorized knowledge, as can the use of semantically unrelated prose.

2 Related Work

Yu et al. (2023) characterize how LMs arbitrate between memorized knowledge and in-context contradictions; Li et al. (2025) extend this line of work with the ParaConflict benchmark and a mechanistic analysis of attention heads. They find that "context heads" and "memory heads" can switch roles when the degree of conflict changes slightly, so Yu et al. (2023)'s observation that some heads promote context or memorized knowledge does not generalize well and intervening on a single head may not robustly steer models toward either source. They also show that multiple interventions are not necessarily additive, and propose a method to improve intervention robustness.

Longpre et al. (2021) formalize the contextual–parametric knowledge conflict and provide a framework for evaluating model behavior; the setup of Yu et al. (2023) is essentially what they call *corpus substitution*. They also propose a mitigation technique based on continued training.

Xie et al. (2024) introduce a framework to systematically construct counterevidence to the parametric memory of LMs. They find that model behavior depends strongly on coherently presented in-prompt evidence.

The context–memory conflict we study is one kind of *knowledge conflict*. Xu et al. (2024) taxonomize knowledge conflicts into three categories: context–memory conflict (context contradicting parametric knowledge), *inter-context* conflict (contradictions among different pieces of context, e.g. noisy or misinformed retrieved documents in RAG settings), and *intra-memory* conflict (inconsistencies within the model's own parametric knowledge, surfacing as divergent answers to differently phrased queries). While we focus on the first category, the analysis we provide could also benefit the study of other types of conflicts.

3 Methods

3.1 Task formulation

We investigate LM behavior when prompted with a *substitution conflict*: a counterfactual statement

followed by a query for the same fact. We use the template from Yu et al. (2023) unless noted otherwise:

```
The capital of {country} is {distractor}.
Q: What is the capital of {country}? A:
```

For each prompt the model must choose between the memorized capital c_{mem} and the in-context distractor c_{ctx}. We report $P(\text{in-context}) = P(c_{ctx} \mid prompt)$ and $P(\text{memorized}) = P(c_{mem} \mid prompt)$.

3.2 Experimental design

Datasets We build prompts from the ParaConflict Dataset (Li et al., 2025). The WORLD CAPITAL category is essentially equivalent to the dataset used by Yu et al. (2023), but only provides one distractor per country. We extend it following Yu et al. (2023): for every country with gold capital c_{mem} and every other country's capital c'_{mem}, we set $\{distractor\} := c'_{mem}$ in the template. We omit pairs where c'_{mem} is also a correct answer for the queried country. This yields n=47,299 prompts covering 218 countries.

We also study five additional ParaConflict categories (ATHLETE SPORT (n=532), BOOK AUTHOR (n=500), COMPANY HEADQUARTER (n=500), COMPANY FOUNDER (n=203), OFFICIAL LANGUAGE (n=193)) to test relation specificity. However, the main focus of this paper is on the WORLD CAPITAL category, because it is the most studied in prior work.

Context-length and filler tokens. Yu et al. (2023) focused on substitution conflicts of the form `conflict|question` which fails to capture the wide range of contexts in which these conflicts can arise. To investigate the effect of adding coherent context before the question, ParaConflict provides a `Coherent Conflict` variant: a fluent passage of ≈120 tokens that elaborates the counterfactual claim. Across all 31 models we evaluate, replacing the substitution template with this coherent passage raises $P(\text{in-context})$ on every model (median $\Delta = +0.19$; range $+0.01$–$+0.57$; Appendix G, Table 5). In the same direction, coherent supporting context can strongly steer models toward the in-context answer (Xie et al., 2024).

To isolate this effect, we insert a configurable filler *between* the conflict statement and the question while holding the conflict and query fixed. Filler length is measured in *tokens* under the model's tokenizer and tiled or truncated to the target length. Unless stated otherwise, the conflict sits in *tail* position (`[conflict] [filler] [question]`, i.e., separated from the question by filler); we also test *front* (`[filler] [conflict] [question]`, immediately before the question) and *middle* (`[½filler] [conflict] [½filler] [question]`) placements at selected lengths. We use the following filler types:

- **prose** — a fixed, AI-generated English paragraph (plant-cell biology) with no country or capital names;
- **capital_noise** — per-prompt sequences of true statements `The capital of X is Y.` sampled from the country pool, excluding the queried country and the distractor's country;
- **capital_noise_false** — same surface form as **capital_noise** but with shuffled (false) country–capital pairings. Queried country's capital and distractor capital are also excluded;
- **wiki_*** — Wikipedia country descriptions (for the queried `country`, the `distractor_country` whose actual capital is the distractor, or an `unrelated_country`) with the relevant capital redacted;

We truncate context lengths to $L \in \{0, 32, 128, 512, 1024\}$ tokens as closely as possible while maintaining correct sentence structure.

Template and conflict ablations. We vary the question template: `qa` template from (Yu et al., 2023) vs. `bare` continuation (`"The capital of {country} is"`), plus `possessive` (`"{country}'s capital is"`), `learned` (`"I learned that the capital of {country} is"`), and `of_course` (`"Of course, the capital of {country} is"`) variants; the conflict template: `standard` vs. `negated`: `"...is not {distractor}."`; and the number of conflict copies $K \in \{1, 2, 4, 8\}$ at L=0. Full results for the alternative question templates are in Appendix K; negated-conflict and conflict-copy ablations are in Appendices L and M.

3.3 Models

Pythia. We evaluate the full EleutherAI Pythia suite (160M–12B) (Biderman et al., 2023), including 6.9B and 12B models beyond the sizes studied

in Yu et al. (2023). All Pythia models are trained on 300B tokens.

GPT-2. We include the OpenAI GPT-2 series (Radford et al., 2019): `small` (124M), `medium` (355M), `large` (774M), and `xl` (1.5B). The training data size is not disclosed.

Qwen3. We also include Qwen3 (Yang et al., 2025) dense models at 0.6B, 1.7B, 4B, 8B, 14B, and 32B (post-trained, instruction-tuned models) and matching Qwen3-Base pretrained models from 0.6B through 14B (no public 32B-Base release). Comparing base vs. post-trained isolates the effect of post-training from pretraining scale and architecture. All Qwen3 models are reportedly trained on 36T tokens.

Ministral-3. We evaluate Mistral's Ministral-3 (Liu et al., 2026) dense models at 3B, 8B, and 14B in three post-training stages: `Base`, `Instruct`, and `Reasoning`. Ministral-3 models are reportedly trained on 1T to 3T tokens.

3.4 Answer classification

For each prompt we label a response as **memorized** if the model favors the gold capital and **in-context** if it favors the distractor.

Generative classifier. Following Yu et al. (2023), we decode up to 10 new tokens greedily and classify by substring match against gold aliases and the distractor.

Log-probability classifier. We verified the correlation of the generative classifier with a more direct evaluation that scores the two candidates by log-probability of the continuation `{city}` (leading space included) given the prompt:

$$\hat{y} = \arg\max_{y \in \{c_{mem}, c_{ctx}\}} \log P(y \mid \text{prompt}),$$

where c_{mem} is the memorized capital and c_{ctx} the distractor. Unlike the 10-token generative classifier, this teacher-forced sequence score evaluates the complete target continuation and is immune to answer truncation and conversational scratchpad preambles (Appendix N). The correlation between the generative and log-probability classifiers is high and discussed in Appendix H.

3.5 Training-frequency proxy

Yu et al. (2023) relate memorization rates to entities' frequency in the Pile (Gao et al., 2020), Pythia's 825 GB pretraining corpus. We approximate entity frequency with the `wordfreq`[2] package's Zipf score on a mixed web/Wikipedia/news corpus. This proxy has the advantage of simplicity of reproduction and allows a fair comparison between models with open-access and private training data. Scores are computed per country (subject) and per distractor country; leading articles (`the`) are stripped so that `the United Kingdom` and `France` are comparable. Full per-entity Zipf scores for all six ParaConflict categories appear in Appendix J. We bin subjects by Zipf and plot $P(\text{memorized})$ vs. $P(\text{in-context})$ by bin, replicating the paper's Figure 3 and parts of Figure 2.

3.6 Mechanistic analysis: logit lens

To inspect when the preference between memorized and in-context answers becomes directly decodable in the residual stream, we apply the *logit lens* (Nostalgebraist, 2020): for each prompt we run a forward pass, project every transformer block output through the model's final normalization and vocabulary unembedding, and read off the log-probability of the *first token* of the memorized capital and of the distractor at the last sequence position. We aggregate mean per-layer gaps $\Delta_\ell = \log P(c_{ctx}) - \log P(c_{mem})$ over n=500 prompts per condition on all models for the World Capital task.

4 Results

4.1 Model sizes and memorization

While Yu et al. (2023) do not state this explicitly, their results suggest that bigger models tend to use memorized knowledge more than smaller models. We studied this phenomenon on all ParaConflict relation categories (Figure 1) and measured the correlation between model size and $P(\text{mem})$ (Table 1; by-family breakdown in Appendix), and the effect of post-training on $P(\text{mem})$ (Table 2; per-model breakdown in Appendix). Overall, we find that model size is significantly positively correlated with $P(\text{mem})$, despite high variance across model families and post-training stages. Post-training can either increase or decrease $P(\text{mem})$ depending on the model family and the relation category, with Qwen3 post-training generally decreasing $P(\text{mem})$ and Ministral-3 Instruct post-training generally increasing it. Thus, post-training seems to have inconsistent effects across model families.

[2] https://github.com/rspeer/wordfreq

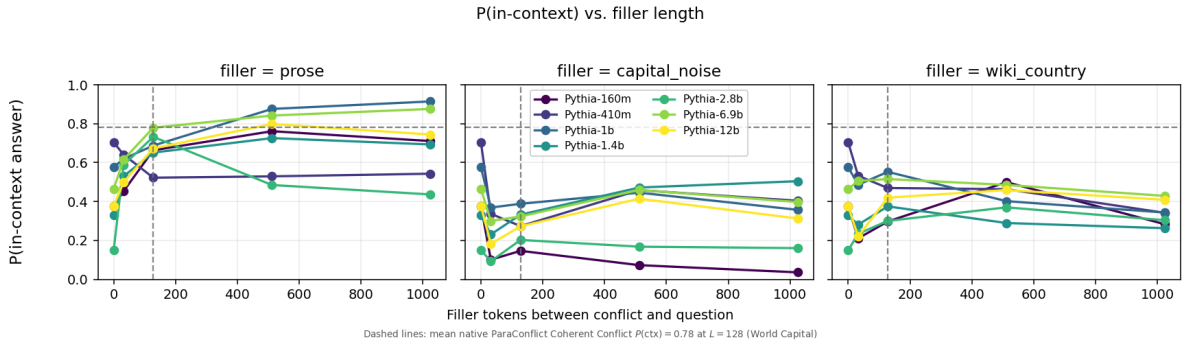

Figure 2: P(in-context) vs. filler length for different fillers across Pythia sizes.

To determine whether low P(mem) under conflict reflects in-context reliance or simply an absence of memorized knowledge, we evaluate a baseline of factual recall without conflict using the bare continuation prompt (P(correct | clean prompt); e.g., *"The capital of X is"*, *"X plays the sport of"*; Figures 4 and 5 in Appendix A). Across all 31 models and all six relations, baseline factual recall scales monotonically with parameter size (e.g., Pythia overall accuracy scales from 13 % at 160M to 53 % at 12B; Qwen3-Base from 28 % at 0.6B to 64 % at 14B; Ministral-3 Base reaches 78 %). Importantly, this clean baseline reveals that the near-zero memorization rates observed under substitution conflict for certain relations (e.g., Book Author and Company Founder) in smaller models stem from an absence of parametric storage (0 % recall on Book Author for Pythia $\leq$ 2.8B and GPT-2) rather than purely robust context following. Conversely, Official Language and World Capital have high baseline recall ($\geq$85 % for models $\geq$1.4B), confirming that their behavior under conflict represents a genuine competition between memorized knowledge and in-context distractors.

Table 1: Correlation of P(mem) with model size, pooled across all 31 models. We report the mean memorization rate $\bar{P}$(mem), Spearman ρ between parameter count and P(mem), and two-sided p-values. Positive ρ means larger models memorize more than smaller ones. n is number of relation instances of each category.

Category	n	$\bar{P}$(mem)	ρ	p
World Capital	47,299	0.30	+0.48	0.006
Athlete Sport	532	0.22	+0.80	<0.001
Book Author	500	0.18	+0.81	<0.001
Company Headquarter	500	0.07	+0.77	<0.001
Company Founder	203	0.07	+0.81	<0.001
Official Language	193	0.58	+0.43	0.016
All (pooled)		0.24	+0.53	<0.001

Table 2: Effect of post-training on P(mem). Positive ΔP(mem) means post-training increases memorization.

	ΔP(mem) vs. Base (pp)		
Category	Qwen3	Ministral-3 Instruct	Ministral-3 Reasoning
World Capital	−10.0	+41.4	−13.8
Athlete Sport	−11.8	+7.1	−3.0
Book Author	−11.4	+38.0	+4.5
Company Headquarter	+0.1	+12.1	−2.4
Company Founder	−11.1	+17.1	+2.8
Official Language	−31.8	+10.0	−18.1
All (pooled)	−12.7	+20.9	−5.0

4.2 Country frequency and memorization

Yu et al. (2023) argue that, when context and memory conflict, models should rely more on memorized answers for *high-frequency* queried countries, reflecting stronger factual storage for entities seen more often during pretraining. We test this hypothesis on the full world-capitals dataset across 31 models. For each model we correlate per-country `wordfreq` Zipf with P(mem) and confirm the trend with ten equal-count Zipf deciles. The hypothesis is supported at least partially: in 17 of 31 models, country-level Spearman $\rho \geq 0.3$ and the highest-frequency decile exceeds the lowest by at least 5 percentage points; Pythia (except 160M), GPT-2 medium–XL, and most Ministral variants follow the Yu et al. pattern. Counterexamples notably include the largest model tested, Qwen3-32B ($\rho \approx 0$). Further analyses are reported in Appendix I (Table 6 and Figure 9).

4.3 Context length and filler tokens

Figure 2 shows generative-classifier outcomes for three representative fillers (`prose`, `capital_noise`, `wiki_country`), context lengths $L \in \{0, 32, 128, 512, 1024\}$, and Pythia models

(160M–12B; n=2000 prompts per cell). For medium and large models (Pythia ≥1.4B), inserting unrelated `prose` filler between conflict and question sharply increases P(in-context) (e.g., from 29% at L=0 to 93% at L=128 for 2.8B), showing that even semantically inert text can mimic the effect of a coherent passage. In contrast, for smaller models (160M and 410M, the grey and green lines in the left panel), increasing filler length steadily dilutes context retention, pulling P(in-context) downward. Meanwhile, `capital_noise` and `wiki_country` suppress in-context answers across virtually all sizes as length grows. The full six-filler grid is in Appendix B. Evaluating L=128 unrelated prose across all 31 models (Appendix C, Figure 7) reveals that models outside Pythia exhibit inconsistent effects. While positive shifts appear on a few models (e.g., Qwen3-Base-4B: +16.4 pp; Ministral-3 Reasoning-8B: +22.8 pp), other families often move strongly in the opposite direction, with unrelated filler decreasing P(in-context) by up to −28.2 pp in Qwen3 and −38.8 pp in Ministral-3 as memorized priors re-assert themselves. Thus, unlike coherent supporting context which elevates P(in-context) universally (Table 5), the boost from unrelated prose is not consistent across model families. Conflict-position ablations (`front`/`middle`/`tail`) are in Appendix D. Question template ablations across all 31 models (Appendix K, Figure 11) further reveal that prompt framing alone can swing P(mem) by up to 95.0 pp (e.g., `qa` vs. `of_course` in Qwen3-14B), demonstrating that phrasing sensitivity is architecture-general across modern models.

4.4 Late-layer direct decodability: cross-family logit lens

Figure 3 tracks the separation between memorized and in-context answers across model families using the logit-lens protocol from §3.6. Across all model families, Δ_ℓ separates decisively only in the second half of the network. This aligns with observations by Meng et al. (2022), who showed via causal tracing that while subject representations are processed in middle MLP layers, the factual prediction is formed and transferred to the residual stream primarily in later layers.

5 Discussion

Our experiments support five main findings:

1. Model size is positively correlated with reliance on memorized knowledge, but the effect is non-linear and strongly family-dependent (e.g. it is stronger for Qwen3 than for the other families studied).

2. Subject frequency in the training data is positively correlated with P(mem) for most models, reinforcing Yu et al. (2023); however, the correlation vanishes for some models, notably Qwen3-14B and 32B.

3. Post-training shifts the memory–context trade-off substantially but inconsistently: Qwen3 post-training and Ministral-3 Reasoning generally decrease P(mem), while Ministral-3 Instruct increases it, with per-size reversals of up to 60 pp on individual relations.

4. The question template alone can change P(mem) by up to 95 pp at fixed model and conflict statement across all modern families tested (Figure 11).

5. A coherent passage supporting the context raises P(in-context) on every model tested; on Pythia, semantically unrelated prose of the same length has a comparable effect, but across other families unrelated prose exhibits inconsistent effects, sometimes shifting strongly in the opposite direction (Figure 7).

Taken together, findings 3–5 suggest that memory–context arbitration is not a stable, model-level property: within a single model, question phrasing, filler content, and conflict position can swing the outcome by tens of percentage points. Characterizations based on a single template and dataset therefore risk overstating generality.

Besides model specificity, Qwen3 counterexamples to the frequency effect admit two readings. Models trained on far more data may store even rare entities robustly enough that frequency ceases to discriminate, making the effect a signature of undertrained knowledge. Alternatively, our `wordfreq` proxy may diverge from the actual pretraining distribution—a limitation shared by any frequency analysis of closed-data models.

6 Conclusion

We reproduced and extended Yu et al. (2023)'s study of factual recall under substitution conflicts across 31 models, six ParaConflict relation types,

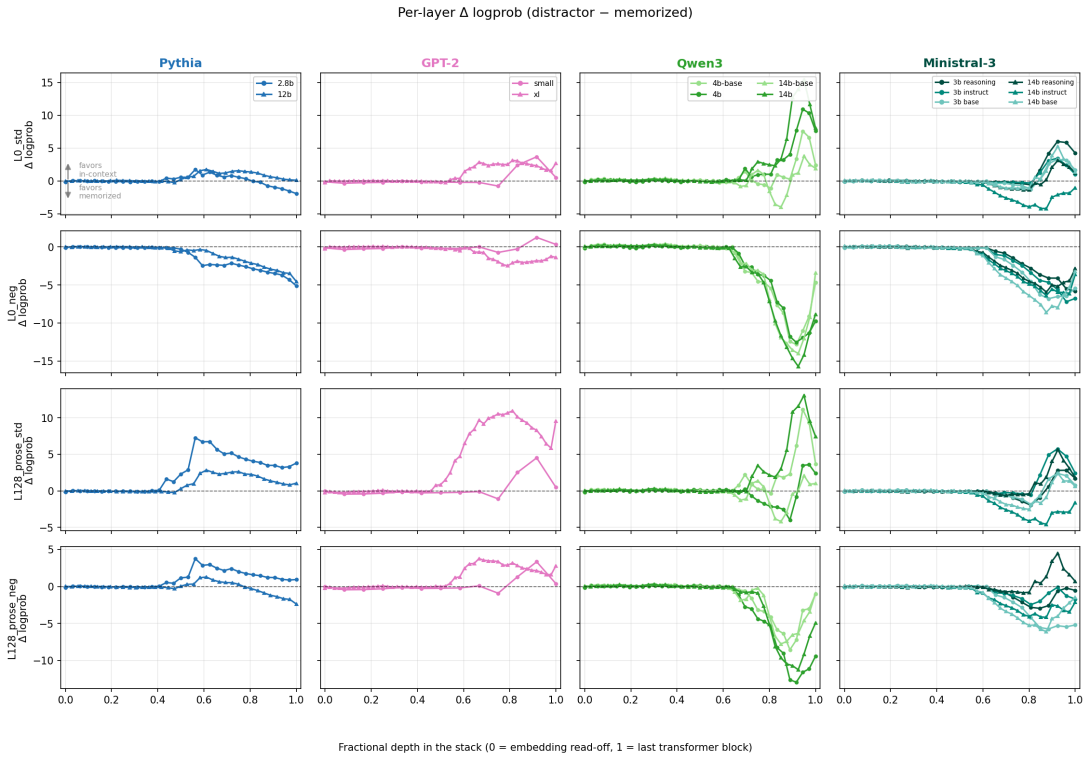

Figure 3: Logit lens on the world-capitals dataset, indicating at each layer if a model leans towards the in-context or the memorized answer by decoding hidden states into vocabulary space. Rows: four matched conditions (no filler vs. L=128 prose; standard vs. negated conflict).

and targeted prompt ablations. Most original trends hold at scale: larger models and higher-frequency entities tend to favor memorized answers. Nevertheless several conclusions are prompt- and family-sensitive: entity-frequency effects vanish on some Qwen3 sizes, post-training shifts the memory–context trade-off inconsistently, question phrasing alone can swing $P(\text{mem})$ by up to 95 percentage points, and semantically unrelated filler can mimic coherent supporting context. A cross-family logit-lens analysis reveals that preference for memorized versus in-context answers becomes directly decodable primarily in late network layers. Together, these findings clarify the boundaries of prior claims and establish prompt design as a decisive factor in memory–context arbitration. Extending our ablation framework to broader conflict benchmarks such as ConflictBank (Su et al., 2024) is a natural next step.

Limitations

Despite extending the original study, there are still restrictions to the present work, along multiple dimensions: (a) conflicts investigated here are limited to conflicts between in-context information and memorized knowledge by the model; internal model or context conflicts might be elicited in comparable ways or demand different methodologies. (b) the type of conflict is factual, assuming a universal stable ground truth, and ignore e.g. temporal aspects (world capitals can change, but are relatively stable in general, at least in the available written data that feed current model training.) (c) we show a lot of variance associated to certain factors, for instance phrasing of the conflict, but it remains unclear what general linguistic concepts are responsible for the variations. Finally, the approach is limited to models for which we can access prediction probabilities, so currently only open-weight models, and for practical reasons we capped the model sizes.

Acknowledgments

This work was supported by the Artificial and Natural Intelligence Toulouse Institute (ANITI). ANITI is funded by the France 2030 program under the Grant agreement ANR-23-IACL-0002. This research is also supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme. Computational resources were provided

by a French government grant managed by the Agence Nationale de la Recherche under the "Investissements d'avenir" program (reference "ANR-21-ESRE-0051").

References

Zeyuan Allen-Zhu and Yuanzhi Li. 2025. Physics of language models: Part 3.3, knowledge capacity scaling laws. In *International Conference on Learning Representations*, volume 2025, pages 14937–14946.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, and 1 others. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, and 1 others. 2020. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Gaotang Li, Yuzhong Chen, and Hanghang Tong. 2025. Taming knowledge conflicts in language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 34074–34104. PMLR.

Alexander H Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, and 1 others. 2026. Ministral 3. *arXiv preprint arXiv:2601.08584*.

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

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Nostalgebraist. 2020. Interpreting gpt: The logit lens. *Blog Post*.

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, and 7 others. 2022. In-context learning and induction heads. *Transformer Circuits Thread*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Matthew L Smith, Jonathan P Shock, Samuel T Segun, Iyiola E Olatunji, and Tegawendé F Bissyandé. 2026. Predictable confabulations: Factual recall by LLMs scales with model size and topic frequency. *arXiv preprint arXiv:2605.18732*.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024. Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in LLMs. In *Advances in Neural Information Processing Systems*, volume 37.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Qinan Yu, Jack Merullo, and Ellie Pavlick. 2023. Characterizing mechanisms for factual recall in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9924–9959, Singapore. Association for Computational Linguistics.

A Baseline Factual Recall Without Conflict

This appendix supplements §4.1. To determine whether low $P(\text{mem})$ under substitution conflict is driven by context-following preference or by lack of parametric knowledge, we evaluate zero-shot factual recall without conflict (n=2146 prompts per model across all six categories) using the bare continuation clean prompt from the ParaConflict benchmark. Figure 4 visualizes the scaling behavior across parameter counts for each relation category under generative substring evaluation, while Figure 5 visualizes the corresponding accuracy under the log-probability classifier ($\arg\max_{y \in \{c_{\text{mem}},\, c_{\text{ctx}}\}} \log P(y \mid \text{clean prompt})$).

Notably, although small models have a low or near-zero generative factual recall rate on some datasets (such as Book Author or Company Founder; Figure 4), Figure 5 shows that they still prefer the gold answer over a random distractor when the distractor does not appear in the prompt (accuracies remain $\geq 50\%$ across all models and categories). This indicates that the models' parametric representations favor the correct entity over an unprompted distractor.

B Context-length: all fillers

This appendix supplements Figure 2 in §4.3. Figure 6 shows the complete generative-classifier grid over all six filler types at $L \in \{0, 32, 128, 512, 1024\}$ (Pythia 160M–12B; n=2000 per cell). Surprisingly, the unrelated `prose` filler steer models towards the in-context answer, sometimes more than the coherent paragraph supporting the counterfactual claim (e.g for Pythia 1b and 6.7B with >512 filler tokens). Other fillers show mixed effects: a filler with text from the wikipedia page of the subject country, with the relevant capital redacted (`wiki_country`) don't appear to influence the in-context-memory arbitration; one with the page of the distractor country (`wiki_distractor_country`) pushes the model towards ICL, but not significantly more than one from an unrelated country (`wiki_unrelated_country`).

C Cross-family generalization of unrelated prose ($L = 128$)

To test whether the in-context boost from semantically unrelated prose observed on Pythia generalizes across other architectures, we evaluate all 31 models on n=2,000 World Capital queries comparing the standard zero-filler prompt (L=0) against an insertion of L=128 tokens of biology prose between the conflict statement and the question (tail position; identical to the setup in §4.3). Figure 7 plots the percentage-point change $\Delta P(\text{in-context}) = P(\text{in-context} \mid L{=}128) - P(\text{in-context} \mid L{=}0)$ for each model.

Across models outside Pythia, the effect is inconsistent:

- On Pythia, prose filler above 1B consistently increases in-context reliance (by up to $+63.8$ pp on 2.8B). A similar boost appears on a few other models, such as Qwen3-Base-4B ($+16.4$ pp), Qwen3-1.7B ($+18.2$ pp), and Ministral-3 Reasoning-8B ($+22.8$ pp).
- However, for many models across Qwen3 and Ministral-3, the effect moves strongly in the opposite direction: inserting unrelated prose reduces $P(\text{in-context})$ by up to -28.2 pp (Qwen3-4B) and -38.8 pp (Ministral-3 Base 3B), diluting the conflict statement and allowing the memorized prior to dominate.

Therefore, while coherent supporting context consistently raises $P(\text{in-context})$ across all 31 models (Finding 5, Table 5), unrelated prose does not produce a consistent boost outside Pythia.

D Conflict position: front, middle, and tail

This appendix supplements the context-length analysis in §4.3. The previous experiments holds the conflict in *tail* position (`conflict | filler | question`). Here we vary where the conflict statement sits relative to the filler and question at $L \in \{128, 512\}$ tokens, using the paper `qa` template, `prose` and `capital_noise` fillers, and the log-probability classifier (n=2,000 prompts per cell; Pythia 160M–6.9B). The three placements are:

- **tail** — `conflict | filler | question` (default elsewhere in the paper);
- **front** — `filler | conflict | question` (conflict immediately before the question);

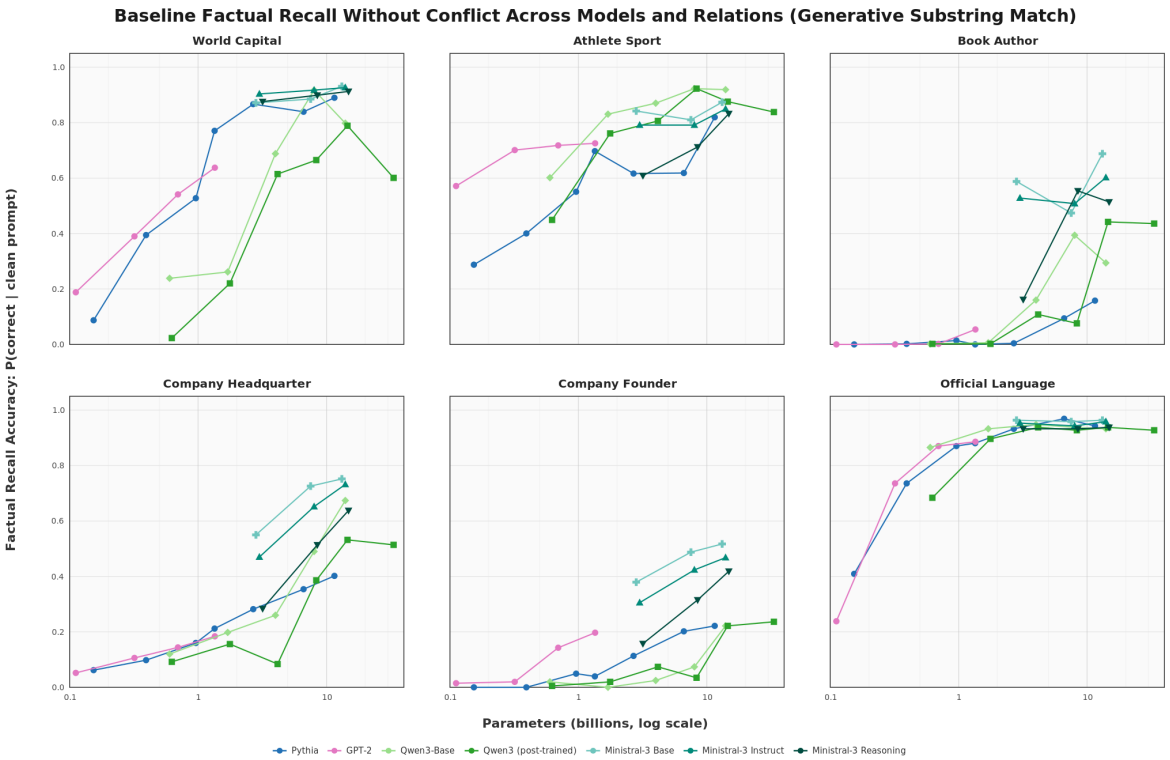

Figure 4: Baseline factual recall accuracy without conflict (P(correct | clean prompt); generative substring match) across model size for each ParaConflict relation category.

- **middle** — ½`filler` | `conflict` | ½`filler` | `question` (conflict embedded at the filler midpoint).

Figure 8 shows P(in-context) across all three positions. With `capital_noise`, placing the conflict in `middle` is typically the worst cell for in-context binding—a lost-in-the-middle-like pattern (Liu et al., 2024)—with competing capital statements on both sides (e.g. Pythia-1.4B at L=512: 7.8 % vs. 29.6 % at `tail`).

`Prose` behaves differently: inserting filler *between* conflict and question (`tail`) steers models toward in-context at medium-to-large scale, whereas placing the conflict immediately before the question (`front`) often restores memorization (Pythia-2.8B at L=128: 92.5 % → 40.7 %). This is hard to reconcile with a simple induction-head recency story (Olsson et al., 2022), which would predict stronger in-context use when the conflict is last before the query.

E Post-training effects by model size

This appendix supplements Table 2 in §4.1. Table 3 reports ΔP(mem) for each size-matched post-training contrast separately, without averaging across model scale. The Qwen3-8B vs. Qwen3-Base-8B flip on book authors (-59.8 pp) and the Ministral-3-8B Instruct gain on the same relation ($+61.6$ pp) illustrate that pooled post-training effects can hide sharp size-specific reversals.

F Model size vs. memorization by family

This appendix supplements Table 1 in §4.1. Table 4 shows the size–memorization Spearman correlation separately within each model family. The all-model trend in Table 1 masks family-specific structure: Qwen3-Base is uniformly positive ($\bar{\rho} = +0.85$), whereas Pythia inverts on company headquarters ($\rho = -0.56$) and GPT-2 has $P(\text{mem}) = 0$ on three relations at every size (undefined ρ because on these tasks it never predicts the memorized response).

G Effect of Coherent Conflict

Table 5 compares `Substitution Conflict` and `Coherent Conflict` prompts on all six ParaConflict relation categories (n=2146 per model; generative classifier). Replacing the minimal substitution template with the dataset's fluent coherent passage raises P(in-context) on every evaluated

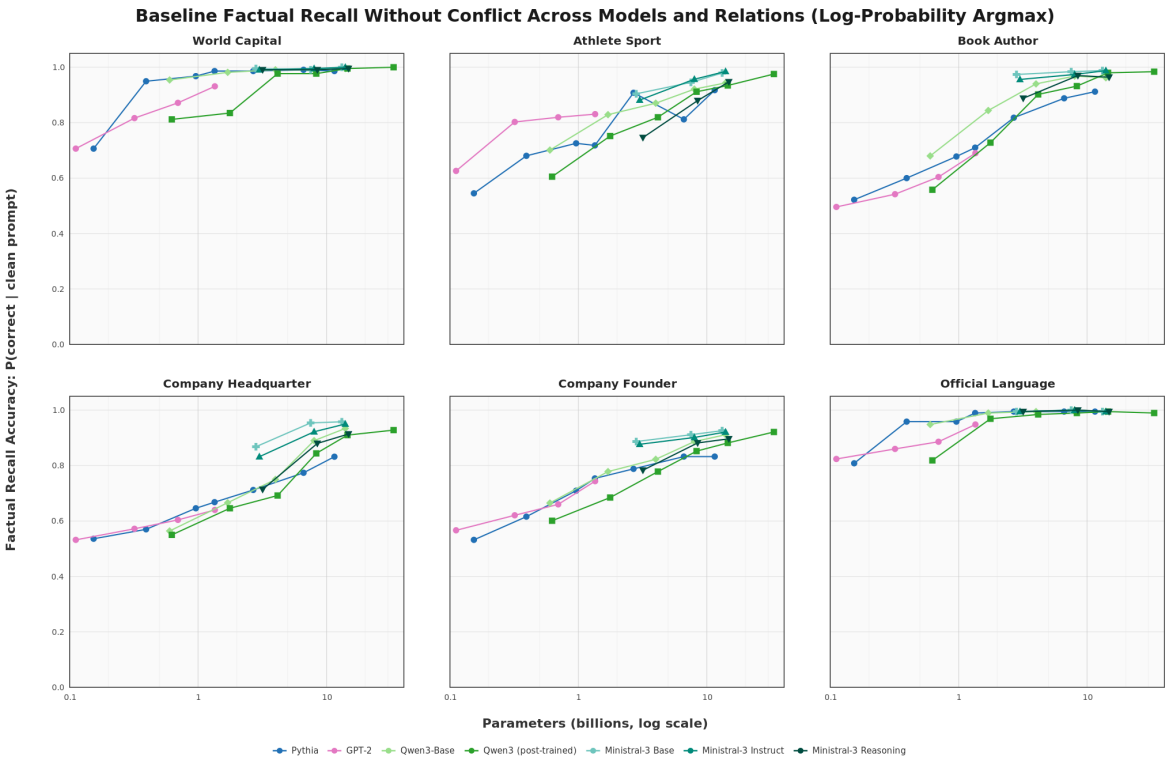

Figure 5: Baseline factual recall accuracy without conflict ($P(\text{correct} \mid \text{clean prompt})$; log-probability argmax) across model size for each ParaConflict relation category.

model ($\Delta > 0$ for all 31; median $\Delta = +0.19$; mean $\Delta = +0.21$; range $+0.01$–$+0.57$).

H Correlation Between Generative and Log-Probability Classifiers

The two classifiers' $P(\text{mem})$ estimates are very tightly correlated (Pearson $r \approx 0.98$); $P(\text{ctx})$ less so ($r \approx 0.73$), mainly because greedy decoding sometimes emits a third answer that the log-probability classifier cannot assign. With filler tokens in the prompt, the correlation tends to decrease more for $P(\text{ctx})$ due to the increased likelihood of generating a third answer.

I Country frequency and memorization: full results

This appendix supplements §4.2.

Table 6 lists all 31 models and associated statistics; Figure 9 visualizes the same data.

J ParaConflict entity word frequencies

This appendix supplements the training-frequency proxy in §3.5. For each of the six ParaConflict relation categories we collect unique *subjects* (queried entities) and *memorized answers* (gold objects from the `Answer` field), then score each string with the `wordfreq` Zipf proxy (leading articles stripped; same normalization as the country-frequency analysis). Table 7 reports the mean Zipf score over unique entities in each role.

K Alternative question template and cross-family phrasing sensitivity

This appendix supplements the template ablation described in §3.2. First, to evaluate interaction with context length, we evaluate all Pythia sizes at $L \in \{0, 128\}$ with standard conflict, prose filler in tail position (K=1), and the log-probability classifier (n=2,000). Figure 10 shows that framing tokens can swing $P(\text{in-context})$ by ≥ 80 percentage points at fixed conflict and filler. `learned` saturates near 100 % in-context at L=128 for most sizes, whereas `of_course` suppresses in-context to $\leq$20 % on Pythia $\geq$1.4B even with 128-token prose filler. `possessive` strongly pushes the models to low $P(\text{in-context})$ without a filler but has the opposite effect with the 128-token filler.

Cross-family generalization across all 31 models. To test whether phrasing sensitivity is an artifact of the Pythia suite, we extend the five-template

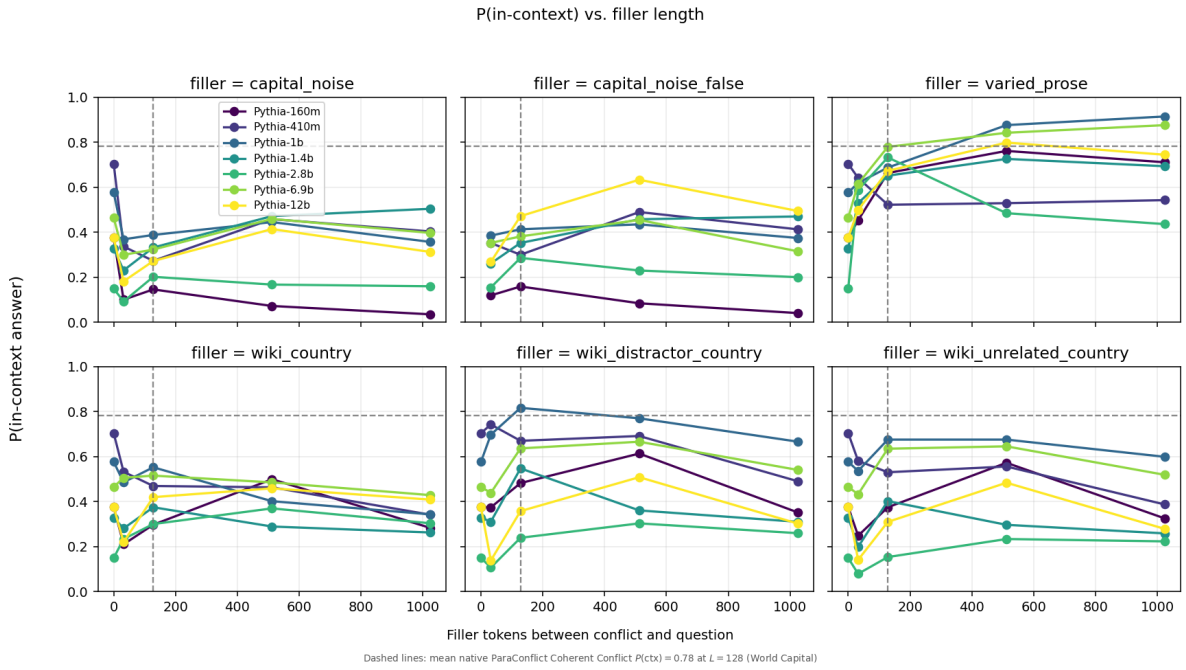

Figure 6: P(in-context) vs. filler length for every filler type and Pythia size.

Table 3: Per-model effect of post-training on P(mem) (percentage-point change, post-trained minus Base) for size-matched models on ParaConflict `Substitution Conflict` prompts. No significant trend change is observed across model size for either Qwen3 or Ministral-3.

	Qwen3 vs. Base					Ministral-3 Instruct vs. Base			Ministral-3 Reasoning vs. Base		
Category	0.6B	1.7B	4B	8B	14B	3B	8B	14B	3B	8B	14B
World Capital	−11.0	−8.8	−9.0	+5.2	−26.2	+56.5	+45.3	+22.3	−21.9	−12.8	−6.8
Athlete Sport	−0.2	−3.2	+1.1	−13.7	−43.0	+5.1	+7.0	+9.4	−3.0	−13.0	+7.0
Book Author	+0.0	+3.4	−7.8	−59.8	+7.2	−3.4	+61.6	+55.6	−11.2	+32.2	−7.4
Company Headquarter	+0.2	+1.0	+0.4	−6.0	+4.8	+8.0	+9.2	+19.2	+0.2	−1.8	−5.6
Company Founder	+0.0	−6.4	−9.4	−10.3	−29.6	+1.0	+27.6	+22.7	−2.5	+14.8	−3.9
Official Language	−62.2	−29.0	−3.1	−55.4	−9.3	+10.9	+15.0	+4.1	+8.8	−17.6	−45.6
Mean (relations)	−12.2	−7.2	−4.6	−23.3	−16.0	+13.0	+27.6	+22.2	−4.9	+0.3	−10.4

evaluation (`qa`, `bare`, `possessive`, `learned`, `of_course`) across all 31 models in our study at L=0 (n=2,000 prompts per cell; 155 evaluations total). Figure 11 reports the resulting maximum template swings $\Delta_{\max} = \max P(\text{mem}) - \min P(\text{mem})$ across the five templates.

While older, smaller architectures (GPT-2) display modest swings (6.6 to 16.4 pp), modern pretrained and post-trained models exhibit massive phrasing sensitivity across every family:

- **Pythia**: swings range from 22.6 pp (160M) to 89.8 pp (12B), growing with scale.
- **Qwen3-Base**: swings reach 83.3 pp (1.7B), 91.5 pp (8B), and 71.5 pp (14B).
- **Qwen3 (post-trained)**: swings reach 92.7 pp (4B), 94.3 pp (8B), 95.0 pp (14B), and 93.0 pp (32B). Models that appear almost completely immune to memorized knowledge under the standard `qa` scaffold (0.0% for 14B, 2.1% for 32B) invert entirely under `of_course` (jumping to 95.0% and 95.2% memorization, respectively) and under `bare` (71.0% for 14B).
- **Ministral-3**: swings reach 81.3 pp in Base (8B), 67.4 pp in Instruct (3B), and 82.4 pp in Reasoning (14B).

These results confirm that prompt phrasing sensitivity is an architecture-general phenomenon and that measuring conflict arbitration from any single question scaffold can yield highly misleading

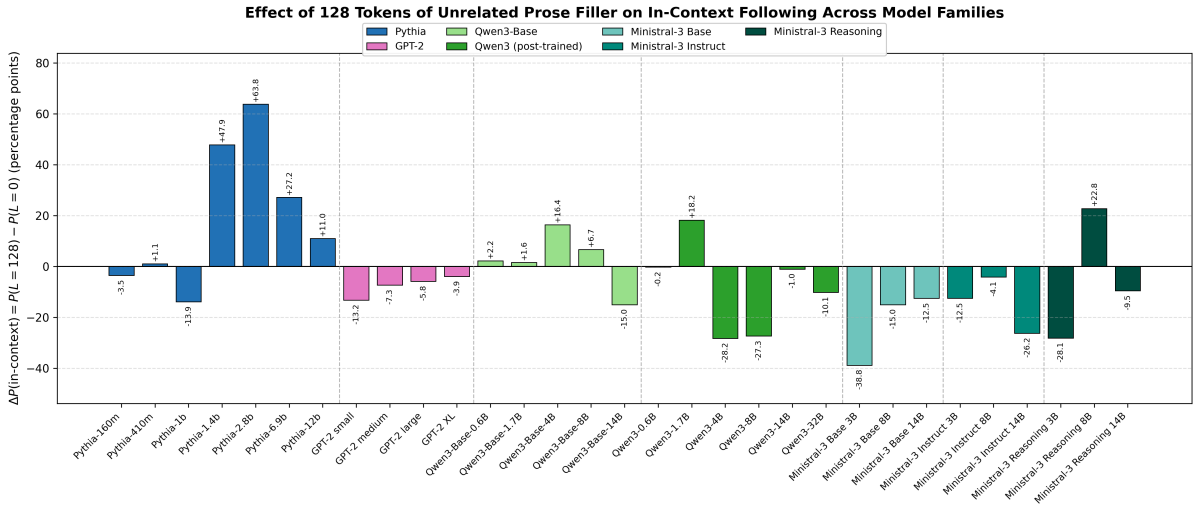

Figure 7: Cross-family effect of inserting L=128 tokens of unrelated biology prose between the conflict statement and the question across all 31 models (n=2,000 World Capital prompts, log-probability classifier). $\Delta P(\text{in-context}) = P(\text{in-context} \mid L{=}128) - P(\text{in-context} \mid L{=}0)$. Colors match Figure 1 for each model family.

Table 4: Spearman ρ between parameter count and $P(\text{mem})$ by model family.

Family (n_{models})	World Capital	Athlete Sport	Book Author	Co. HQ	Co. Founder	Official Lang.	$\bar{P}(\text{mem})$	$\bar{\rho}$
Pythia (7)	+0.82	+0.68	+0.63	−0.56	+0.76	+0.68	0.18	+0.50
GPT-2 (4)	+1.00	−0.20	—	—	—	−0.20	0.06	+0.20
Qwen3-Base (5)	+0.90	+0.90	+0.70	+1.00	+0.90	+0.70	0.31	+0.85
Qwen3 (6)	+0.14	+0.94	+0.77	+0.94	+0.93	+0.60	0.21	+0.72
Ministral-3 Base (3)	+1.00	+0.50	+0.50	+1.00	+0.50	−0.50	0.27	+0.50
Ministral-3 Instruct (3)	+1.00	+0.50	+1.00	+1.00	+0.50	−0.50	0.48	+0.58
Ministral-3 Reasoning (3)	+1.00	+1.00	+0.50	+1.00	+0.50	−1.00	0.22	+0.50

conclusions about model-level memory resistance.

L Negated conflict statement

This appendix supplements the conflict-template ablation in §3.2. We replace `...is {distractor}.` with `...is not {distractor}.` under the paper qa template, tail position (K=1), the log-probability classifier, and n=2,000 prompts at $L \in \{0, 128\}$ with prose or `capital_noise` filler. Table 8 reports $P(\text{in-context})$ under standard vs. negated conflict; cells show standard → negated (Δ). Negative Δ indicates a shift toward the memorized capital. At L=0 the two fillers agree (no intervening tokens); negation has little effect on Pythia-160M but grows with scale, reaching $\Delta = -42$ for Pythia-12B. At L=128, both fillers weaken this effect, `capital_noise` more than `prose`.

M Conflict repetition

This appendix supplements the conflict-copy ablation in §3.2. With no intervening filler (L=0), we repeat the standard conflict statement $K \in \{1, 2, 4, 8\}$ times before the qa question, holding all else fixed (log-probability classifier, n=2,000 prompts, Pythia suite).

Figure 12 shows that a single extra copy (K=2) pushes every size to 85–97 % in-context; K=4 and K=8 plateau near saturation. Repeated same-template assertions therefore reinforce in-context binding even without recency from filler tokens. This is a much stronger effect than `capital_noise` filler.

N Model Checkpoints and Inference Interface

Table 9 lists all 31 evaluated checkpoints across the Pythia, GPT-2, Qwen3, and Ministral-3 families. Evaluations use native `bfloat16` (Qwen3,

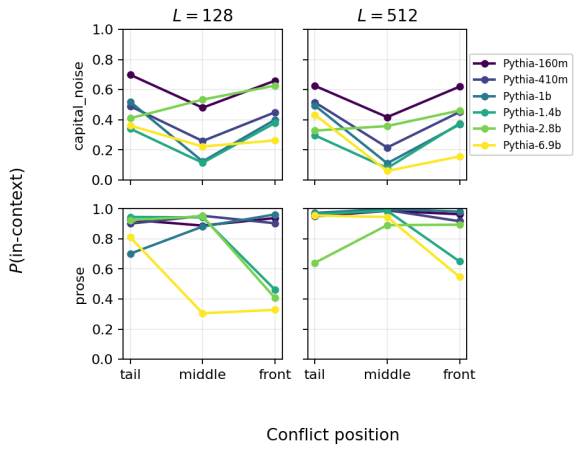

Figure 8: Effect of the position of the conflict in the prompt.

Ministral-3), `float16` (Pythia-6.9B, 12B), and `float32` (smaller Pythia, GPT-2).

Prompt formatting All models are evaluated on identical raw completion prompts (e.g., `"The capital of Poland is London. Q: What is the capital of Poland? A:"`) without chat templates (e.g., ChatML `<|im_start|>` or `[INST]`). As shown in §4.1, prompt phrasing alone swings $P(\text{mem})$ by up to 95 pp; wrapping instruct/reasoning models in chat templates while feeding raw text to base models would conflate post-training effects with prompt framing. Raw evaluation maintains token-level input parity across stages and architectures.

Reasoning models and generation limits. Generative evaluation uses deterministic greedy decoding (T=0, max_new_tokens=10) with substring matching against gold aliases. We acknowledge that this uniform evaluation design choice does not equally test reasoning models under the conditions for which they are designed, such as extended generation budgets or multi-step thinking processes. Rather than introducing model-specific generation limits or specialized reasoning scaffolds, we maintain identical decoding constraints across all 31 base, instruction, and reasoning models to preserve controlled comparability, even though this uniform 10-token regime restricts reasoning models to direct short-horizon completion.

Table 5: Native ParaConflict: $P(\text{in-context})$ under Substitution vs. Coherent Conflict (generative classifier; six categories; n=2146 per model). $\Delta = P(\text{ctx} \mid \text{Coh}) - P(\text{ctx} \mid \text{Sub})$.

Model	$P(\text{ctx} \mid \text{Sub})$	$P(\text{ctx} \mid \text{Coh})$	Δ
Pythia			
Pythia-160m	0.71	0.82	+0.11
Pythia-410m	0.86	0.92	+0.06
Pythia-1b	0.76	0.92	+0.16
Pythia-1.4b	0.85	0.92	+0.08
Pythia-2.8b	0.76	0.93	+0.17
Pythia-6.9b	0.70	0.94	+0.24
Pythia-12b	0.81	0.92	+0.12
GPT-2			
GPT-2 small	0.90	0.92	+0.02
GPT-2 medium	0.94	0.95	+0.01
GPT-2 large	0.88	0.94	+0.05
GPT-2 xl	0.93	0.94	+0.01
Qwen3			
Qwen3-0.6B	0.76	0.97	+0.21
Qwen3-1.7B	0.80	0.93	+0.14
Qwen3-4B	0.64	0.90	+0.27
Qwen3-8B	0.71	0.94	+0.23
Qwen3-14B	0.49	0.96	+0.47
Qwen3-32B	0.48	0.94	+0.47
Qwen3-Base			
Qwen3-Base-0.6B	0.89	0.95	+0.07
Qwen3-Base-1.7B	0.77	0.96	+0.18
Qwen3-Base-4B	0.61	0.95	+0.34
Qwen3-Base-8B	0.39	0.97	+0.57
Qwen3-Base-14B	0.40	0.96	+0.56
Ministral-3			
Ministral-3 Instruct 3B	0.67	0.71	+0.05
Ministral-3 Instruct 8B	0.44	0.76	+0.32
Ministral-3 Instruct 14B	0.44	0.75	+0.31
Ministral-3 Reasoning 3B	0.78	0.97	+0.19
Ministral-3 Reasoning 8B	0.67	0.97	+0.31
Ministral-3 Reasoning 14B	0.81	0.97	+0.16
Ministral-3 Base 3B	0.79	0.97	+0.19
Ministral-3 Base 8B	0.76	0.97	+0.22
Ministral-3 Base 14B	0.73	0.97	+0.24

Table 6: $P(\text{mem})$ by queried-country `wordfreq` Zipf percentile bin (10 equal-count bins; $n{\approx}4{,}730$ prompts per bin per model). ρ: Spearman correlation between per-country frequency and $P(\text{mem})$; Δ: $P(\text{mem})$ in the highest minus lowest bin (percentage points). Verdicts follow §4.2: *supports* if $\rho \geq 0.3$ and $\Delta \geq 5$ pp.

Model	Q10	Q20	Q30	Q40	Q50	Q60	Q70	Q80	Q90	Q100	ρ	Δ pp	Verdict
GPT-2 small	8.8	4.5	4.7	4.6	0.0	12.3	1.7	6.5	3.5	5.8	+0.430	-2.9	flat
GPT-2 medium	4.4	2.6	1.1	4.1	1.1	12.9	8.0	12.1	17.0	9.2	+0.553	+4.7	weak support
GPT-2 large	4.9	2.4	0.6	5.2	6.2	14.1	17.4	25.2	32.5	22.0	+0.613	+17.1	**supports**
GPT-2 XL	5.8	7.8	8.6	18.5	29.3	46.9	42.4	44.0	55.8	54.7	+0.642	+48.9	**supports**
Ministral-3 Base-3b	16.6	28.5	26.7	31.9	26.4	34.7	29.9	42.5	25.8	39.1	+0.242	+22.5	weak support
Ministral-3 Base-8b	24.5	41.5	38.2	45.1	35.5	52.3	48.1	44.9	44.4	51.7	+0.228	+27.2	weak support
Ministral-3 Base-14b	32.4	59.7	52.7	70.1	72.7	75.8	71.8	72.5	81.3	80.8	+0.383	+48.4	**supports**
Ministral-3 Instruct-3b	59.1	88.3	71.4	91.1	97.2	97.1	88.2	84.1	93.3	96.3	+0.366	+37.2	**supports**
Ministral-3 Instruct-8b	60.8	93.6	80.4	93.6	96.9	92.4	89.6	86.4	93.1	93.1	+0.061	+32.4	weak support
Ministral-3 Instruct-14b	62.6	90.6	79.9	98.8	98.0	95.3	90.0	93.1	92.6	93.4	-0.036	+30.9	flat
Ministral-3 Reasoning-3b	3.8	5.3	11.4	7.1	7.3	9.6	8.7	8.5	8.0	13.5	+0.284	+9.8	weak support
Ministral-3 Reasoning-8b	11.2	19.4	21.9	23.1	29.2	32.3	37.9	37.5	41.1	45.4	+0.521	+34.2	**supports**
Ministral-3 Reasoning-14b	36.4	59.5	56.1	64.7	63.6	64.6	64.0	65.4	64.5	63.9	+0.201	+27.5	weak support
Pythia-160m	5.3	4.2	1.6	3.6	0.0	12.1	0.1	5.0	1.0	1.9	+0.274	-3.4	flat
Pythia-410m	3.5	1.7	3.0	5.9	5.2	18.0	8.2	18.8	14.7	18.0	+0.543	+14.5	**supports**
Pythia-1b	6.0	3.3	5.3	7.6	9.1	23.2	21.0	20.3	26.1	30.4	+0.510	+24.4	**supports**
Pythia-1.4b	10.8	8.9	15.0	31.4	36.6	50.2	48.6	58.4	61.2	67.1	+0.571	+56.3	**supports**
Pythia-2.8b	18.2	37.3	50.1	71.3	76.5	71.6	82.5	79.6	75.6	80.2	+0.513	+62.1	**supports**
Pythia-6.9b	10.4	7.1	16.8	25.1	29.0	38.3	37.5	51.0	57.7	65.3	+0.602	+54.9	**supports**
Pythia-12b	10.9	14.2	16.1	34.5	37.7	54.4	58.9	68.5	68.9	74.5	+0.625	+63.7	**supports**
Qwen3-0.6b	1.2	1.4	2.4	2.0	0.1	7.2	1.5	4.1	4.7	5.4	+0.451	+4.2	weak support
Qwen3-1.7b	9.1	4.5	5.0	4.6	0.6	13.1	0.5	5.3	0.8	1.3	+0.288	-7.8	flat
Qwen3-4b	2.1	2.1	2.9	9.1	8.8	9.7	13.8	14.0	23.4	35.7	+0.679	+33.6	**supports**
Qwen3-8b	8.7	12.3	22.1	29.0	31.7	35.3	34.8	30.8	36.6	44.6	+0.436	+35.9	**supports**
Qwen3-14b	2.4	2.1	1.2	1.9	0.8	2.2	0.5	0.9	0.9	0.3	+0.009	-2.1	flat
Qwen3-32b	11.4	11.4	11.7	12.1	12.6	12.6	15.9	13.3	10.8	9.3	+0.073	-2.0	flat
Qwen3-Base-0.6b	3.1	1.5	2.6	5.8	9.8	13.3	14.5	21.5	24.4	43.3	+0.716	+40.2	**supports**
Qwen3-Base-1.7b	2.0	5.0	6.9	6.9	8.1	8.7	15.7	17.0	25.2	36.7	+0.661	+34.7	**supports**
Qwen3-Base-4b	2.0	1.1	3.3	6.9	5.9	10.5	26.0	36.8	52.1	65.9	+0.678	+63.9	**supports**
Qwen3-Base-8b	8.6	6.2	8.4	16.5	24.2	23.1	32.7	32.1	40.6	41.3	+0.581	+32.7	**supports**
Qwen3-Base-14b	25.3	29.8	31.6	30.1	31.8	30.0	20.9	28.0	23.8	24.0	-0.022	-1.4	weak invert

Table 7: Mean `wordfreq` Zipf per unique entity by Para-Conflict category. Each cell averages Zipf over distinct subject or memorized-answer strings; leading `the` is stripped before scoring.

	Subject		Memorized answer	
Category	n	$\bar{z}$	n	$\bar{z}$
World Capital	218	3.58	260	2.67
Athlete Sport	532	2.93	33	4.09
Book Author	500	3.42	279	2.68
Company Headquarter	500	3.02	242	3.37
Company Founder	203	2.93	255	2.25
Official Language	192	3.68	162	2.72

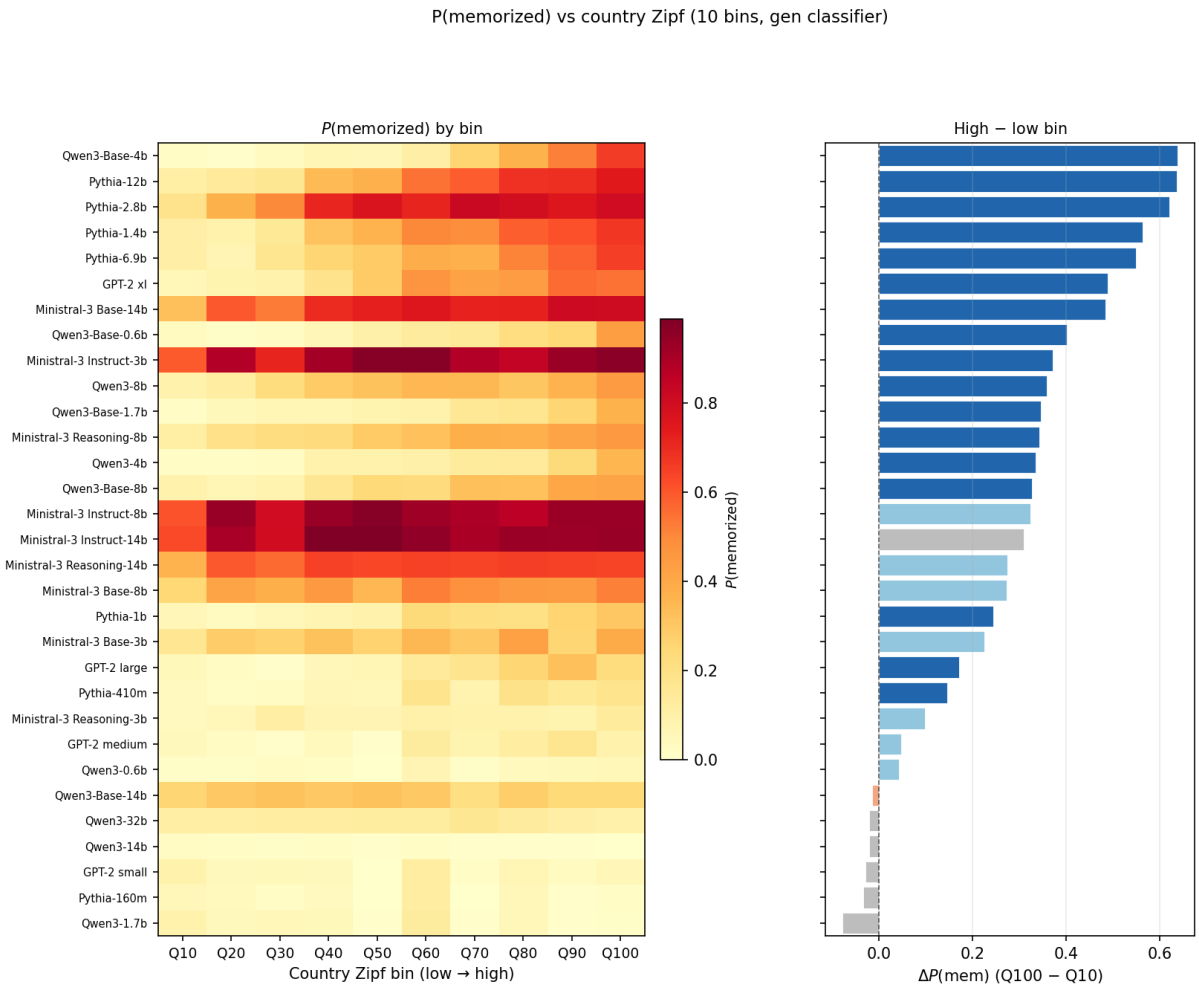

Figure 9: Country-frequency memorization across model families (generative classifier). **Left:** $P(\text{memorized})$ in ten country-Zipf percentile bins (Q10 = rarest countries; Q100 = most frequent). **Right:** $\Delta P(\text{memorized})$ between Q100 and Q10. Rows share a common y-axis and are sorted by Δ (ascending); bar color encodes the verdict in Table 6 (blue = supports the Yu et al. trend, red = inverts or weakly inverts, gray = flat).

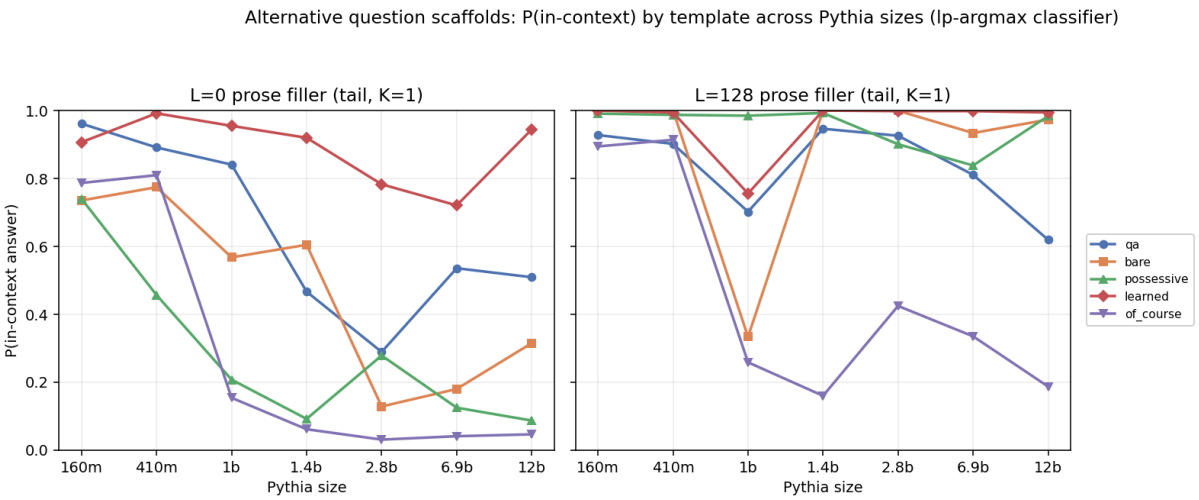

Figure 10: $P(\text{in-context})$ increasingly depends on the question template with Pythia model size.

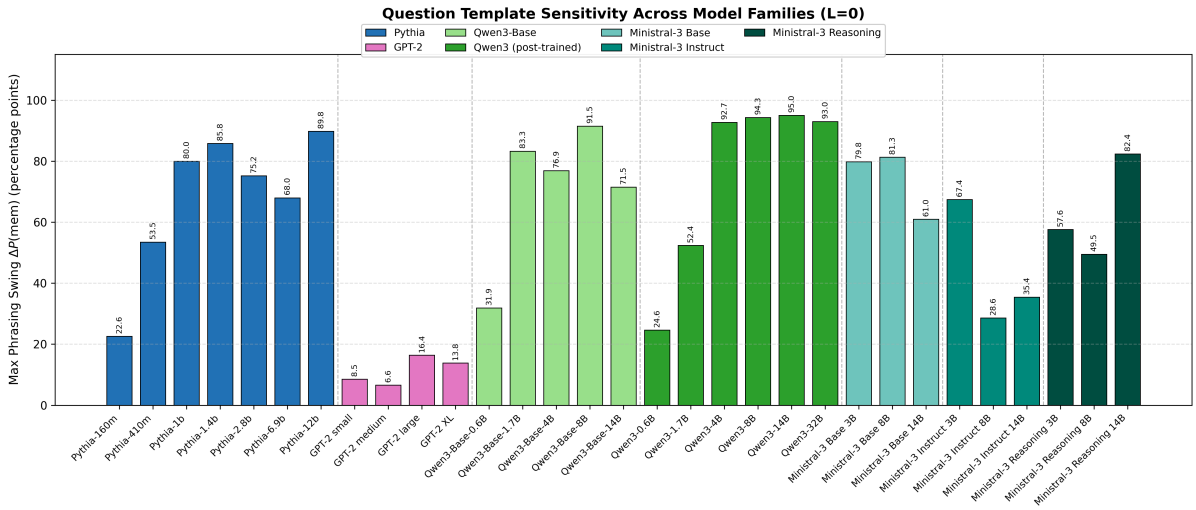

Figure 11: Maximum phrasing swing $\Delta_{\max} = \max P(\text{mem}) - \min P(\text{mem})$ across five question templates for all 31 models at L=0 (n=2,000).

Table 8: Negated vs. standard conflict (log-probability classifier). Each cell gives $P(\text{in-context} \mid \text{standard}) \rightarrow P(\text{in-context} \mid \text{negated})$ (Δ).

Size	L=0	L=128 (prose)	L=128 (`capital_noise`)
160m	0.96 → 0.97 (+0.01)	0.93 → 0.94 (+0.02)	0.70 → 0.76 (+0.06)
410m	0.89 → 0.81 (−0.08)	0.90 → 0.84 (−0.06)	0.49 → 0.46 (−0.03)
1b	0.84 → 0.68 (−0.16)	0.70 → 0.51 (−0.20)	0.52 → 0.40 (−0.12)
1.4b	0.47 → 0.27 (−0.20)	0.95 → 0.83 (−0.11)	0.34 → 0.26 (−0.08)
2.8b	0.29 → 0.09 (−0.20)	0.93 → 0.64 (−0.29)	0.41 → 0.20 (−0.21)
6.9b	0.54 → 0.23 (−0.30)	0.81 → 0.52 (−0.29)	0.36 → 0.24 (−0.12)
12b	**0.51 → 0.09 (−0.42)**	**0.62 → 0.26 (−0.36)**	**0.31 → 0.19 (−0.12)**

Table 9: Evaluated model catalogue across all 31 checkpoints with exact Hugging Face repository identifiers.

Family	Stage	Hugging Face Checkpoint	Params	Vocab	Pretraining	Eval Dtype
Pythia	Pretrained	`EleutherAI/pythia-{160m. . . 12b}`	0.16B–12B	50,304	300B (The Pile)	`float32/16`
GPT-2	Pretrained	`openai-community/gpt2{,-medium,-large,-xl}`	0.12B–1.5B	50,257	WebText	`float32`
Qwen3-Base	Pretrained	`Qwen/Qwen3-{0.6B. . . 14B}-Base`	0.6B–14B	151,936	36T	`bfloat16`
Qwen3	Post-trained	`Qwen/Qwen3-{0.6B. . . 32B}`	0.6B–32B	151,936	36T	`bfloat16`
Ministral-3	Base	`mistralai/Ministral-3-{3B,8B,14B}-Base-2512`	3B, 8B, 14B	131,072	1T–3T	`bfloat16`
Ministral-3	Instruct	`mistralai/Ministral-3-{3B,8B,14B}-Instruct-2512`	3B, 8B, 14B	131,072	1T–3T	`bfloat16`
Ministral-3	Reasoning	`mistralai/Ministral-3-{3B,8B,14B}-Reasoning-2512`	3B, 8B, 14B	131,072	1T–3T	`bfloat16`

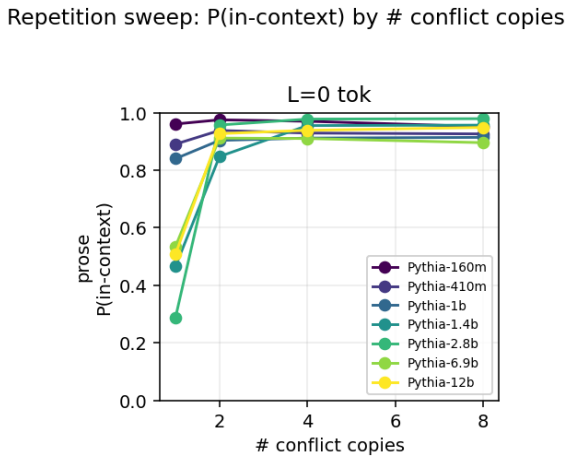

Figure 12: Conflict repetition at $L{=}0$ (log-probability classifier). $P(\text{in-context})$ vs. Pythia size for $K \in \{1, 2, 4, 8\}$ redundant conflict copies.