BENCHMARKING KNOWLEDGE ACQUISITION WITH CAUSALLY-RELATED SYNTHETIC FUTURE EVENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models rely on static pre-training corpora, resulting in outdated knowledge and motivating methods for post-hoc knowledge integration. Existing approaches for evaluating knowledge updates either suffer from rapid contamination or rely on counterfactual edits that conflict with a model's existing knowledge. In this work, we introduce CASCADE, a model pipeline that generates causally interconnected fictional future events while preserving local and global consistency across a large event graph. We use CASCADE to construct a benchmark spanning a taxonomy of question reasoning types, avoiding contamination while enabling controlled, fine-grained evaluation of knowledge acquisition and reasoning. Using CASCADE, we benchmark retrieval on proprietary and open-weight models and continued pretraining (CPT) on open-weight models, finding that retrieval consistently outperforms parametric knowledge injection and that reasoning across multiple linked events is substantially harder than reasoning on a single-event. We show that reasoning-focused reinforcement learning improves accuracy under retrieval, and that retrieval methods degrade less as the number of events increases compared to CPT. These results position CASCADE as a scalable, contamination-resistant testbed for knowledge updating and multi-event reasoning.

1 INTRODUCTION

Large Language Models (LLMs) are increasingly used as parametric knowledge bases for knowledge-intensive questions (Anthropic, 2025; Comanici et al., 2025; Singh et al., 2026). However, real-world knowledge is rarely static, as facts and relationships evolve continuously over time. Evaluating how effectively LLMs use new information outside their existing knowledge remains a critical open challenge in language model research (Liška et al., 2022; Thede et al., 2025).

Existing work on knowledge acquisition falls into three paradigms based on how new knowledge is collected. **Counterfactual datasets** (Zhong et al., 2025; Meng et al., 2022) perturb existing facts (e.g., swapping historical figures). Despite being new, these edited facts could conflict unnaturally with pre-trained knowledge, causing benchmarks to inadvertently measure both genuine knowledge updates and memorized noise (Singh, 2026; Park et al., 2025). **Temporal datasets** (Kasai et al., 2023; Liška et al., 2022) rely on recent events, but they quickly become obsolete as model knowledge cutoffs advance. Since LLMs have different cutoffs, fair comparison requires repeatedly collecting new data and re-benchmarking models, which is costly. **Synthetic future event benchmarks** (Park et al., 2025; Glockner et al., 2025; Zheng et al., 2026) sidestep both issues by constructing consistent fictional future events, capturing causal relationships among facts within individual events. However, these events are typically generated in isolation, so such benchmarks cannot test whether models integrate knowledge across interconnected events, as real-world questions often require.

To this end, we propose CASCADE. Beyond causal relationships between facts in individual isolated events, CASCADE introduces causal relationships among events, creating an *integrated* fictional world. CASCADE links event topics through a shared causal graph and represents the resulting world as a corpus of *long-form, interlinked articles* in the style of Wikipedia, mirroring how knowledge appears in pre-training data. These articles cover the events and their associated entities, which enables evaluation of how models utilize information distributed across documents.

Comprehensively evaluating knowledge acquisition requires distinguishing the kinds of reasoning models can perform over newly introduced facts, beyond fact memorization alone. We therefore pair

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.

CASCADE with a fine-grained question taxonomy and generation framework (§4.1). The taxonomy spans four categories: single-event, multi-event, unanswerable, and open-ended questions, which cover reasoning capabilities including factual recall, multihop reasoning, and causal inference (Lin et al., 2025; Glockner et al., 2025; Zheng et al., 2026). In particular, **multi-event questions** require reasoning over knowledge distributed across interconnected events through causal chains, temporal relationships, and aggregate outcomes, while unanswerable and open-ended questions test whether models can recognize insufficient evidence and provide grounded explanations, respectively.

We benchmark GPT-5.2 (Singh et al., 2026), Gemini-2.5 (Comanici et al., 2025), Qwen3 (Yang et al., 2025a), and LLaMA-3.1 (Grattafiori et al., 2024) on events generated by CASCADE (§6). We find that parameter-based knowledge integration is largely ineffective at our benchmark's scale and factual density: performance degrades sharply as the number of events and their interdependencies grow. Retrieval-based methods perform best overall, though they remain constrained by retrieval budget. Fine-tuning for event-based reasoning improves how models reason over retrieved evidence, yielding accuracy gains on our benchmark. Across all methods, models struggle most with questions requiring reasoning across multiple events or combining fictional and real knowledge.

In summary, our key contributions are as follows:

- We develop a multi-stage framework that generates synthetic new knowledge insertion benchmarks while maintaining strict local and global consistency across the knowledge base.
- We present a comprehensive taxonomy of knowledge update types, enabling precise, fine-grained diagnostic evaluation of event-based reasoning for new knowledge.
- We conduct extensive evaluations across state-of-the-art LLMs, uncovering key insights into how models resolve, propagate, or retain conflicting parametric information.

2 RELATED WORK

New Knowledge Evaluation. Prior work has evaluated LLM temporal consistency with explicit cutoffs (Liu et al., 2026b) and temporal drift using neologisms (Zheng et al., 2024), new named entities (Rijhwani & Preotiuc-Pietro, 2020; Liu et al., 2024b; Thede et al., 2025), evolving real-world events (Vu et al., 2024; Kasai et al., 2023; Liu et al., 2026a; Kim et al., 2026; Amalvy & Huang, 2026), and counterfactual information (Meng et al., 2022; Li et al., 2023; Zhong et al., 2023; 2025). Previous work has also explored episodic experiences (Huet et al., 2025) and hypothetical-but-plausible future events (Park et al., 2025; Glockner et al., 2025; Zheng et al., 2026). In particular, PARALLELEVENTS (Zheng et al., 2026) represents fictional events as knowledge graphs with causal and multihop questions. However, such benchmarks treat events independently and do not present new knowledge as naturally occurring long-form text. Our benchmark pairs interconnected future events with detailed articles, enabling evaluation of integrated knowledge acquisition and reasoning.

Synthetic Text Generation. Synthetic data has been widely explored for pre-training and domain adaptation (Eldan & Li, 2023; Yang et al., 2024) and post-training objectives such as instruction following and reasoning (Lambert et al., 2025; Chan et al., 2025; Muennighoff et al., 2025). Prior analyses have examined its scaling, quality, and novelty (Kang et al., 2025; Liu et al., 2024a; Yang et al., 2024). It has also been applied to temporal and event-based data, including time-series modeling (Rousseau et al., 2025), novel event detection (Du et al., 2024), temporal reasoning (Kenneweg et al., 2025), and temporal graph extraction (Amalvy & Huang, 2026). Our work synthesizes future events as long-form text, yielding a contamination-resistant testbed for knowledge updating.

3 CASCADE: GENERATION OF CAUSALLY CONNECTED EVENTS

To evaluate model capabilities in acquiring and reasoning over novel knowledge without the confound of knowledge conflicts, we introduce CASCADE (**Ca**usal **S**ynthetic **C**reation of **A**nchored, **D**ynamic **E**vents), a method to generate interconnected fictional future events. As illustrated in Figure 1, CASCADE comprises a pipeline for creating a causal graph $G = (V, E)$, where nodes $V = \{v_1, \ldots, v_N\}$ correspond to fictional events and directed edges E encode causal relations, starting from a curated set of N event types obtained from Wikidata and current events (e.g., cultural movements). CASCADE realizes these events by expanding the causal graph into a corpus of Wikipedia articles $a_i \in \mathcal{A}$ and entities $u_j \in \mathcal{U}$ in the events, with hyperlinks connecting the articles.

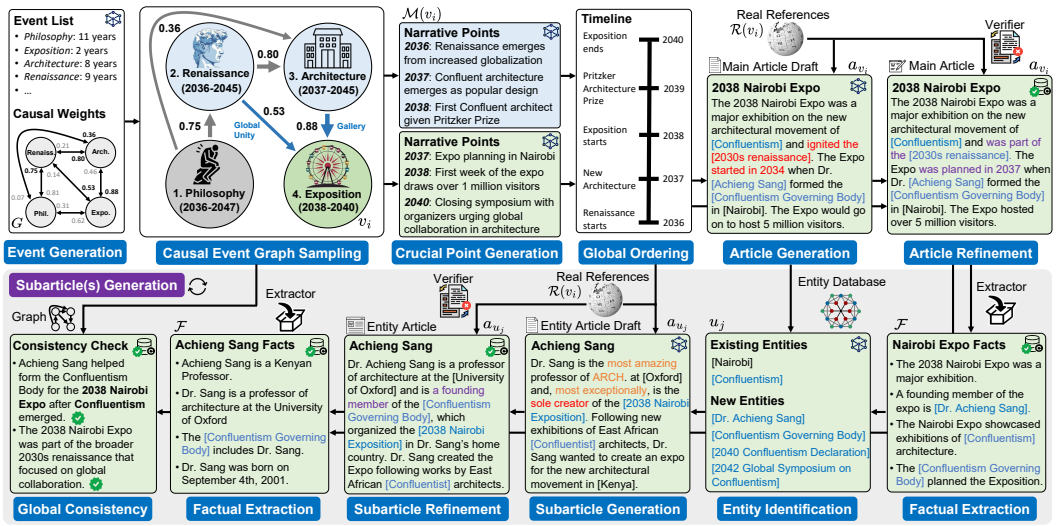


Figure 1: Overview of CASCADE. Starting from curated event types, we sample a causal event graph with plausibility weights, event durations, and causal rationales. Each event is expanded into timestamped narrative milestones, which are integrated into generated event articles and hyperlinked entity articles. Articles are then checked for consistency and rewritten to correct factual errors and non-encyclopedic phrasing, resulting in a hyperlinked corpus of long-form Wikipedia-style articles.

Temporal Ordering and Causality. We randomly partition events into non-overlapping, consecutive time periods (e.g., decades) via δ, where $\delta(v_i) = [\delta_1(v_i), \delta_2(v_i)]$ is the period assigned to event type v_i, to prune the $N!$ possible event orderings. An LLM scorer π_{score} assigns each candidate edge a plausibility weight $w_{ij} = \pi_{\text{score}}(v_i \to v_j) \cdot \mathbb{1}[\delta_1(v_i) \le \delta_1(v_j)]$, zeroing out edges from later to earlier periods. Starting from a fully connected graph, we sample edges e_{ij} with $w_{ij} > \epsilon$, where ϵ is a minimum threshold enforcing sparsity, such that the resulting graph is acyclic. Each sampled edge is annotated with an LLM-generated causal rationale to keep narratives consistent. We topologically sort events and assign each v_i a start year $\mathit{start}(v_i) \in \left[\max\left(\delta_1(v_i),\ \max_{v_j \in p(v_i)} \mathit{start}(v_j)\right),\ \delta_2(v_i)\right]$ and an end year $\mathit{end}(v_i) = \mathit{start}(v_i) + \Delta(v_i)$, where $p(v_i)$ is the set of parent events and $\Delta(v_i)$ is an LLM-specified event duration. This construction ensures that each event starts within its assigned period and no earlier than its parents and may end after its time period, allowing events to overlap.

Event Realization. Each event $v_i \in V$ is realized as an article $a_{v_i} \in \mathcal{A}$ via LLM generation,

$$a_{v_i} \sim \pi_{\text{gen}}\big(\mathcal{M}(v_i),\ r_{v_i}\big),$$

conditioned on two sources: event metadata $\mathcal{M}(v_i)$, comprising event attributes (type, duration, sentiment, themes, etc.), timestamped narrative milestones (detailed, LLM-generated sub-event points), and the causal rationales for the one-hop neighborhood of parent and child events; and a real Wikipedia article $r_{v_i} \sim \mathcal{R}(v_i)$, where $\mathcal{R}(v_i)$ is the set of articles on events of the same type as v_i that serve as demonstrations of narrative arc and style. Milestone timestamps enforce detailed event generation and temporal consistency, while parent/child rationales support causal grounding and forward references (e.g., the 2008 Olympics mentioning the 2012 Olympics), as in real Wikipedia.

Entity Enrichment. Following Wikipedia's convention of hyperlinking entities (Shi et al., 2021; Yang et al., 2025b), we use an LLM to estimate the saliency of $u \in \mathcal{U}(v_i)$, entities (e.g., organizations) mentioned in a_{v_i}, and select the top $K{=}20$ entities, which are split into new entities $\mathcal{U}_{\text{new}}(v_i)$ and existing entities $\mathcal{U}_{\text{real}}(v_i)$. To keep articles mutually consistent, we maintain a fact base $\mathcal{F}$ of atomic factual claims, initialized as $\mathcal{F} = \emptyset$ and updated after each article a_x is generated:

$$\mathcal{F} \leftarrow \mathcal{F} \cup \pi_{\text{extract}}(a_x).$$

For each new entity, we generate its article conditioned on the facts $\mathcal{F}[u_j] \subseteq \mathcal{F}$ relevant to u_j and a sampled reference article $r_{u_j} \sim \mathcal{R}(u_j)$:

$$a_{u_j} \sim \pi_{\text{gen}}\big(\mathcal{F}[u_j],\ r_{u_j}\big), \quad \forall u_j \in \mathcal{U}_{\text{new}}(v_i).$$

For existing entities, we instead update their articles $a^{\text{real}}_{u_j}$, preserving their content and structure:

$$a_{u_j} \sim \pi_{\text{gen}}\big(\mathcal{F}[u_j],\ a^{\text{real}}_{u_j}\big), \quad \forall u_j \in \mathcal{U}_{\text{real}}(v_i).$$



Figure 2: Taxonomy of CASCADE question reasoning types. Questions span four categories: **Single-Event** questions probe one event in isolation; **Multi-Event** questions probe multiple events that are causally or temporally linked; **Unanswerable** questions test refusal on unsupported or invalid information; and **Open-Ended** questions elicit long-form responses.

Article Revision. All articles undergo iterative revision to enforce stylistic and factual consistency. We compute a set of stylistic features (e.g., noun ratio) for each generated article a_x with $x \in \{v_i\} \cup \mathcal{U}(v_i)$, and convert them to z-scores relative to the set of Wikipedia reference articles $\mathcal{R}(x)$. Each z-score measures deviation from typical Wikipedia style, and we revise a_x whenever any feature is atypical ($|z| > 1.96$) (Burrows, 2002). Revision repeats until no outliers remain or an iteration limit is reached. Articles are then verified and revised against the fact base $\mathcal{F}$ built from event and entity articles. A final global check ensures all articles remain coherent with their parents.

4 BENCHMARK QUESTION GENERATION

We introduce a framework that transforms the synthetic events in CASCADE into benchmark questions. The framework comprises modules for generating different types of questions, each targeting complex reasoning over new event information. Each question-answer pair is generated as $(q, y) \sim \pi_{\text{question}}(F, A)$, where $F \subseteq \mathcal{F}$ is a sampled set of atomic claims indexed by category, entity, and timestamp, and $A \subseteq \mathcal{A}$ a sampled set of article sections, both selected for the target question type. Depending on the question type, the generator is grounded in facts, article passages, or both: facts supply the precision needed for numerical and temporal reasoning (e.g., ordering, set addition), while passages supply the event and entity context needed for contextual questions.

4.1 EVENT REASONING QUESTION TAXONOMY

Existing event-centric benchmarks tend to isolate narrow reasoning behaviors (e.g., factual recall), rather than jointly evaluating reasoning over new events and its integration with existing knowledge. We address this gap by synthesizing question types from prior work (Lin et al., 2025; Glockner et al., 2025; Zheng et al., 2026) into a taxonomy of four event-based question categories (Figure 2), covering both answerable and unanswerable questions. **Single-Event** questions probe one CASCADE event in isolation, spanning factual recall, reasoning over its entities and their links to the existing world, and temporal tracking. **Multi-Event** questions mirror these categories but link a focal event to causally or temporally adjacent events, requiring reasoning over event chains, shared entities, and dependencies. **Unanswerable** questions probe safe refusal on invalid premises, unsupported inferences, and unavailable information. Finally, **Open-Ended** questions elicit long-form generations about the events and are graded with a rubric on core content and bias.

4.2 BENCHMARK VERIFICATION

All question-answer pairs pass through multi-stage verification before entering the benchmark. A groundedness filter uses an independent judge $c^* \sim \pi_{\text{judge}}(q, y, \mathcal{A})$ to accept candidate questions-

answers pairs (q, y), cross-referencing each pair against articles $a \in \mathcal{A}$ to eliminate out-of-context or hallucinated content. Advanced categories undergo additional structural checks. For instance, a judge flags any Multi-Source Synthesis question answerable from just a single relevant article as insufficiently distributed. Quantitative questions undergo arithmetic verification, $c \sim \pi_{\text{reasoning}}(q, y, \mathcal{F})$, checking y against the relevant entries in $\mathcal{F}$. Questions failing any stage are discarded. We further audit the benchmark by having an independent judge classify incorrect LLM predictions as model, grading, or question construction errors, and remove the latter.

5 Experimental Setup

Data. In addition to the 50 events generated by CASCADE, we include a held-out *real* split of 50 historical events to measure knowledge retention on existing knowledge, and a disjoint *real_train* split of 50 real-world events providing reasoning-focused training data. The real, historical test set complements the fictional future event set, letting us analyze model behavior changes after training, such as catastrophic forgetting or calibration shifts. Each split is generated with the same question generation pipeline for single-event, multi-event, unanswerable, and open-ended questions, yielding 23,450 fictional questions (7,507 / 4,722 / 6,833 / 4,388) and 10,416 held-out real test questions (1,983 / 2,859 / 3,130 / 2,444). Correctness of single- and multi-event questions is judged with an LLM against the gold answer, while the abstention rate is measured for unanswerable questions. For open-ended questions, we evaluate answers with a rubric-based metric (details in Appendix G.1).

Models. We evaluate the open-weight models of **LLaMA-3.1-8B-Instruct** (Grattafiori et al., 2024), **Qwen3-8B**, and **Qwen3-32B** (Yang et al., 2025a) on all knowledge integration methods, and evaluate retrieval methods on the proprietary models of **GPT-5.2** and **Gemini-2.5-Pro**.

Knowledge Integration Approaches. We compare approaches that supply new knowledge either in context or in model weights. **Retrieval**-based approaches provide new knowledge in context at inference time. All methods retrieve from our corpus of 926 fictional Wikipedia-style articles using BM25. *RAG (articles)* retrieves the top-k=10 full articles, and *RAG (paragraphs)* retrieves the top-k=128 paragraphs, which matches the total context length of 10 articles. We also benchmark IRCoT (Trivedi et al., 2023), which interleaves multiple rounds of retrieval with chain-of-thought reasoning.

We also supply new knowledge parametrically with **continual pre-training (CPT)** methods, which train open-weight models on the 926 articles for 10 epochs with next-token prediction, aiming to inject the new knowledge directly into model weights. Since CPT can degrade instruction following, we also apply *mid-training* using *real_train* after CPT to recover answering and abstention behavior.

For open-weight models, we also experiment with **event-reasoning mid-training**, which teaches models when and how to reason and abstain over *retrieved evidence*. All reasoning-training methods use the 11,430 supervised QA instances in *real_train* (2,963 / 2,778 / 2,419 / 3,270 across the four question families, with 326 held out for development). **DPO** (Rafailov et al., 2024) pairs each gold response with a dispreferred one (an incorrect answer or confident hallucination), teaching the model to prefer grounded answers and abstention over fabrication. **OPD** (Agarwal et al., 2024) distills each 8B student from a frozen same-family teacher (Qwen3-32B and LLaMA-3.1-70B-Instruct, respectively) on the student's own rollouts, providing dense token-level supervision. We add a forward-KL term on the gold answer y: $\mathcal{L} = \mathcal{L}_{\text{OPD}} + \lambda \sum_t \mathrm{KL}\big(T(\cdot \mid x, y_{<t}) \,\|\, S(\cdot \mid x, y_{<t})\big)$, with $\lambda = 0.1$. **GRPO** (Shao et al., 2024) applies on-policy RL from the base model, directly optimizing answer correctness as judged by an LLM (GPT-4.1) against the gold answer.

Inference. During inference on the fictional events, retrieval and reasoning-training methods receive retrieved passages, while CPT-based methods see only the question (closed-book), since they are meant to encode new knowledge in their weights. On real events, all methods are evaluated closed-book, providing a reference for pre-trained knowledge and revealing how each method affects knowledge forgetting and abstention behaviors.

6 Results

Retrieval Methods. Table 1 provides the main results on our benchmark. We first establish retrieval-based baselines. Among proprietary models, GPT-5.2 reaches 73.4% single-event accuracy using retrieval with articles on fictional events, ahead of Gemini-2.5-Pro (64.0%); among

open-weight models, Qwen3-8B outperforms LLaMA-3.1-8B by 9.4 points on average. Single-event accuracy is largely insensitive to retrieval unit of paragraphs or articles. On open-ended questions, GPT-5.2 and Qwen3-8B achieve comparable rubric scores, while Gemini-2.5-Pro struggles, achieving only 31.5 points. We also experiment with an oracle setting where relevant event articles are given to models directly, which improves model performance by less than 2 points, indicating that our retrieval setup introduces limited noise to the results.

Multi-event questions are the hardest question category overall, and unanswerable questions are especially difficult for open-weight models. Multi-event questions, which our causal-graph construction is uniquely able to generate at scale, highlight a large and consistent reasoning capability gap: every model scores substantially lower on multi-event questions than on single-event questions (e.g., GPT-5.2 achieves 71.3 points on single-event but only 53.5 on multi-event questions), as multi-event reasoning types (e.g., causal chaining, aggregation) require merging knowledge across events, increasing informational load. Specifically, multi-event questions reference 2.25× more distinct entities (7.98 vs. 3.55) pulled from nearly 3× as many source documents (2.93 vs. 1.25) than single-event questions. Another challenge emerges for unanswerable questions, particularly among open-weight models: Qwen3 models abstain correctly only around 60% of the time, far behind GPT-5.2 (76.5%) and Gemini-2.5-Pro (89.9%). This suggests open-weight models are less able to recognize when retrieved evidence is insufficient to support an answer (§6.2).

Method	Single	Unanswer	Open	Multi
GPT-5.2				
RAG (articles)	73.4	**76.5**	**72.4**	46.8
RAG (paragraphs)	71.3	74.5	69.8	53.5
Agentic RAG	**74.8**	36.3	**72.4**	**56.6**
Gemini-2.5 Pro				
RAG (articles)	**64.0**	89.1	**31.5**	29.7
RAG (paragraphs)	59.9	**89.9**	26.2	**39.3**
Qwen 3 32B				
RAG (articles)	68.8	59.2	70.4	46.8
DPO	**70.8**	**66.5**	**82.7**	**49.6**
Qwen 3 8B				
RAG (articles)	62.5	61.4	63.5	41.9
RAG (paragraphs)	59.1	59.3	48.6	**47.8**
IRCoT	58.1	53.0	22.3	40.7
Agentic RAG	41.7	64.5	48.6	29.3
DPO	**65.3**	**77.0**	**75.0**	39.4
OPD	64.5	53.8	60.0	43.8
GRPO	63.8	35.0	60.2	42.9
CPT	14.5	9.6	3.3	16.8
CPT→DPO	17.3	10.7	1.7	15.1
CPT→OPD	13.6	22.6	1.2	7.0
LLaMA-3.1-8B				
RAG (articles)	53.0	59.8	15.1	30.2
RAG (paragraphs)	48.0	62.1	15.5	33.9
IRCoT	50.3	48.0	6.1	**34.2**
DPO	**54.2**	55.0	**17.8**	31.2
OPD	53.9	36.2	11.0	32.9
GRPO	48.4	**67.6**	12.1	29.9
CPT	4.5	9.8	0.1	3.8
CPT→DPO	1.4	12.0	0.1	2.2
CPT→OPD	2.8	51.4	0.1	1.5

Table 1: Results on fictional future events: **Single** event, **Unanswer**able, **Open**-ended, **Multi**-event. Method types are colored: retrieval, RL training, CPT. **Best/**Second-best scores per model are highlighted.

Continued Pre-training. Full-parameter CPT performs poorly, with Qwen3-8B and LLaMA-3.1-8B averaging just 9.9% single- and multi-event accuracy on fictional events. While prior work shows effective parametric knowledge injection using synthetic, supervised datasets (Park et al., 2025; Zheng et al., 2026), which are extremely costly to construct for CASCADE's 926 documents, *unsupervised training baselines* are insufficient to encode new knowledge. Chaining CPT with event-reasoning training recovers neither question answering accuracy nor abstention on unanswerable questions, with the best model reaching only 17.3 points on single-event questions. Post-training yields coherent answers, but 61.1% refer to events before 2025, suggesting that CPT fails to encode the new knowledge, leaving models to fall back on pre-trained knowledge. Our follow-up analysis shows CPT accuracy drops as event count increases (§7.1), indicating that denser facts and causal interdependencies make new knowledge harder to retain.

Event Reasoning Training. *Event-reasoning training with DPO improves on base retrieval by teaching models to reason better over retrieved evidence.* For Qwen3-8B, DPO and OPD give the strongest fictional single-event accuracy (65.3 and 64.5 pts) and multi-event accuracy (39.4 and 43.8 pts), while GRPO trails on single-event accuracy. The three methods diverge sharply on unanswerable questions, however: DPO raises fictional abstention to 77.0%, while OPD and GRPO abstain far less (see Appendix O). The same ranking holds for LLaMA-3.1-8B and Qwen3-32B, with Qwen3-32B under DPO exceeding retrieval alone on both fictional single-event (+4.0 pts) and multi-event (+2.8 pts) accuracy. Given the effectiveness of retrieval, teaching models to better reason over retrieved context proves more beneficial than teaching them to memorize knowledge parametrically.

Agentic retrieval. Since many multi-event questions require synthesizing information, we test agentic retrieval (up to 8 rounds, with query refinement) on GPT-5.2 and Qwen3 8B. Qwen sees no bene-

fit: fictional single-event accuracy falls to 41.7 pts and multi-event to 31.8 pts, as the model exhibits only 1.04 average tool calls across all questions, suggesting its poor agentic tool use capability. GPT-5.2 shows the opposite pattern, with the agentic system reaching 74.8 pts for single-event and 57.1 pts for multi-event accuracy, as iterative retrieval lets the model gather and compose evidence across events. However, unanswerable abstention drops sharply to 36.8%, which is linked to multi-retrieval, as restricting retrieval to only 3 rounds raises abstention back to 73.4%, suggesting that additional rounds surface misleading evidence that leads models to answer instead of abstain.

6.1 Results on Existing Knowledge

To assess knowledge insertion relative to existing parametric knowledge, we compare performance on CASCADE fictional future events against parametrically-known real historical events (Table 2). *Fictional future events are substantially harder across nearly every setting*: averaged across models, the out-of-the-box accuracy on real events exceeds fictional events by 10.8, 24.4, and 44.2 points for single-event, open-ended, and multi-event questions. Among event-reasoning methods, OPD best retains existing knowledge, achieving the highest real-event accuracy. Fictional events remain harder under OPD, with real-event accuracy exceeding fictional by 7.0, 28.2, and 26.8 points on single-event, open-ended, and multi-event questions. Compared to fictional events, real events benefit from latent parametric knowledge, remaining easier and well-maintained even after reasoning training. Notably, while multi-event questions are hardest on fictional events, single- and multi-event accuracy are comparable on real events, as the relevant facts are equally well memorized.

Method	Single	Unanswer	Open	Multi
GPT-5.2				
OOB	83.0 (+9.6)	59.1 (–17.4)	83.3 (+10.9)	81.2 (+34.4)
Gemini-2.5 Pro				
OOB	85.6 (+21.6)	73.4 (–15.7)	81.1 (+49.6)	78.1 (+48.4)
Qwen 3 32B				
OOB	**81.1** (+12.3)	36.1 (–23.1)	**84.2** (+13.8)	85.9 (+39.1)
DPO	78.1 (+7.3)	**54.3** (–12.2)	79.4 (–3.3)	**89.7** (+40.1)
Qwen 3 8B				
OOB	72.6 (+10.1)	39.7 (–21.7)	**74.6** (+11.1)	71.9 (+30.0)
DPO	64.2 (–1.1)	64.4 (–12.6)	73.5 (–1.5)	**76.7** (+37.3)
OPD	74.1 (+9.6)	34.5 (–19.3)	70.0 (+10.0)	74.1 (+30.3)
GRPO	61.4 (–2.4)	**88.9** (+53.9)	74.0 (+13.8)	71.5 (+28.6)
LLaMA-3.1-8B Instruct				
OOB	53.5 (+0.5)	82.4 (+22.6)	51.6 (+36.5)	55.0 (+24.8)
DPO	56.7 (+2.5)	14.4 (–40.6)	54.2 (+36.4)	55.5 (+24.3)
OPD	**58.2** (+4.3)	82.8 (+46.6)	**57.4** (+46.4)	**56.2** (+23.3)
GRPO	57.5 (+9.1)	**96.4** (+28.8)	49.4 (+37.3)	49.7 (+19.8)

Table 2: Performance on real historical events. **OOB**: out of the box parametric model. (Δ) is relative to the same method on fictional events (Table 1); OOB is compared to RAG (articles). Method types are colored by row: retrieval, RL training. **Best**/Second best scores per model are highlighted.

On real historical events, *models appear to overestimate their parametric knowledge, leading to degraded abstention behavior.* Even before additional training, GPT-5.2, Gemini-2.5, and Qwen3 8B/32B abstain substantially less often on real events, with an average drop of 20.6 points (see Appendix N for an analysis of why). LLaMA-3.1-8B is the exception, with abstention *rising* from 55.9% to 82.4%; it also abstains more on answerable real-event questions (Table 18), reflecting a conservative default to abstention. Training does not consistently improve abstention, as DPO improves Qwen3 8B but degrades LLaMA-3.1-8B, while OPD improves LLaMA-3.1-8B but degrades Qwen3 8B. Together, fictional and real events probe complementary aspects of knowledge acquisition and usage: fictional events isolate the acquisition of genuinely new knowledge, while real events surface pre-existing and training-related issues.

6.2 Event Question Type Breakdown

Figure 3 reports accuracy on 6 difficult question types, with the full breakdown by all eight reasoning modes given in the Appendix (Tables 6, 7, 8, and 9). Among single-event questions, **aggregative** and **multi-hop** questions exhibit the largest fictional-event penalty (19.8 points on average). Comparative and causal questions show the same general pattern with smaller gaps (Table 6). These question types require models to compose multiple pieces of newly introduced knowledge, sometimes together with existing real-world knowledge, rather than retrieve a single fact directly. Real-events demand the same compositional reasoning: GPT-5.2 produces explicit stepwise reasoning on 64.7% of aggregative questions and 54.3% of multi-hop questions (Table 20). Among training methods, DPO and OPD give the most consistent gains on these composition-intensive question types.

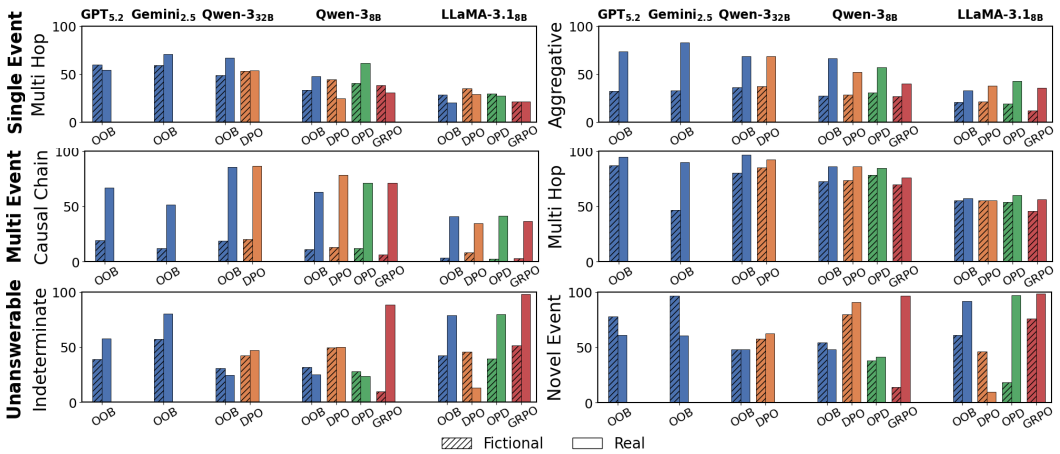


Figure 3: Accuracy on real and fictional events in CASCADE for six challenging reasoning types in Single-Event, Multi-Event, and Unanswerable questions. Fictional events incur the largest penalty on reasoning types that require combining new and existing knowledge.

For multi-event questions, **causal chain** questions are particularly difficult on fictional events across models and methods. For example, Qwen3-8B with RAG achieves only 11.0% accuracy. **Long-chain multi-hop** is the strongest category by comparison and also benefits most from reasoning training, with **OPD** lifting Qwen3-8B from 72.2 to 78.2 pts. On real events, scores are substantially higher across all reasoning types, with **DPO** providing the most consistent gains over the base model, most notably on causal chain (Qwen3-8B: 63.0 pts→78.4 pts). Long-chain multi-hop is especially easy on real events (e.g., 94.4 pts for GPT-5.2) since the consequences linking known historical events are extensively documented in pre-training data.

For unanswerable questions, **indeterminate** and **novel event** are the hardest fictional categories (Qwen3-8B RAG (articles): 31.8% and 54.0% abstention). Indeterminate questions, which ask for a precise value the evidence does not specify, are difficult even for GPT-5.2, which abstains only 38.8% of the time, suggesting that models overestimate what their retrieved evidence supports. **DPO** raises Qwen3-8B abstention on both (to 49.6% and 79.8%), while **OPD** falls below the RAG baseline. On real events, abstention is substantially lower across most categories, and **DPO** again gives the most consistent gains over retrieval for Qwen3-8B (indeterminate: 47.8%→90.6%).

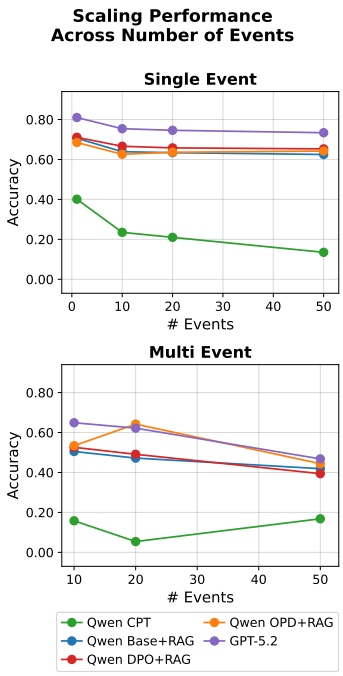


Figure 4: Performance of Qwen3-8B and GPT-5.2 as the number of inserted events scales from 1 to 50, for single-event and multi-event questions.

7 ANALYSIS

7.1 SCALING WITH THE NUMBER OF EVENTS

By controlling the scale of events CASCADE generates, we can systematically increase the size and interdependency density of the fictional world. As the event count increases, the underlying causal graph grows from approximately 599 facts with 1 event to 5,637, 11,130, and 28,711 facts with 10, 20, and 50 events, respectively. Figure 4 leverages this controllability to examine how knowledge acquisition and reasoning performance change as the world scales. For single-event questions, parametric methods degrade sharply as the world scales. With CPT, Qwen's accuracy drops from 40.1 with 1 event to 13.5 with 50 events. In contrast, RAG-based methods remain substantially more robust, with out-of-the-box Qwen and its reasoning-finetuned variants staying within the 62–71 accuracy range and GPT-5.2 dropping from 81 to 73. Multi-event performance is consistently lower than single-event performance and exhibits a stronger downward trend as event count increases (e.g., OPD+RAG: 64.2 → 44.4 pts from 20 to 50 events).

Method	Fict. Single		Fict. Unanswerable		Real Single		Real Unanswerable	
	Brier ↓	ACE ↓	Brier ↓	ACE ↓	Brier ↓	ACE ↓	Brier ↓	ACE ↓
GPT-5.2	**0.194**	**0.130**	**0.103**	**0.104**	**0.105**	**0.073**	**0.166**	**0.125**
Gemini-2.5-Pro	0.319	0.287	0.138	0.125	0.137	0.120	0.238	0.241
Qwen3-32B Base	**0.249**	0.234	0.303	0.286	**0.169**	0.138	0.417	0.423
Qwen3-32B DPO	0.232	**0.199**	**0.260**	**0.215**	0.175	**0.111**	**0.347**	**0.296**
Qwen3-8B Base	0.289	0.316	0.315	0.300	**0.218**	**0.146**	**0.362**	**0.346**
Qwen3-8B DPO	**0.278**	**0.264**	**0.271**	**0.273**	0.224	0.201	0.388	0.385
Qwen3-8B OPD	0.294	0.302	0.419	0.403	0.249	0.186	0.444	0.447
Qwen3-8B CPT → DPO	0.414	0.423	0.356	0.274	0.289	0.258	0.378	0.351
LLaMA-3.1-8B Instruct Base	0.325	0.305	0.391	0.401	0.268	0.235	0.484	0.449

Table 3: Brier score and Adaptive Calibration Error (ACE, 8 equal-mass bins) across single-event and unanswerable questions for fictional and real events. Row color indicates method family: base/RAG, reasoning-training, and CPT. The best value in each model's block is **highlighted**. Overall, GPT-5.2 is the best calibrated model, while DPO yields modest but inconsistent calibration gains.

7.2 Calibration

We evaluate model calibration on single-event questions from Cascade using model-stated confidence, reporting Brier score (MSE against the binary outcome), which reflects both calibration and accuracy, and ACE (Adaptive Calibration Error), which isolates calibration quality (see Appendix H.1 for details). As a reference point, a model that always states 50% confidence obtains a Brier score of 0.25, so higher scores indicate confidence that is less informative than no signal at all. Table 3 reports the results. *GPT-5.2 is the best-calibrated model*, achieving the lowest Brier score (0.194 fictional; 0.105 real) and ACE (0.130 fictional; 0.073 real) on both splits. These scores are comparable to or better than those of models explicitly trained for calibration, which reach Brier scores of 0.14–0.21 on the QA and math tasks they were trained on (Damani et al., 2026). In contrast, Gemini-2.5 is noticeably worse calibrated on both splits and exceeds 0.25 on fictional events. Among fine-tuning recipes, *DPO offers a modest calibration benefit* compared to the base model, reducing calibration error for Qwen3-32B on both splits and for Qwen3-8B on fictional events. In contrast, OPD does not improve calibration, and CPT is the worst calibrated overall, exceeding a score of 0.25 and compounding its poor accuracy.

On unanswerable questions, GPT-5.2 remains the best-calibrated model (Brier 0.103 fictional; 0.166 real), with Gemini-2.5-Pro and the open-weight models trailing by a wide margin, mirroring the single-event pattern. Conditioning confidence on whether a model abstained or hallucinated reveals that *hallucination is typically a confident act*: most models report high mean confidence on their own fabricated answers, and for several models the large majority of hallucinated answers are rated at 70% confidence or above. This shows that confidence, while informative in aggregate, cannot be used to detect individual hallucinations, since fabricated answers are stated as confidently as correct ones across models and methods. *DPO improves calibration inconsistently*: Qwen3-32B's Brier score improves on both splits (-0.043 fictional, -0.070 real), though the gap in confidence between correct abstentions and hallucinations narrows (0.108 → 0.058 fictional). Additionally, Qwen3-8B improves on unanswerable questions for fictional events but worsens on real events, where it is on average 13.9% more confident when hallucinating than when correctly abstaining.

8 Conclusion

We introduce Cascade, a multi-model pipeline that generates contamination-resistant benchmarks of causally interconnected fictional future events while preserving local and global consistency at scale. We benchmark several methods for integrating new knowledge, finding that reasoning across causally linked events is substantially harder than single-event reasoning for every model. Continued pretraining catastrophically forgets both new and existing knowledge due to the factual density of the benchmark, whereas fine-tuning for event-based reasoning with DPO improves calibration and modestly improves accuracy under retrieval by reasoning better over evidence documents. Our real-event comparison further shows models substantially overestimate their own knowledge, fabricating answers with high confidence rather than abstaining, a failure intrinsic to how models handle familiar versus new information. Overall, Cascade is a scalable method for producing rigorous, reusable benchmarks to diagnose how LLMs acquire, retain, and reason over evolving knowledge.

AI USE STATEMENT

In this work, we used generative AI tools as a core part of our methodology, for software development, and for writing assistance. We did not use generative AI tools for literature review or for generating research ideas.

Methodology. CASCADE is an LLM-based pipeline, so generative models produce the benchmark itself. This includes causal edge scoring and rationales, event and entity articles, fact extraction, question–answer pairs, and the consistency, groundedness, and arithmetic verification stages (Sections 3–4, Appendix C). LLMs are also used as evaluation instruments: as judges for answer grading, GRPO rewards (GPT-4.1), and benchmark auditing, and for creating candidate rubric criteria (Appendix G.1). We reviewed this generated content in depth. We manually evaluated corpus quality across seven dimensions (Appendix E.2), validated rubric-based grading against blind human judgments (Appendix G.2), manually categorized model errors (Appendix K), and removed questions flagged as ungrounded, unanswerable, or mis-graded during auditing (Section 4).

Code. We used AI coding assistants to help implement the CASCADE pipeline and our training and evaluation code. The authors reviewed all AI-assisted code extensively by manually inspecting the resulting code and model outputs during training and evaluation.

Writing. We used AI tools to improve the clarity and polish of the manuscript. The authors verified all of the final text in the manuscript. The authors wrote and verified all technical content, results, and claims.

We have reviewed all AI-assisted work, and we take full responsibility for the final content of this paper, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

CASCADE generates fictional events, so there is a potential risk of creating plausible but false text. We reduce this potential risk in two main ways. First, all events are dated in the future. As Appendix J shows, fabricated events dated before a model's knowledge cutoff are endorsed as real far more often (28.0% vs. 2.9%) and are closer in kind to disinformation. Future-dated events cannot be mistaken for verified records and become self-evidently false once their dates pass. Second, new people and organizations central to each event are wholly fictional. Real-world entities (e.g., cities, universities) appear only as settings or as existing articles updated with fictional future developments. We avoid attributing fabricated statements or actions to real, identifiable individuals. The synthetic corpus generation is intended only for research on knowledge updating and not as a source of information about the world.

Several event types cover sensitive themes, such as a pandemic event. Generated text may reproduce framings present in the generator models or reference articles. Our human evaluation (Appendix E.2) found lower neutrality scores for crises events like pandemics, and we flag this as a known limitation. Human annotations for corpus quality and rubric validation (Appendices E.2 and G.1) were performed by in-house annotators. No personal data was collected. All knowledge-injection and training methods are general-purpose and publicly documented.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix C describe the CASCADE generation pipeline, including causal graph sampling, temporal ordering, event and entity realization, the fact database, and the style and consistency revision loops. All stochastic steps use fixed random seeds, so the causal graph, global timeline, and corpus can be regenerated exactly. Section 4 and Section 4.1 specify the question generation framework, the full reasoning-type taxonomy, and the verification stages. Section 5 and Appendix F give the data splits, evaluated models, retrieval configurations, and all training hyperparameters (learning rates, schedules, epochs, LoRA settings, OPD λ sweep, and GRPO reward configuration, including the fix for the empty-completion failure). Appendix G details rubric construction, weighting, and human meta-evaluation for open-ended grading. Results for proprietary models (GPT-5.2, Gemini-2.5-Pro) and the GPT-4.1 judge depend on API versions that may change.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes, 2024. URL https://arxiv.org/abs/2306.13649.

Arthur Amalvy and Hen-Hsen Huang. Beyond known facts: Generating unseen temporal knowledge to address data contamination in llm evaluation, 2026. URL https://arxiv.org/abs/2601.13658.

Anthropic. Claude model family. https://www.anthropic.com/claude, 2025. Computer software / Large Language Model.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3, 1950. URL https://api.semanticscholar.org/CorpusID:122906757.

John Burrows. 'delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267–287, 09 2002. ISSN 0268-1145. doi: 10.1093/llc/17.3.267. URL https://doi.org/10.1093/llc/17.3.267.

Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling synthetic data creation with 1,000,000,000 personas, 2025. URL https://arxiv.org/abs/2406.20094.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, Krishna Haridasan, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. Beyond binary rewards: Training lms to reason about their uncertainty, 2026. URL https://arxiv.org/abs/2507.16806.

Huidong Du, Hao Sun, Pengyuan Liu, and Dong Yu. Generate-then-revise: An effective synthetic training data generation framework for event detection. In *China National Conference on Chinese Computational Linguistics*, pp. 57–72, Taiyuan, China, 2024. Springer Nature Singapore. doi: 10.1007/978-981-97-8367-0_4. URL https://aclanthology.org/2024.ccl-1.78/.

Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english?, 2023. URL https://arxiv.org/abs/2305.07759.

Yiyang Feng, Zeming Chen, Haotian Wu, Jiawei Zhou, and Antoine Bosselut. Tracking the limits of knowledge propagation: How llms fail at multi-step reasoning with conflicting knowledge, 2026. URL https://arxiv.org/abs/2601.15495.

Max Glockner, Xiang Jiang, Leonardo F. R. Ribeiro, Iryna Gurevych, and Markus Dreyer. Neoqa: Evidence-based question answering with generated news events, 2025. URL https://arxiv.org/abs/2505.05949.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Hengrui Gu, Kaixiong Zhou, Xiaotian Han, Ninghao Liu, Ruobing Wang, and Xin Wang. Pokemqa: Programmable knowledge editing for multi-hop question answering, 2024. URL https://arxiv.org/abs/2312.15194.

Pavan Holur, Shreyas Rajesh, David Chong, and Vwani Roychowdhury. Creating an ai observer: Generative semantic workspaces, 2024. URL https://arxiv.org/abs/2406.04555.

Alexis Huet, Zied Ben-Houidi, and Dario Rossi. Episodic memories generation and evaluation benchmark for large language models, 2025. URL https://arxiv.org/abs/2501.13121.

Feiyang Kang, Newsha Ardalani, Michael Kuchnik, Youssef Emad, Mostafa Elhoushi, Shubhabrata Sengupta, Shang-Wen Li, Ramya Raghavendra, Ruoxi Jia, and Carole-Jean Wu. Demystifying synthetic data in llm pre-training: A systematic study of scaling laws, benefits, and pitfalls, 2025. URL https://arxiv.org/abs/2510.01631.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. Realtime qa: What's the answer right now?, 2023. URL https://arxiv.org/abs/2207.13332.

Svenja Kenneweg, Jörg Deigmöller, Philipp Cimiano, and Julian Eggert. Traveler: A benchmark for evaluating temporal reasoning across vague, implicit and explicit references, 2025. URL https://arxiv.org/abs/2505.01325.

Jiyeon Kim, Hyunji Lee, Dylan Zhou, Sue Hyun Park, Seunghyun Yoon, Trung Bui, Franck Dernoncourt, Sungmin Cha, and Minjoon Seo. Can large language models keep up? benchmarking online adaptation to continual knowledge streams, 2026. URL https://arxiv.org/abs/2603.07392.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training, 2025. URL https://arxiv.org/abs/2411.15124.

Zichao Li, Ines Arous, Siva Reddy, and Jackie C. K. Cheung. Evaluating dependencies in fact editing for language models: Specificity and implication awareness, 2023. URL https://arxiv.org/abs/2312.01858.

Teng Lin, Yuyu Luo, Honglin Zhang, Jicheng Zhang, Chunlin Liu, Kaishun Wu, and Nan Tang. Mebench: Benchmarking large language models for cross-document multi-entity question answering, 2025. URL https://arxiv.org/abs/2502.18993.

Hanbing Liu, Lang Cao, and Yang Li. RAG or learning? understanding the limits of LLM adaptation under continuous knowledge drift in the real world. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 11234–11252, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.546. URL https://aclanthology.org/2026.findings-acl.546/.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. Best practices and lessons learned on synthetic data for language models, 2024a. URL https://arxiv.org/abs/2404.07503.

Yachuan Liu, Xiaochun Wei, Lin Shi, Xinnuo Li, Bohan Zhang, Paramveer Dhillon, and Qiaozhu Mei. ExAnte: A benchmark for ex-ante inference in large language models. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1551–1571, Rabat, Morocco, March 2026b. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.72. URL https://aclanthology.org/2026.eacl-long.72/.

Yu Lu Liu, Su Lin Blodgett, Jackie Chi, K. cheung, Q. Liao, Alexandra Olteanu, and Ziang Xiao. Ecbd: Evidence-centered benchmark design for nlp, 2024b. URL https://arxiv.org/abs/2406.08723.

Adam Liška, Tomáš Kočiský, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien de Masson d'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-McMahon, Sophia Austin, Phil Blunsom, and Angeliki Lazaridou. Streamingqa: A benchmark for adaptation to new knowledge over time in question answering models, 2022. URL https://arxiv.org/abs/2205.11388.

Yinyi Luo, Zhexian Zhou, Hao Chen, Kai Qiu, Marios Savvides, Sharon Li, and Jindong Wang. Knowledgesmith: Uncovering knowledge updating in llms with model editing and unlearning, 2026. URL https://arxiv.org/abs/2510.02392.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt, 2022. URL https://arxiv.org/abs/2202.05262.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer, 2023. URL https://arxiv.org/abs/2210.07229.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale, 2021. URL https://arxiv.org/abs/2110.11309.

Tatiana Moteu Ngoli, N'Dah Jean Kouagou, Hamada M. Zahera, and Axel-Cyrille Ngonga Ngomo. *Benchmarking Knowledge Editing Using Logical Rules*, pp. 41–56. Springer Nature Switzerland, October 2025. ISBN 9783032095305. doi: 10.1007/978-3-032-09530-5_3. URL http://dx.doi.org/10.1007/978-3-032-09530-5_3.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Fei-Fei Li, Hanna Hajishirzi, Luke S. Zettlemoyer, Percy Liang, Emmanuel J. Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling, 2025. URL https://arxiv.org/abs/2501.19393.

Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

Shankar Padmanabhan, Mustafa Omer Gul, and Tanya Goyal. Updating parametric knowledge with context distillation retains post-training capabilities, 2026. URL https://arxiv.org/abs/2602.16093.

Core Francisco Park, Zechen Zhang, and Hidenori Tanaka. *New News*: System-2 fine-tuning for robust integration of new knowledge, 2025. URL https://arxiv.org/abs/2505.01812.

Siyuan Qi, Bangcheng Yang, Kailin Jiang, Xiaobo Wang, Jiaqi Li, Yifan Zhong, Yaodong Yang, and Zilong Zheng. In-context editing: Learning knowledge from self-induced distributions, 2025. URL https://arxiv.org/abs/2406.11194.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

Shreyas Rajesh, Pavan Holur, Chenda Duan, David Chong, and Vwani Roychowdhury. Beyond fact retrieval: Episodic memory for rag with generative semantic workspaces, 2026a. URL https://arxiv.org/abs/2511.07587.

Shreyas Rajesh, Pavan Holur, Mehmet Yigit Turali, Chenda Duan, and Vwani Roychowdhury. Panini: Continual learning in token space via structured memory, 2026b. URL https://arxiv.org/abs/2602.15156.

Shruti Rijhwani and Daniel Preotiuc-Pietro. Temporally-informed analysis of named entity recognition. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Annual Meeting of the Association for Computational Linguistics*, pp. 7605–7617, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.680. URL https://aclanthology.org/2020.acl-main.680/.

Cécile Rousseau, Tobia Boschi, Giandomenico Cornacchia, Dhaval Salwala, Alessandra Pascale, and Juan Bernabé-Moreno. Forging time series with language: A large language model approach to synthetic data generation, 2025. URL https://arxiv.org/abs/2505.17103.

Amit Rozner, Barak Battash, Lior Wolf, and Ofir Lindenbaum. Knowledge editing in language models via adapted direct preference optimization, 2024. URL https://arxiv.org/abs/2406.09920.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

William F. Shen, Xinchi Qiu, Chenxi Whitehouse, Lisa Alazraki, Shashwat Goel, Francesco Barbieri, Timon Willi, Akhil Mathur, and Ilias Leontiadis. Rethinking rubric generation for improving llm judge and reward modeling for open-ended tasks, 2026. URL https://arxiv.org/abs/2602.05125.

Weijia Shi, Mandar Joshi, and Luke Zettlemoyer. DESCGEN: A distantly supervised datasetfor generating entity descriptions. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 415–427, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.35. URL https://aclanthology.org/2021.acl-long.35/.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, et al. Openai gpt-5 system card, 2026. URL https://arxiv.org/abs/2601.03267.

Aditya Pratap Singh. On scope classification and current knowledge-editing benchmarks: A negative result, with inlay as a gradient-free case study, 2026. URL https://arxiv.org/abs/2608.26292.

Charlie Snell, Dan Klein, and Ruiqi Zhong. Learning by distilling context, 2022. URL https://arxiv.org/abs/2209.15189.

Lukas Thede, Karsten Roth, Matthias Bethge, Zeynep Akata, and Thomas Hartvigsen. WikiBigEdit: Understanding the limits of lifelong knowledge editing in LLMs. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 59326–59354. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/thede25a.html.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions, 2023. URL https://arxiv.org/abs/2212.10509.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, et al. Freshllms: Refreshing large language models with search engine augmentation, 2024.

Hao Xiong, Chuanyuan Tan, and Wenliang Chen. Is fine-tuning an effective solution? reassessing knowledge editing for unstructured data, 2025. URL https://arxiv.org/abs/2506.09672.

Takayuki Yamamoto and Daisuke Kawahara. Pasta: A paraphrasing and self-training approach for knowledge updating in llms, 2026. URL https://arxiv.org/abs/2606.28898.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025a. URL https://arxiv.org/abs/2505.09388.

Zhongyu Yang, Jun Chen, Dannong Xu, Junjie Fei, Xiaoqian Shen, Liangbing Zhao, Chun-Mei Feng, and Mohamed Elhoseiny. Wikiautogen: Towards multi-modal wikipedia-style article generation, 2025b. URL https://arxiv.org/abs/2503.19065.

Zitong Yang, Neil Band, Shuangping Li, Emmanuel Candès, and Tatsunori Hashimoto. Synthetic continued pretraining, 2024. URL https://arxiv.org/abs/2409.07431.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, et al. A comprehensive study of knowledge editing for large language models, 2024. URL https://arxiv.org/abs/2401.01286.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. Can we edit factual knowledge by in-context learning?, 2023. URL https://arxiv.org/abs/2305.12740.

Jonathan Zheng, Alan Ritter, and Wei Xu. Neo-bench: Evaluating robustness of large language models with neologisms, 2024. URL https://arxiv.org/abs/2402.12261.

Jonathan Zheng, Zirui Shao, Alan Ritter, and Wei Xu. Synthetic worlds for temporal evaluation and knowledge updating in llms, 2026. URL https://arxiv.org/abs/2609.00184.

Shaochen Zhong, Yi-Meng Lu, Lize Shao, Bhargav Bhushanam, Xiaocong Du, Yixin Wan, Yucheng Shi, D. Zha, Yiwei Wang, Ninghao Liu, et al. MQuAKE-remastered: Multi-hop knowledge editing can only be advanced with reliable evaluations. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=m9wG6ai2Xk.

Zexuan Zhong, Zhengxuan Wu, Christopher D. Manning, Christopher Potts, and Danqi Chen. Mquake: Assessing knowledge editing in language models via multi-hop questions, 2023. URL https://arxiv.org/abs/2305.14795.

A Limitations

Our parametric knowledge-injection baselines are primarily unsupervised (CPT on raw articles, optionally followed by event-reasoning training), whereas prior work shows parametric insertion is substantially more effective with supervised or synthetically augmented data such as generated QA pairs and paraphrases (Park et al., 2025; Yang et al., 2024; Zheng et al., 2026); because constructing such data at CASCADE's factual density is costly, we evaluate these approaches only in limited, largely LoRA-based form (QA-SFT, System-2 fine-tuning, DiSC) and do not evaluate knowledge-editing methods (Meng et al., 2023; Mitchell et al., 2021), so our finding that retrieval outperforms parametric insertion should be read as a comparison against unsupervised insertion rather than the strongest possible parametric approach. Beyond this, the corpus and questions are LLM-generated and may inherit generator biases: generated articles remain more lexically complex than real Wikipedia (grade level 19.6 vs. 14.4), human evaluation covers only five events, and crisis-related events score lower on neutrality. Grading relies on LLM judges with binary scoring, which likely understates differences on set-valued questions where most errors are partial, and open-ended rubric agreement with human judgments is weaker for timing and multi-answer questions. The benchmark is English-only and Wikipedia-style, built from a single 50-event causal graph, and the real-event comparison is a reference rather than a controlled counterpart, since real events are evaluated closed-book with a different question distribution and may be partly answerable through memorized shortcuts. Finally, future-dated events will lose their novelty as knowledge cutoffs advance, though CASCADE can be regenerated with later dates.

B Additional Related Work

Knowledge Updating Methods. A broad literature explores incorporating new or modified knowledge into LLMs through model editing (Mitchell et al., 2021; Meng et al., 2022; 2023), in-context and retrieval-based methods (Zheng et al., 2023; Qi et al., 2025; Gu et al., 2024; Zhong et al., 2025), and fine-tuning or RL (Xiong et al., 2025; Zhang et al., 2024; Rozner et al., 2024). Recent work has explored synthetic-data-based updating from news and general knowledge (Luo et al., 2026; Park et al., 2025; Yamamoto & Kawahara, 2026; Zheng et al., 2026), context distillation (Snell et al., 2022; Padmanabhan et al., 2026). Propagation beyond directly updated facts has drawn attention through multi-hop and logical reasoning benchmarks (Feng et al., 2026; Moteu Ngoli et al., 2025). Episodic-memory approaches separately represent temporally structured experiences for reasoning and event anticipation (Huet et al., 2025; Rajesh et al., 2026a; Holur et al., 2024; Rajesh et al., 2026b). Our work complements these directions by evaluating knowledge acquisition on reasoning over newly introduced future events and existing knowledge.

C Additional Details of the Cascade Generation Pipeline

Reproducibility. All stochastic steps — decade partitioning, causal edge sampling, event duration sampling, and entity saliency selection — use fixed random seeds, making the sampled causal graph, global timeline, and resulting corpus fully reproducible.

Type-specific convention distillation. For each event type, we distill a set of conventions from the reference Wikipedia articles of that type, which are then provided to the generator alongside the reference articles themselves. These conventions capture regularities that individual demonstrations convey only implicitly: naming patterns (e.g., whether titles follow a year–location–type template such as "2011 Tōhoku earthquake" or a named-instance form), the composition of linked entities (e.g., the relative proportion of individuals, organizations, locations, and prior events typically referenced), and the typical scale and severity of events of that type (e.g., ranges for participants, geographic extent, or social impact). Distilling these conventions per type keeps generated articles plausible relative to their category, so that, for instance, a fictional regional election does not report nationwide turnout figures, and a fictional earthquake article links to entities in proportions resembling those of real earthquake articles.

Canonical fact database. After each event or entity article is generated, a structured extraction stage updates a fact database for the fictional world, recording each fact's semantic category, participating entities, temporal information when available, and a normalized textual representation. Hyperlinked entities are classified as newly introduced fictional entities, previously established fictional entities inherited from parent events, real-world entities requiring an updated article, or unchanged real-world entities. This database constrains all downstream generation: only facts relevant to a new article are retrieved and injected into its prompt, ensuring previously established information is never contradicted. The database updates incrementally after each article is finalized, keeping later articles consistent with earlier ones.

Entity timeline inference. Before generating dedicated entity articles, an LLM infers each candidate entity's type, creation year, optional termination year, and narrative role from the main event article alone. These inferred attributes become shared constraints for subsequent entity generation, later enforcing that an event cannot reference an entity before its inferred creation year.

Markovian conditioning and context scoping. We restrict causal context to an event or entity's direct one-hop neighborhood (immediate parents and children) rather than the full ancestor/descendant chain, bounding prompt growth as the causal graph deepens. We also cap the number of parent entities or events explicitly referenced in an article, describing remaining dependencies more generically rather than by name.

Canonical name resolution. Entities are always resolved using their exact canonical names from the database, preventing inconsistent surface forms for the same fictional concept across articles.

Iterative revision loops. Event and entity articles pass through up to two verification loops before finalization: a style-fidelity check and a factual-consistency check against the fact database. The style-fidelity check computes 41 features per article and scores each via leave-one-out typicality against approximately 1000 reference Wikipedia articles, averaged into an overall typicality score; articles falling below a typicality threshold of 0.05 are flagged and trigger a rewrite. Eight of these features (hyperlink and citation-density measures) are computed but excluded from scoring, since the plaintext reference set contains no wiki markup for comparison. The remaining features span:

- **Structural:** word count, section count, average section length, maximum heading depth, number of list items, and presence of a substantial lead section.
- **Lexical:** type-token ratio, average word length, proportion of long words (>6 characters), and stopword ratio.
- **Sentence-level:** sentence count, average sentence length, and sentence-length standard deviation.
- **Part-of-speech:** noun, verb, adjective, and superlative ratios; a passive-voice indicator; first- and second-person pronoun ratios; and named-entity (proper noun) density.

- **Readability:** grade level, averaged across Flesch–Kincaid, Gunning Fog, Coleman–Liau, and the Automated Readability Index.
- **Discourse and register:** temporal-expression density, discourse-connective density, parenthetical density, paragraph length and count, sentence-opener variety, and densities of peacock (promotional) and weasel (vague-attribution) language.

Entity articles undergo an additional check against their originating event article and any entity articles referencing them. Failed checks trigger a rewrite and re-check, terminating on a pass or after eight iterations. A final postprocessing pass standardizes Wikipedia-style headers and formatting across all articles.

D Question Taxonomy

The question reasoning types fall under the main categories of single event, multi event, unanswerable, and open-ended questions.

Single Event questions cover 8 main categories:

- **Single-Hop**: Bounded retrieval or deduction within a single source or fact.
- **Multi-Source**: Factual synthesis across sub-articles, such as shared attributes or a timeline.
- **Multi-Hop**: Real-world entity chains with each hop anchored to a verifiable attribute.
- **Aggregative**: Set reduction over entities, via enumeration or computation of a shared attribute.
- **Causal**: Cause-effect relationships, including multi-step chains and unstated consequences.
- **Temporal**: Chronological tracking, ordering, or interval computation over the event timeline.
- **Comparative**: Entities/Event comparison on a stated metric with real or other fictional events.
- **Perspectival/Normative**: Report of agent's documented internal state (stance/intent).

Multi Event questions cover 6 main categories:

- **Shared entity**: A recurring entity tracked across event transitions to test thematic coherence.
- **Inferential**: An unstated outcome emerging only by combining facts from independent events.
- **Causal chain**: A multi-step causal path traced across events, including intermediate steps.
- **Aggregative**: Set aggregation (unions, intersections, counts) over facts from multiple events.
- **Temporal**: Sequencing and interval comparison across distinct event timelines.
- **Long-chain multi-hop**: Reasoning across three or more sequential events, requiring the full causal trajectory tracked explicitly.

Unanswerable questions cover 7 main categories:

- **Missing information**: Requested detail absent from the text and facts.
- **Nonexistent entity**: An actor, object, or location absent from the synthetic world.
- **Novel event**: A plausible, fictionalized sub-incident absent from the source.
- **False presupposition**: A factual error embedded in the premise, precluding validation.
- **Temporally impossible**: A sequence or timing violating the established timeline.
- **Indeterminate**: A precise value where the source gives only approximations or ranges.
- **Spurious correlation**: A coincidental link between unrelated occurrences framed as causal.

Open-ended questions cover 5 main categories:

- **Significance**: Assessment of the event's broader importance.
- **Timing**: Explanation of why an incident occurred at that point rather than earlier or later.
- **Evaluation**: Value judgments synthesized from at least three conflicting stakeholder positions.
- **Multi-Answer**: Questions with multiple valid answers, such as several causes of an event.
- **Counter-policy**: Reasoning on outcomes under alternative decision in the event.

E META-EVALUATION OF CASCADE

E.1 AUTOMATIC EVALUATION

To assess whether the generated corpus is stylistically indistinguishable from authentic Wikipedia prose, each article is scored against a fixed reference distribution of 883 real Wikipedia articles using a leave-one-out (LOO) typicality procedure. For each of the 33 linguistic and structural features (lexical, sentence-level, POS-based, and Wikipedia-convention markers), a generated article's absolute z-score is computed relative to the reference distribution; the same $|z|$ is then recomputed for every reference article after holding it out from the rest of the reference set, yielding a null distribution of how extreme a real Wikipedia article can look on this feature. Typicality is defined as the fraction of held-out reference articles at least as extreme as the generated article, so a value of 1.0 means the generated article falls comfortably within the range of ordinary Wikipedia variation, while values near 0 flag the article as a genuine stylistic outlier relative to what real Wikipedia articles themselves exhibit under the same leave-one-out test. This self-referential design is deliberate: because the threshold for "extreme" is calibrated from the reference corpus's own internal variability rather than an arbitrary cutoff, a feature is only treated as truly divergent when it falls outside the range that real Wikipedia articles routinely produce among themselves. Following standard practice, features scoring below 0.05 typicality—i.e., more extreme than 95% of real Wikipedia articles on that same LOO test—are flagged as atypical and considered the only differences significant enough to warrant targeted correction; on average, just 2.4 of the 33 features (about 7%) per article cross this threshold, meaning the overwhelming majority of stylistic dimensions in the generated corpus are statistically indistinguishable from genuine Wikipedia writing.

Broken down by article type, main articles score noticeably lower on average typicality (mean 0.503) than sub-articles (mean 0.625), a gap that is plausible rather than concerning: a main article serves as the anchor overview page for an entire fictional event, and is therefore longer and structurally denser than a typical Wikipedia entry, while sub-articles—covering the people, organizations, and places pulled into the event—more closely resemble the length and structure of a Wikipedia page.

Across the 50 event types, average typicality ranges narrowly from 0.587 to 0.658, a spread of just 0.07 on a 0–1 scale, indicating that no single event type stands out as a systematic stylistic failure relative to the rest of the dataset. Crisis- and disaster-adjacent events such as migration_crisis, medical_breakthrough, climate_disaster, and debt_crisis cluster at the low end of the distribution, while narrative and cultural event types such as cultural_renaissance, global_sporting_event, new_philosophy, and new_literature cluster at the high end, suggesting that emotionally charged or rapidly evolving subject matter is mildly harder for the generation pipeline to render in a stylistically typical register than more narrative or descriptive subject matter.

Decomposing the same overall 0.619 headline by feature rather than by article reveals which of the 33 linguistic dimensions drive the corpus's typicality score. The dataset is essentially indistinguishable from real Wikipedia on tone and register: first- and second-person avoidance, peacock, weasel, and attribution phrasing, superlative density, passive voice, heading depth, list usage, and the presence of a substantial lead all score at or above 0.7 mean typicality, several approaching a perfect 1.0. The systematic weak points instead concentrate in word- and sentence-level style, with avg_word_length, long_word_ratio, adj_ratio, and readability_grade_level sitting far below every other feature, and a secondary dip in parenthetical_density_per_sent.

Critically, checking the direction of divergence for the four most atypical features against the reference means shows the generated prose is not simplified relative to real Wikipedia, as might be assumed, but rather consistently more complex: generated articles average 6.15 characters per word against a real-Wikipedia mean of 5.14, and 44.5% long words (seven or more characters) versus 29.8% in the reference corpus, with this pattern holding in essentially all 926 scored articles. The same holds for adjective usage (10.2% versus 5.9% of tokens) and overall readability, where the generated corpus averages a 19.6 grade level, corresponding to graduate-level reading difficulty, compared to 14.4, or college-level, for real Wikipedia, again in 925 of 926 articles. The one exception to this "more complex" pattern is parenthetical density, where the corpus under-uses parentheticals relative to real Wikipedia (0.09 versus 0.69 per sentence), meaning generated articles omit the brief inline asides—dates, alternate names, short clarifications—that real Wikipedia prose uses frequently. Thus, the generation pipeline over-produces dense, high-register vocabulary and under-produces the

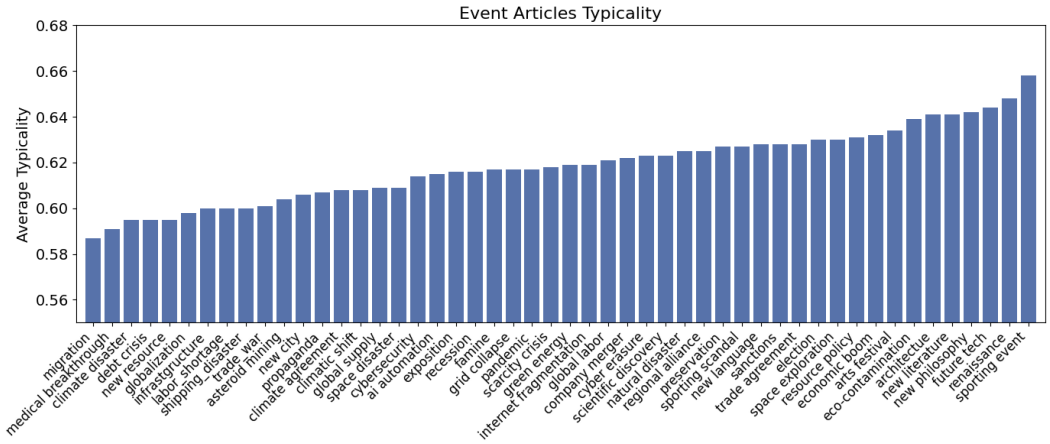


Figure 5: Average stylistic typicality of main articles and sub-articles for each of the 50 event types, sorted from lowest to highest. Typicality is the fraction of held-out real Wikipedia articles that are at least as extreme as the generated article, averaged over 33 features, so higher values are more Wikipedia-like. Scores fall in a narrow range (0.587–0.658), with crisis-related events at the low end and cultural events at the high end.

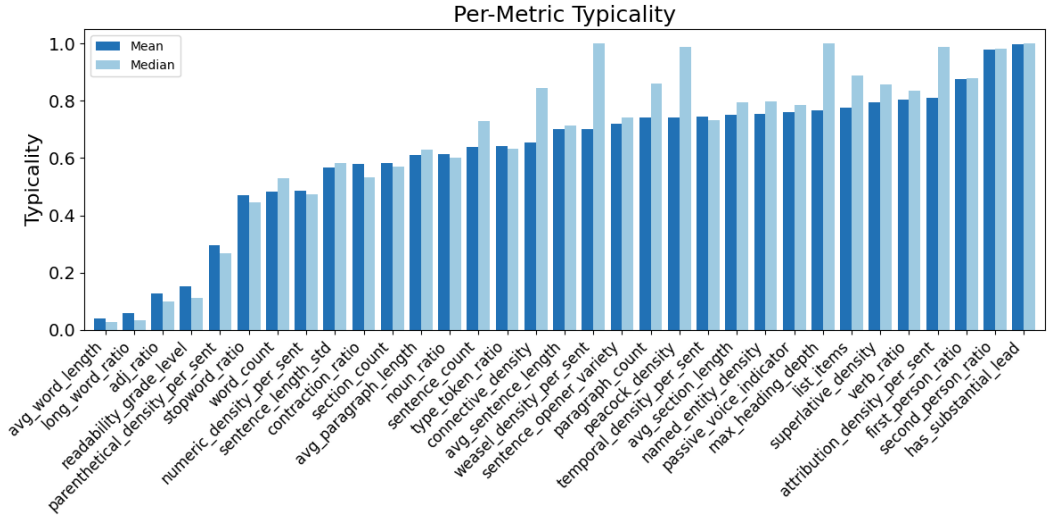


Figure 6: Mean typicality of each of the 33 stylistic features, averaged across all 926 generated articles and sorted from lowest to highest. Tone and register features closely match real Wikipedia, while word-level complexity features (average word length, long-word ratio, adjective ratio, and readability grade level) are the least typical, reflecting prose that is more complex than real Wikipedia.

small clarifying asides that make Wikipedia prose easier to parse, even though its broader structural and tonal conventions are already well-calibrated.

E.2 Manual Evaluation

Across five news events—each annotated from 15–20 source articles on a 1–5 Likert scale—the hand-annotated evaluation yields an overall mean score of **4.11** across all seven quality dimensions, with per-event averages of **3.86** for Pandemic, **4.29** for Propaganda, **4.29** for Regional Alliance, **4.29** for Trade War, and **3.86** for Migration Crisis, indicating that Propaganda, Regional Alliance, and Trade War were judged most consistently high-quality while Pandemic and Migration Crisis lagged slightly behind. Breaking this down by dimension, Adherence to Wikipedia Conventions

scored highest overall (mean **4.8**), reflecting strong structural and stylistic consistency across all events, followed by Entity Type Diversity, Information Density, and Factual Consistency (all mean **4.2**), while Redundancy scored lowest (mean **3.4**), suggesting repeated or overlapping information was the most persistent weakness across the corpus; Temporal Consistency (**4.0**) and Neutrality (**4.0**) fell in the middle, with Migration Crisis and Pandemic notably dragging down Neutrality and Factual/Temporal Consistency respectively, likely due to the more contentious, rapidly evolving, or emotionally charged nature of coverage for those two events compared to the more procedurally reported Propaganda, Regional Alliance, and Trade War stories.

Event	Entity Type Diversity	Temporal Consistency	Factual Consistency	Redundancy	Wikipedia Conventions	Neutrality	Info. Density
Pandemic	3	4	4	3	5	5	3
Propaganda	4	5	4	4	5	4	4
Regional Alliance	4	4	5	4	5	4	4
Trade War	5	4	5	3	4	4	5
Migration Crisis	5	3	3	3	5	3	5
Mean	4.2	4.0	4.2	3.4	4.8	4.0	4.2

Table 4: Human evaluation of five CASCADE events on seven quality dimensions, scored on a 1–5 Likert scale (higher is better). Each event was annotated using its 15–20 generated articles.

F TRAINING HYPERPARAMETERS

All runs used seed 12345, a linear warmup of 5–10 steps, and either a cosine schedule (preference-based objectives: DPO, OPD, OPD+gold, GRPO, contrastive distillation) or constant-with-warmup schedule (supervised objectives: CPT, QA-SFT, DISC). All runs trained for 10 epochs unless otherwise specified.

CPT. Learning rate 1.5×10^{-4}, constant-with-warmup, 8192-token context, 10 epochs, for both full-parameter and LoRA variants, over the fictional-event source articles.

QA-SFT / System-2 SFT. Same configuration as CPT (1.5×10^{-4}, constant-with-warmup, 8192 tokens, 10 epochs), differing only in training corpus (QA pairs or rationales vs. raw articles).

DPO. Sigmoid preference loss, $\beta = 0.1$ against a frozen reference, learning rate 5×10^{-7}, cosine decay, 10 epochs on `real_train`. Fixed across all starting checkpoints (base, full CPT, LoRA-CPT) to isolate the effect of pretraining. The LoRA variant uses learning rate 5×10^{-6}; all else unchanged. The two DPO controls (chosen/rejected swapped) use identical hyperparameters but train for 1 epoch.

OPD. Student trained on its own rollouts scored by a frozen teacher (Qwen3-32B or Llama-3.1-70B-Instruct), temperature 1.0, up to 512 new tokens, group size 1, rollout batch 32, over `real_train`. Learning rate 1×10^{-6}, cosine decay, 10 epochs. We add a λ-weighted forward-KL term on the gold answer,

$$\mathcal{L} = \mathcal{L}_{\text{OPD}} + \lambda \cdot \mathrm{KL}\big(T(\text{gold} \mid x) \,\|\, S(\text{gold} \mid x)\big),$$

with context extended to 2048 tokens. Main result uses $\lambda = 0.1$; we conduct a sweep over $\lambda \in \{0.1, 0.5, 1.0, 2.0\}$ and find that $\lambda = 0.1$ results in the highest accuracy and lowest degradations in abstention on our development set.

GRPO. Fully on-policy: policy and reference both initialized from the same 8B base, reference used only for the KL penalty. Group size 8, 32 prompts/batch (256 rollouts/step), temperature 1.0, top-p 1.0, minimum generation length 16 tokens. Reward from an LLM-as-judge (GPT-4.1) against the gold answer. Trained for max_steps$= 1755$ ($\approx$5 passes over the 11,430-prompt set), the closest exposure-matched analogue to 10 epochs. Learning rate 1×10^{-6}, cosine decay.

An early run collapsed to empty completions, which a judge misconfiguration scored as valid declines on the unanswerable subset (27.6% of prompts). Fixed by enforcing the minimum length,

Method	Fictional, Future Events				Real, Historical Events (parametric)			
	Single	Unanswerable	Open	Multi	Single	Unanswerable	Open	Multi
Qwen3-8B								
OPD Only	64.2	47.8	54.3	**44.4**	72.3	33.2	69.5	72.0
OPD+Gold Answer ($\lambda = 0.1$)	**64.5**	**53.8**	**60.0**	43.8	**74.1**	**34.5**	**70.0**	**74.1**
Llama-3.1-8B								
OPD Only	**54.3**	**42.4**	10.0	**33.6**	56.7	**87.8**	53.9	53.6
OPD+Gold Answer($\lambda = 0.1$)	54.0	36.2	**11.1**	32.9	**58.2**	82.8	**57.4**	**56.2**

Table 5: OPD with and without the gold-answer KL term ($\lambda = 0.1$) for Qwen3-8B and Llama-3.1-8B on fictional and real (closed-book) events. Values are accuracy, except for unanswerable questions, which report abstention rate. Adding the gold term gives small accuracy gains on most settings, but its effect on abstention depends on the model: it increases abstention for Qwen3-8B and decreases it for Llama-3.1-8B.

zeroing reward for empty completions without judge invocation, and tightening the judge rubric to require an explicit statement of epistemic limitation.

DISC. Distillation temperature $T = 2.0$ over prefix/suffix pairs against the model's own frozen initialization, learning rate 1×10^{-5}, constant-with-warmup, 8192-token context, 1 epoch (matching the source paper's schedule).

LoRA configuration. All LoRA variants (CPT, DPO, QA-SFT, and their chained combinations) use rank 64, $\alpha = 128$. Only DPO-LoRA's learning rate departs from its full-parameter counterpart; all other LoRA runs retain the full-parameter schedule and epoch count.

F.1 On Policy Distillation Details

We compare OPD without the gold term on LLaMA-3.1-8B-Instruct and Qwen3-8B. Adding the gold-KL term gives a small, consistent accuracy gain on single/multi-event questions in real and fictional question sets, but its effect on abstention is model-dependent: it *raises* correct-refusal for Qwen ($47.8\% \rightarrow 53.8\%$ fictional) but *lowers* it for Llama (real: $87.8\% \rightarrow 82.8\%$; fictional: $42.4\% \rightarrow 36.2\%$).

Comparing the four λ values (Qwen only, fictional/RAG k=10): $\lambda = 0.1$ has the best single-event accuracy at 64.5%, ahead of $\lambda = 1.0$ (64.1%), $\lambda = 0.5$ (63.8%), and $\lambda = 2.0$ (63.5%). On multi-event, $\lambda = 2.0$ edges out the rest at 44.8%, with $\lambda = 0.5$ at 44.3%, $\lambda = 0.1$ at 43.8%, and $\lambda = 1.0$ lowest at 43.5%. On abstention, $\lambda = 1.0$ is highest at 55.4%, followed by $\lambda = 0.1$ at 53.8%, then $\lambda = 0.5$ at 46.4% and $\lambda = 2.0$ at 46.3%. Taking single-event, multi-event, and abstention together, $\lambda = 0.1$ is the best overall choice: it achieves the best single-event accuracy and is second on abstention rate with only $\lambda = 1.0$ being better.

G Open Ended Rubric Construction

We follow the rubric-generation and validation methodology of Shen et al. (2026) to build and check the per-type rubrics used for grading open-ended questions.

G.1 Rubric Generation

Open-ended questions admit no single correct string to match against: answers vary widely in phrasing, structure, and emphasis, and a good answer may satisfy several distinct forms of "correctness" at once (factual accuracy, causal reasoning, temporal grounding, completeness) while a superficially fluent answer may satisfy none. Fixed multi-criterion rubrics let us decompose "is this answer good" into a set of independently checkable properties, which both makes grading more holistic than a single scalar judgment and makes it decidable by an LLM judge rather than requiring exact-match or reference-answer comparison. Each open-ended question type (significance, timing, evaluation,

multi-answer vector, counter-policy) is therefore graded against a fixed rubric of weighted binary criteria, constructed offline in five stages using gpt-4.1 as the brainstorming model.

(1) Contrastive brainstorming: for ∼50 questions per type, paired with their event documents and 8 sample answers spanning a range of models, an LLM identifies candidate criteria that separate strong from weak answers in the pool; five independent passes yield ∼100 candidates per type.

(2) Clustering and consolidation: candidates are clustered into 13–20 semantically coherent groups per type, weighted by how many of the five runs surfaced each cluster, then consolidated and further merged across types into 10 universal evaluation dimensions.

(3) Breadth validation: the consolidated criteria are checked against ∼250 questions from ∼38 events for applicability (a real answer can meaningfully pass or fail) and specificity (pass/fail is decidable from the source documents); criteria failing either test are dropped or split.

(4) RRD refinement: adapting prior work's Recursive Rubric Decomposition (Shen et al., 2026), criteria are scored against a pool of sampled answers and iteratively decomposed (pass rate ≥ 0.75 is too coarse), filtered for redundancy ($\geq$70% semantic overlap with an existing criterion is rejected), and checked for polarity (a criterion is discarded if it favors a known-weak model over a known-strong one, or cannot distinguish a strong answer from a copy corrupted with 2–3 injected errors); this loop terminates after four iterations, ten cumulative rejections, or convergence.

(5) Weighting: each criterion receives a whitened-uniform weight derived from the covariance structure of the binary answer-by-criterion score matrix X, estimated from temperature-1.0 rollouts (not strong-vs-weak contrasts, which collapse onto a single quality axis). After centering and regularized shrinkage ($\Sigma_{\text{reg}} = 0.9\Sigma + 0.1\,\text{diag}(\Sigma)$), weights are set as $w \propto \Sigma_{\text{reg}}^{-1/2}\mathbf{1}$ (negatives clipped, normalized to sum to 1), so that criteria that co-vary with others are down-weighted relative to those that vary independently. Each criterion is stored with both this whitened-uniform weight and a plain uniform weight ($1/n$); a separate rubric set for real-event questions is adapted from these with uniform weights only. At grading time, an LLM judge returns a binary pass/fail vector per answer (abstentions score 0), and we compute both a uniform score (fraction of criteria passed) and a whitened-uniform score (weighted sum).

G.2 Meta-Evaluation Against Human Judgment

To verify that the automatic grader tracks human judgment, we compare grader and human rankings of candidate answers within each question, since the property that matters for benchmarking is relative ordering rather than absolute score. For all question types (counter-policy, significance, timing, evaluation, multi-answer vector), we sample 20 questions each, with 8 candidate answers per question spanning GPT-5.2, Gemini-2.5-Pro, and Qwen3-8B under four retrieval/training conditions (oracle-article RAG, article-chunk RAG, paragraph-chunk RAG, and continued-pretraining for Qwen), giving 160 answers per type and 100 question-level correlations overall. A blind annotator scores every answer on a 1–5 scale (half-points allowed), unaware of both the source system and the grader's score. For each question we compute the Spearman correlation between the human score vector and the grader's uniform and whitened-uniform score vectors; per-type correlations are pooled via Fisher z-transformation, with significance from a one-sample t-test on the z values.

All correlations are significant at $p < 10^{-5}$ (most $p < 10^{-10}$), and agreement is decent-to-high across the board, indicating the rubrics are reliable for evaluation. Agreement is strongest for counter-policy ($r \approx 0.78$), moderate for significance and evaluation ($r \approx 0.67$), and weakest for timing ($r \approx 0.48$–0.56) and multi-answer vector ($r \approx 0.54$), consistent with timing questions depending on chains of dated preconditions that both humans and the grader score inconsistently. The uniform and whitened-uniform weightings perform comparably: uniform slightly outperforms weighted on counter-policy, significance, multi-answer vector, and evaluation (by 0.01–0.04), and whitened-uniform helps only on timing (0.48 → 0.56). Across all question types, the difference vanishes (0.66 vs. 0.66). Given this, we adopt the uniform score as our primary metric for its simplicity and independence from covariance-pool composition.

Method	Fictional, Future Events									Real, Historical Events								
	S-Hop	M-Src	M-Hop	Agg.	Cause	Comp.	Temp.	Persp.	Total	S-Hop	M-Src	M-Hop	Agg.	Cause	Comp.	Temp.	Persp.	Total
GPT-5.2																		
Full Information	**97.4**	**87.9**	58.1	**35.6**	69.3	**74.5**	**79.3**	97.4	73.2	93.7	89.5	54.6	73.8	91.9	93.3	48.6	100.0	83.0
RAG (articles)	96.4	87.1	**60.1**	32.6	**73.9**	72.6	77.6	**98.6**	**73.4**	**93.7**	**89.5**	**54.6**	**73.8**	**91.9**	**93.3**	**48.6**	**100.0**	**83.0**
RAG (paragraphs)	96.1	86.7	55.3	28.4	72.1	68.2	77.5	97.8	71.3	93.7	89.5	54.6	73.8	91.9	93.3	48.6	100.0	83.0
Gemini-2.5 Pro																		
Full Information	87.5	**70.9**	59.0	**36.1**	53.6	**47.7**	**75.8**	**96.6**	63.8	93.7	88.5	71.1	83.3	87.6	94.4	64.2	98.8	85.6
RAG (articles)	**89.8**	70.0	**59.7**	32.8	**56.8**	**47.7**	74.1	**96.6**	**64.0**	**93.7**	**88.5**	**71.1**	**83.3**	**87.6**	**94.4**	**64.2**	**98.8**	**85.6**
RAG (paragraphs)	85.1	67.0	53.0	30.1	54.9	34.2	74.1	95.4	59.9	93.7	88.5	71.1	83.3	87.6	94.4	64.2	98.8	85.6
Qwen 3 32B																		
RAG (articles)	**95.0**	82.4	49.1	36.1	70.2	**66.4**	67.6	98.0	68.8	**83.2**	**84.3**	**67.4**	**69.0**	87.2	**88.8**	**52.0**	**97.6**	**81.1**
DPO	94.4	**85.8**	**53.3**	**37.7**	**73.5**	65.4	**69.4**	**99.2**	**70.8**	79.6	84.0	53.9	**69.0**	**88.4**	87.9	43.9	94.0	78.1
Qwen 3 8B																		
Full Information	92.0	77.4	39.0	29.6	65.9	55.7	69.4	96.8	63.7	70.7	76.3	47.7	66.7	84.1	85.4	33.1	94.0	72.6
RAG (articles)	92.0	76.9	33.4	27.7	66.1	57.6	66.9	96.2	62.5	70.7	76.3	47.7	**66.7**	**84.1**	**85.4**	33.1	94.0	72.6
RAG (paragraphs)	90.5	76.0	28.6	20.2	63.4	49.0	66.7	95.0	59.1	70.7	76.3	47.7	66.7	84.1	85.4	33.1	94.0	72.6
IRCoT	83.5	53.1	41.8	30.0	63.9	51.4	66.7	85.8	58.1	70.7	76.3	47.7	66.7	84.1	85.4	33.1	94.0	72.6
DPO	90.8	82.2	44.6	28.7	**70.1**	51.8	69.0	97.8	**65.3**	62.3	75.6	25.0	52.4	79.5	76.7	29.7	81.0	64.2
DPO LoRA	90.3	**82.5**	**45.0**	28.2	68.8	49.9	**70.0**	98.0	65.0	58.1	72.1	31.9	35.7	74.4	75.0	18.2	77.4	61.2
OPD	**93.2**	77.5	40.9	**30.8**	68.6	58.4	65.1	95.6	64.5	**72.3**	**76.7**	61.8	57.1	83.2	84.8	29.1	**95.2**	74.1
GRPO	91.7	78.2	38.3	27.1	68.9	**60.1**	63.6	**98.2**	63.8	57.6	67.6	30.9	40.5	76.9	71.1	25.7	85.7	61.4
CPT	20.8	12.6	5.0	0.4	28.3	3.6	9.7	42.2	14.5	4.2	1.4	2.0	0.1	0.7	0.1	2.7	4.8	1.5
CPT LoRA	15.5	4.5	8.4	0.5	19.3	5.8	9.5	14.6	10.1	62.3	71.4	50.3	54.8	80.7	81.7	31.8	86.9	69.2
CPT → DPO	25.1	23.3	3.8	0.5	32.8	7.9	13.2	30.8	17.3	4.2	2.8	0.3	0.1	7.5	3.4	0.1	4.8	3.8
CPT → OPD	27.5	20.8	3.3	1.1	20.9	8.1	6.3	21.2	13.6	70.2	64.5	**81.6**	**66.7**	82.5	76.4	**61.5**	73.8	**75.2**
CPT LoRA → QA SFT LoRA	34.7	19.5	15.3	0.5	50.9	16.8	10.3	18.0	22.8	42.4	49.1	16.4	19.0	65.0	52.2	4.1	65.5	45.3
CPT LoRA → DPO LoRA	44.2	49.5	25.6	2.9	46.0	28.4	15.0	53.6	32.9	56.5	66.6	24.7	38.1	56.6	64.0	21.6	71.4	52.1
DiSC	39.2	36.8	11.2	0.6	49.4	21.8	16.6	42.8	27.6	61.3	55.7	67.4	42.9	65.0	60.7	37.8	73.8	60.8
System2-FT	3.5	0.1	0.3	0.3	2.7	0.3	1.3	1.4	1.3	0.5	0.1	0.1	0.1	0.4	0.1	0.1	0.1	0.2
LLaMA-3.1-8B Instruct																		
Full Information	82.6	59.9	29.2	**22.9**	58.8	**42.5**	62.0	89.4	53.9	58.1	63.1	20.4	33.3	62.0	68.5	22.3	72.6	53.5
RAG (articles)	83.7	**61.2**	28.9	20.8	57.0	42.4	56.5	89.8	53.0	58.1	63.1	20.4	33.3	62.0	68.5	22.3	72.6	53.5
RAG (paragraphs)	82.4	56.6	24.4	15.5	48.8	30.4	55.9	88.2	48.0	58.1	63.1	20.4	33.3	62.0	68.5	22.3	72.6	53.5
IRCoT	76.0	46.1	32.7	15.2	61.8	41.9	57.6	80.8	50.3	58.1	63.1	20.4	33.3	62.0	68.5	22.3	72.6	53.5
DPO	83.2	59.9	**35.2**	21.7	60.5	38.6	58.1	**90.0**	**54.2**	56.5	**70.4**	**29.3**	38.1	69.2	63.2	20.3	71.4	56.7
DPO LoRA	82.1	55.6	29.3	**22.9**	58.9	40.7	55.4	88.0	52.2	54.5	63.8	26.0	**42.9**	66.0	60.4	27.0	69.0	54.2
OPD	**84.2**	59.0	29.6	19.1	**63.1**	39.3	62.1	86.8	53.9	56.5	69.3	27.3	**42.9**	68.7	**71.9**	23.0	**77.4**	**58.2**
GRPO	78.8	49.1	21.3	12.2	55.8	34.6	**65.7**	84.0	48.4	**62.3**	62.7	21.7	35.7	**72.3**	65.2	**34.5**	76.2	57.5
CPT	5.0	3.2	2.7	0.7	5.1	0.6	4.5	20.8	4.5	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
CPT LoRA	45.8	44.0	15.7	2.4	37.7	27.6	9.8	43.6	27.9	56.5	56.8	17.4	28.6	64.1	**63.2**	14.9	**76.2**	51.1
CPT → DPO	1.9	1.4	3.1	0.1	1.0	0.1	2.0	0.6	1.4	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
CPT → OPD	2.8	0.4	1.3	0.1	8.8	1.3	2.8	1.0	2.8	0.1	0.1	0.1	0.1	0.9	0.1	0.1	0.1	0.3
CPT LoRA → SFT LoRA	23.3	9.7	15.0	0.3	34.1	9.0	6.2	10.2	15.1	47.1	49.1	25.3	19.0	59.4	50.8	8.8	60.7	45.4
CPT LoRA → DPO LoRA	34.2	34.4	21.3	2.5	38.7	19.0	10.2	39.6	25.2	57.1	62.0	**29.3**	26.2	62.9	59.3	16.9	76.2	52.7
DiSC	22.3	10.1	8.9	0.9	18.7	5.7	9.4	21.4	12.1	13.1	6.6	3.3	0.1	8.2	5.1	1.4	9.5	6.5
System2 finetuning	0.1	0.1	0.5	0.1	0.5	0.3	0.8	0.1	0.3	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

Table 6: Fine-grained results on single-event reasoning subtypes (Single-Hop, Multi-Source Synthesis, Multi-Hop, Aggregative, Causal, Comparative, Temporal, Perspectival) for fictional and real events. Row color indicates method family: retrieval/RAG methods, reasoning-training methods, CPT-based methods. Bold marks the best value in each column within each model's block.

H Additional Results

For all models, we find that single-event and multi-event accuracy is largely insensitive to retrieval setup (e.g., 4.6-point spread for Qwen3-8B) and to embedding choice (BGE vs. BM25 differ by only 0.1 points at $k = 10$).

LoRA-Based Methods. We also test LoRA and context-distillation variants on our benchmark: **DPO LoRA**, **QA-SFT LoRA**, **CPT LoRA**, chained variants (**CPT LoRA → DPO/QA-SFT LoRA**), and two context-distillation methods, **Sys2-FT** (Park et al., 2025) and **DiSC** (Padmanabhan et al., 2026), which distill in-context behavior into weights via synthetic QA data and KL-divergence teacher-student training, respectively (generator: GPT-4.1; teacher: Qwen3-32B/LLaMA-3.1-70B). CPT LoRA variants are evaluated parametrically, as in the full-parameter setup.

Tables 6, 7, 9, and 8 show the same acquisition/retention trade-off seen with full-parameter training, but attenuated. CPT LoRA barely acquires new knowledge (10.1% fictional single-event accuracy) while nearly preserving real-event performance (69.2%), confirming that low-rank updates curb catastrophic forgetting relative to full-parameter CPT at the cost of encoding little new information. DPO LoRA shows the opposite profile and is the strongest LoRA method overall: 65.0%

Method	Fictional, Future Events								Real, Historical Events							
	Info.	Ent.	Event	Presup.	Time	Indet.	Corr.	Total	Info.	Ent.	Event	Presup.	Time	Indet.	Corr.	Total
GPT-5.2																
Full Information	**84.2**	**93.7**	**80.3**	**82.6**	**86.3**	**40.7**	**79.0**	**78.3**	79.3	47.6	57.4	41.2	37.8	60.8	81.6	59.1
RAG (articles)	83.0	90.2	77.8	80.3	**86.3**	38.8	78.2	76.5	**79.3**	**47.6**	**57.4**	**41.2**	**37.8**	**60.8**	**81.6**	**59.1**
RAG (paragraphs)	83.0	89.9	71.1	79.3	81.3	40.3	75.9	74.5	79.3	47.6	57.4	41.2	37.8	60.8	81.6	59.1
Gemini-2.5 Pro																
Full Information	**89.4**	98.4	**97.6**	94.7	**94.1**	**60.3**	99.1	**90.7**	61.7	74.2	80.2	79.7	75.6	60.6	82.4	73.4
RAG (articles)	85.7	97.8	96.2	93.8	92.8	57.1	98.8	89.1	**61.7**	**74.2**	**80.2**	**79.7**	**75.6**	**60.6**	**82.4**	**73.4**
RAG (paragraphs)	86.0	**98.5**	96.3	**94.8**	93.4	58.9	**99.6**	89.9	61.7	74.2	80.2	79.7	75.6	60.6	82.4	73.4
Qwen 3 32B																
RAG (articles)	64.8	**76.0**	48.1	54.6	56.4	30.9	**82.3**	59.2	51.9	21.6	24.6	25.8	26.6	47.8	51.0	36.1
DPO	**78.5**	75.3	**57.5**	**73.6**	**59.4**	**42.2**	78.3	**66.5**	**65.2**	**45.0**	**47.2**	**46.1**	**35.6**	**62.2**	**73.2**	**54.3**
Qwen 3 8B																
Full Information	64.7	87.6	56.4	48.5	57.3	26.9	**94.3**	62.5	56.9	23.8	24.8	29.6	27.1	47.8	63.8	39.6
RAG (articles)	63.6	85.5	54.0	48.6	55.9	31.8	88.7	61.4	56.9	23.8	24.8	29.6	27.1	47.8	63.8	39.6
RAG (paragraphs)	65.6	80.4	52.8	47.9	44.5	36.5	86.1	59.3	56.9	23.8	24.8	29.6	27.1	47.8	63.8	39.6
IRCoT	54.1	85.5	50.5	19.6	56.1	29.5	74.2	53.0	56.9	23.8	24.8	29.6	27.1	47.8	63.8	39.6
DPO	83.6	**94.2**	**79.8**	72.9	73.3	49.6	84.3	**77.0**	86.7	61.0	49.8	44.1	38.1	90.6	70.6	64.4
DPO LoRA	**83.9**	89.5	72.7	**77.3**	77.3	50.8	80.4	76.1	74.5	50.2	46.2	39.4	32.9	77.6	49.2	53.8
OPD	59.1	75.8	37.9	41.5	57.2	27.8	76.2	53.8	59.8	18.2	23.4	22.3	23.6	41.4	50.0	34.5
GRPO	52.5	34.7	14.0	47.1	59.3	9.5	28.7	35.0	99.0	96.0	88.4	64.6	70.4	96.6	96.4	88.9
CPT	19.0	0.8	2.0	9.4	2.2	25.7	9.1	9.6	32.9	3.4	5.4	10.1	7.1	7.0	3.0	9.4
CPT LoRA	10.5	2.0	0.1	3.7	36.7	22.5	24.0	14.1	31.7	10.6	11.0	23.5	16.7	45.0	53.4	28.0
CPT → DPO	23.4	2.3	1.8	15.6	15.4	15.1	2.8	10.7	33.1	29.6	24.6	25.2	21.1	36.2	12.8	26.2
CPT → OPD	35.3	12.6	10.2	7.1	18.4	41.2	35.0	22.6	62.6	25.4	28.8	29.6	34.2	52.2	68.4	43.6
CPT LoRA → QA SFT LoRA	62.5	63.4	54.7	77.2	**85.2**	60.3	58.4	65.9	**100.0**	**99.4**	**95.8**	**73.6**	**74.2**	**98.6**	**98.0**	**92.8**
CPT LoRA → DPO LoRA	65.5	35.5	15.0	36.3	58.2	**77.4**	57.1	48.9	81.9	31.4	25.2	22.0	29.0	88.2	65.2	50.4
DiSC	26.9	3.1	2.2	5.6	32.1	30.8	35.4	19.2	70.2	9.0	7.8	20.6	16.2	61.4	56.2	35.0
System2-FT	2.7	2.9	3.1	5.2	2.6	4.4	2.5	3.3	2.4	11.6	8.4	11.0	11.2	0.2	0.2	6.1
LLaMA-3.1-8B Instruct																
Full Information	59.5	81.4	46.6	39.5	36.7	37.5	88.5	55.9	94.3	82.0	78.6	68.1	70.7	91.4	86.0	82.4
RAG (articles)	62.7	83.1	61.0	45.0	37.4	42.2	85.3	59.8	94.3	82.0	78.6	68.1	70.7	91.4	86.0	82.4
RAG (paragraphs)	67.3	**90.5**	67.0	53.1	37.1	35.8	88.7	63.1	94.3	82.0	78.6	68.1	70.7	91.4	86.0	82.4
IRCoT	48.6	76.5	42.2	29.4	39.3	25.1	73.0	48.0	94.3	82.0	78.6	68.1	70.7	91.4	86.0	82.4
DPO	74.9	70.3	46.0	48.6	38.3	45.5	61.2	55.0	28.8	14.8	13.2	10.7	10.4	9.4	13.8	14.4
DPO LoRA	63.2	63.0	30.3	32.8	21.3	41.1	54.3	43.7	27.6	4.2	3.4	7.8	6.8	23.8	6.2	11.4
OPD	48.9	50.6	18.3	22.8	24.3	39.4	49.4	36.2	**100.0**	83.0	79.8	58.0	67.9	96.8	85.0	82.8
GRPO	71.1	87.5	75.8	58.2	55.0	51.5	72.9	67.6	**100.0**	**99.8**	**97.8**	**87.5**	**89.6**	**98.4**	97.4	**96.4**
CPT	32.0	5.1	2.5	10.8	5.7	9.9	3.8	9.8	2.1	0.2	1.0	1.7	2.7	0.6	0.4	1.2
CPT LoRA	78.3	9.0	9.2	35.2	46.0	63.8	58.8	42.4	96.4	26.6	46.4	51.9	44.1	80.0	83.4	61.6
CPT → DPO	35.9	8.3	6.3	11.1	6.7	11.2	5.5	12.0	3.1	0.6	0.2	0.6	3.3	0.2	0.4	1.1
CPT → OPD	84.0	43.2	43.2	45.0	66.1	52.4	28.4	51.4	96.2	56.6	57.8	40.6	60.3	81.8	59.4	65.2
CPT LoRA → QA SFT LoRA	69.5	73.9	80.2	**80.0**	86.0	**80.9**	59.8	75.7	**100.0**	96.2	94.4	76.5	76.2	96.4	**98.6**	92.3
CPT LoRA → DPO LoRA	**87.4**	70.4	**82.6**	67.7	**91.5**	67.6	**93.6**	**80.1**	82.4	38.0	37.4	29.0	27.4	41.0	47.8	43.7
DiSC	57.9	1.8	2.7	16.5	13.6	54.1	29.8	24.7	96.4	4.8	3.6	5.8	7.1	66.6	8.4	27.7
System2-FT	12.8	4.1	3.5	10.9	6.8	8.4	3.3	7.1	20.7	5.0	4.0	18.0	28.5	1.0	1.6	9.9

Table 7: Fine-grained results on unanswerable question subtypes (Missing Information, Non-Existent Entity, Novel Event, False Presupposition, Temporally Impossible, Inherently Indeterminate, Spurious Correlation) for fictional and real events. Row color indicates method family: retrieval/RAG methods, reasoning-training methods, CPT-based methods. Bold marks the best value in each column within each model's block.

fictional single-event and 78.6% open-ended accuracy, while retaining strong real-event performance (61.2% single-event, 80.8% multi-event, the best real multi-event score among Qwen3-8B methods)—indicating that preference optimization on a frozen base acquires new knowledge more effectively than continued pretraining, without full-parameter updates.

Chaining CPT LoRA with a second training stage does not close this acquisition gap and instead erodes real-event retention. For Qwen3-8B, CPT LoRA → DPO LoRA raises fictional single-event accuracy over CPT LoRA alone (32.9% vs. 10.1%) but drops real single-event accuracy from 69.2% to 52.1% and real multi-event accuracy from 75.8% to 59.5%; CPT LoRA → QA-SFT LoRA fares worse still, collapsing fictional open-ended accuracy to 5.3%. LLaMA-3.1-8B shows the same pattern, with both chained variants underperforming direct DPO LoRA on fictional single-event accuracy (25.2% and 15.1% vs. 52.2%) and falling to near-floor fictional multi-event accuracy (3.8%, 4.6%). Unanswerable-abstention accuracy is also highly unstable across chained methods (e.g., CPT LoRA → DPO LoRA: 48.9% fictional vs. 43.7% real abstention for Qwen3-8B vs. LLaMA-3.1-8B), suggesting sequential LoRA composition interacts unpredictably with calibration rather than cleanly combining each stage's benefits.

Method	Fictional, Future Events							Real, Historical Events						
	Ent.	Inf.	Chain	Agg.	Temp.	M-Hop	Total	Ent.	Inf.	Chain	Agg.	Temp.	M-Hop	Total
GPT-5.2														
Full Information	–	–	–	–	–	–	–	99.4	88.8	66.7	63.0	74.8	94.4	81.2
RAG (articles)	43.7	58.5	**19.3**	**11.4**	62.1	87.0	46.8	**99.4**	**88.8**	**66.7**	**63.0**	**74.8**	**94.4**	**81.2**
RAG (paragraphs)	**44.5**	**60.3**	14.5	11.2	**90.4**	**96.8**	**53.5**	99.4	88.8	66.7	63.0	74.8	94.4	81.2
Gemini-2.5 Pro														
Full Information	–	–	–	–	–	–	–	98.0	84.6	51.2	58.8	84.2	89.7	78.1
RAG (articles)	31.2	35.0	**12.0**	**2.5**	49.7	46.7	29.7	**98.0**	**84.6**	**51.2**	**58.8**	**84.2**	**89.7**	**78.1**
RAG (paragraphs)	**34.2**	**38.7**	9.3	1.7	**90.3**	**54.1**	**39.3**	98.0	84.6	51.2	58.8	84.2	89.7	78.1
Qwen 3 32B														
RAG (articles)	34.2	**77.5**	18.8	9.2	59.8	80.3	46.8	**99.0**	**96.2**	85.6	62.4	77.4	**96.5**	85.9
DPO	**37.8**	77.2	**20.4**	**13.2**	**63.3**	**84.7**	**49.6**	97.8	92.8	**86.5**	**81.4**	**87.2**	92.3	**89.7**
Qwen 3 8B														
Full Information	–	–	–	–	–	–	–	94.0	85.2	63.0	38.4	66.0	85.8	71.9
RAG (articles)	33.9	67.9	11.0	6.1	57.7	72.2	41.9	94.0	85.2	63.0	38.4	66.8	85.8	71.9
RAG (paragraphs)	33.4	67.4	10.6	4.1	**81.4**	**85.6**	**47.8**	94.0	85.2	63.0	38.4	66.0	85.8	71.9
IRCoT	**36.7**	63.1	5.3	5.4	64.3	63.9	40.7	94.0	85.2	63.0	38.4	66.0	85.8	71.9
DPO	30.8	58.2	12.8	4.5	56.1	73.5	39.4	91.0	78.6	**78.4**	56.4	71.8	85.8	76.7
DPO LoRA	32.2	55.4	**15.1**	4.8	54.5	76.3	39.6	94.6	**85.4**	77.9	**60.6**	**78.8**	**88.1**	**80.8**
OPD	34.3	73.0	12.1	5.4	57.9	78.2	43.8	**95.2**	85.2	70.9	42.6	67.2	84.6	74.1
GRPO	32.3	**82.8**	6.1	**7.2**	54.6	69.7	42.9	86.0	82.6	70.9	49.6	64.2	76.0	71.5
CPT	21.2	26.0	0.5	0.1	24.8	24.6	16.8	2.0	4.6	3.0	1.8	5.2	0.9	3.0
CPT LoRA	11.8	59.9	2.0	2.2	43.9	68.5	31.8	94.2	**85.4**	64.7	55.2	69.4	85.5	75.8
CPT → DPO	6.2	24.4	0.7	0.1	29.6	27.0	15.1	6.0	7.6	5.1	1.2	18.8	1.4	6.9
CPT → OPD	2.2	12.8	0.2	0.1	19.0	4.1	7.0	56.4	47.8	55.1	24.8	51.8	51.5	47.6
CPT LoRA → QA SFT LoRA	2.0	56.4	1.2	0.1	26.7	4.7	16.6	54.6	47.4	23.7	31.2	43.4	36.6	39.9
CPT LoRA → DPO LoRA	10.1	35.0	4.2	1.0	41.2	65.7	26.1	74.4	61.8	54.7	38.4	62.2	65.7	59.5
DiSC	2.8	41.4	0.3	0.2	29.5	39.9	19.5	69.6	47.8	57.0	36.6	69.4	63.6	57.2
System2-FT	2.4	4.0	0.5	0.2	2.6	6.2	2.6	0.0	0.2	0.0	0.2	0.4	0.0	0.1
LLaMA-3.1-8B Instruct														
Full Information	–	–	–	–	–	–	–	85.6	63.0	40.9	32.0	49.8	57.1	55.0
RAG (articles)	31.1	39.5	3.3	2.1	48.0	55.1	30.2	85.6	63.0	40.9	**32.0**	49.8	57.1	55.0
RAG (paragraphs)	31.9	30.1	3.3	1.6	**67.3**	**66.7**	33.9	85.6	63.0	40.9	32.0	49.8	57.1	55.0
IRCoT	**33.8**	**61.3**	2.5	**2.9**	54.5	43.2	**34.2**	85.6	63.0	40.9	32.0	49.8	57.1	55.0
DPO	28.7	39.0	**8.1**	1.8	52.5	55.2	31.2	**87.0**	**65.0**	34.4	29.2	**59.4**	55.0	55.5
DPO LoRA	27.1	60.8	5.6	1.4	42.3	64.2	33.8	84.0	54.8	26.7	25.6	58.0	34.0	48.0
OPD	29.5	53.7	2.2	1.8	52.0	53.8	32.9	85.2	**65.0**	**41.2**	27.0	56.9	**60.1**	**56.2**
GRPO	30.8	43.6	3.0	1.4	51.2	45.5	29.9	71.0	52.6	36.5	28.4	53.0	55.9	49.7
CPT	4.6	3.1	1.0	0.1	8.7	4.9	3.8	0.1	0.1	0.1	0.1	0.1	0.1	0.1
CPT LoRA	12.5	3.1	2.3	**2.9**	37.0	61.6	19.5	80.8	53.4	34.4	28.4	50.8	51.7	50.3
CPT → DPO	1.9	0.8	0.1	0.1	4.3	6.6	2.2	0.1	0.1	0.1	0.1	0.1	0.1	0.1
CPT → OPD	1.1	3.3	0.1	0.1	3.8	0.1	1.5	0.1	0.1	0.1	0.1	0.4	0.1	0.1
CPT LoRA → QA SFT LoRA	0.3	5.4	0.0	0.1	17.6	1.0	4.6	31.2	34.8	18.4	19.0	47.8	25.9	29.9
CPT LoRA → DPO LoRA	0.4	0.6	0.8	0.1	11.6	8.7	3.8	80.2	57.6	27.0	29.8	58.4	43.6	50.1
DiSC	2.8	12.0	0.0	0.1	18.5	2.7	6.6	5.2	2.4	0.0	1.0	1.0	1.4	1.9
System2-FT	0.4	0.5	0.2	0.1	1.1	0.1	0.4	0.1	0.1	0.1	0.1	0.1	0.1	0.1

Table 8: Fine-grained results on multi-event reasoning subtypes (Shared Entity, Cross-Event Inferential, Cross-Event Causal Chain, Cross-Event Aggregative, Cross-Event Temporal, Long-Chain Multi-Hop) for fictional and real events. Row color indicates method family: retrieval/RAG methods, reasoning-training methods, CPT-based methods. Bold marks the best value in each column within each model's block.

Among context-distillation methods, DiSC clearly outperforms Sys2-FT on both fictional acquisition and real-event retention, though neither matches the base model or DPO LoRA. Sys2-FT collapses almost entirely (single-event accuracy under 3% on both fictional and real events, for both models). DiSC performs much better: 27.6% fictional single-event accuracy on Qwen3-8B with strong real-event retention (60.8% single-event, 57.2% multi-event), and the same ordering on LLaMA-3.1-8B at lower absolute scores (12.1% fictional, 6.5% real single-event). DiSC's real-event retention rivals several LoRA variants, though it still trails the base model's RAG performance.

QA-SFT. We additionally tested QA-SFT under two training regimes: LoRA and full-parameter fine-tuning. **QA-SFT LoRA**, applied on top of retrieval, fails to improve over base retrieval performance for Qwen3-8B: single-event accuracy drops from 62.5% to 55.4% and open-ended rubric score falls sharply from 63.5% to 22.1%, with only a modest gain in unanswerable abstention (61.4% → 78.3%) that comes at the cost of the other three question types. This suggests that supervised fine-tuning on QA pairs alone, without an explicit preference or policy signal, is insufficient to teach the reasoning behavior needed for event-based questions—it appears to bias the model toward abstention rather than genuinely improving its ability to reason over retrieved evidence. **QA-SFT (full**

Method	Fictional, Future Events						Real, Historical Events					
	Sign.	Time	Eval.	M-Ans	Policy	Total	Sign.	Time	Eval.	M-Ans	Policy	Total
GPT-5.2												
Full Information	**71.4**	**66.1**	**88.7**	55.3	**86.1**	**73.4**	89.7	90.6	94.1	88.4	88.7	90.3
RAG (articles)	69.7	65.9	87.1	**55.4**	84.5	72.4	**89.7**	**90.6**	**94.1**	**88.4**	**88.7**	**90.3**
RAG (paragraphs)	64.4	64.0	84.7	53.5	83.2	69.8	89.7	90.6	94.1	88.4	88.7	90.3
Gemini-2.5 Pro												
Full Information	**39.6**	**26.0**	**39.7**	**43.4**	**28.8**	**36.0**	85.2	78.8	86.2	81.6	77.2	81.8
RAG (articles)	32.8	24.3	36.8	40.7	21.0	31.5	**85.2**	**78.8**	**86.2**	**81.6**	**77.2**	**81.8**
RAG (paragraphs)	24.5	23.7	29.8	40.1	10.9	26.2	85.2	78.8	86.2	81.6	77.2	81.8
Qwen 3 32B												
RAG (articles)	75.4	51.8	81.7	**85.9**	52.7	70.4	**93.4**	67.9	**88.4**	**86.2**	**83.9**	**84.2**
DPO	**83.2**	**70.0**	**89.5**	**87.0**	**82.0**	**82.7**	79.9	**74.8**	75.3	85.7	80.6	79.4
Qwen 3 8B												
Full Information	79.1	47.1	87.8	79.3	52.4	70.2	90.1	54.6	78.6	78.8	69.1	74.5
RAG (articles)	69.3	43.9	78.5	76.7	44.4	63.5	**90.1**	**54.6**	78.6	**78.8**	69.1	74.5
RAG (paragraphs)	46.1	38.5	59.6	65.1	30.5	48.6	90.1	54.6	78.6	78.8	69.1	74.5
IRCoT	24.9	13.7	32.1	21.1	18.3	22.3	90.1	54.6	78.6	78.8	69.1	74.5
DPO	68.8	59.7	83.9	78.3	**83.1**	75.0	83.6	46.4	71.7	77.4	86.2	73.5
DPO LoRA	**83.5**	**62.3**	**88.8**	**82.5**	72.7	**78.5**	75.7	48.8	66.3	77.8	76.1	69.3
OPD	66.7	42.6	76.4	76.3	32.6	60.0	88.4	49.1	76.0	75.1	59.8	70.0
GRPO	62.0	31.4	70.7	77.5	54.5	60.2	85.1	44.3	**80.6**	74.2	83.5	74.0
CPT	1.1	3.1	3.2	7.1	2.0	3.3	1.1	0.2	0.4	0.3	0.9	0.6
CPT LoRA	11.1	6.3	26.4	24.9	43.5	22.4	87.2	48.3	80.4	75.8	86.3	**76.0**
CPT → DPO	1.3	0.1	1.5	2.3	3.0	1.7	7.3	1.5	2.5	7.6	6.7	5.2
CPT → OPD	0.5	0.1	1.1	3.3	0.7	1.2	23.9	6.7	13.6	18.3	5.1	13.6
CPT LoRA → QA SFT LoRA	1.6	0.2	2.6	9.0	13.3	5.3	70.6	44.4	65.4	50.8	**91.1**	64.8
CPT LoRA → DPO LoRA	11.9	6.9	18.0	28.7	35.0	20.2	65.6	42.1	50.1	62.3	57.3	55.7
DiSC	1.3	2.0	9.7	18.0	10.3	8.4	60.5	21.4	57.0	56.4	64.5	52.4
System2-FT	0.1	0.1	0.1	0.1	0.2	0.1	0.1	0.1	0.2	0.4	0.2	0.2
LLaMA-3.1-8B Instruct												
Full Information	11.4	6.4	12.9	30.4	6.0	13.8	71.6	28.1	53.3	62.8	40.3	51.6
RAG (articles)	12.1	8.4	13.3	**34.7**	5.3	15.1	71.6	28.1	53.3	62.8	40.3	51.6
RAG (paragraphs)	10.2	9.4	18.3	31.8	6.3	15.5	71.6	28.1	53.3	62.8	40.3	51.6
IRCoT	7.2	2.3	6.3	8.2	5.8	6.1	71.6	28.1	53.3	62.8	40.3	51.6
DPO	13.7	**13.8**	17.9	26.8	16.1	**17.8**	58.5	**46.5**	45.8	**68.4**	50.8	54.1
DPO LoRA	**15.7**	8.8	**21.4**	27.3	12.9	17.6	56.9	28.9	43.3	60.4	58.3	49.9
OPD	7.4	4.7	7.4	29.6	4.5	11.0	**73.0**	33.5	61.7	67.2	49.7	**57.4**
GRPO	10.2	4.3	12.9	25.2	6.0	12.1	64.8	35.7	44.3	65.2	35.7	49.4
CPT	0.1	0.2	0.1	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
CPT LoRA	2.6	5.7	10.4	15.5	**17.2**	10.3	72.9	27.8	**62.3**	54.8	66.5	57.3
CPT → DPO	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
CPT → OPD	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
CPT LoRA → QA SFT LoRA	0.6	0.1	1.6	6.3	8.7	3.5	54.8	40.2	45.4	38.7	**83.9**	52.9
CPT LoRA → DPO LoRA	0.5	0.8	1.1	5.8	6.2	2.9	64.1	40.1	42.2	61.4	51.6	52.1
DiSC	0.1	0.1	0.1	2.2	0.3	0.6	10.6	1.9	5.3	9.8	5.0	6.6
System2-FT	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

Table 9: Fine-grained rubric scores on open-ended question subtypes (Significance, Timing, Evaluation, Multi-Answer Vector, Counter-Policy) for fictional and real events. Row color indicates method family: retrieval/RAG methods, reasoning-training methods, CPT-based methods. Bold marks the best value in each column within each model's block.

parameter) fares far worse: training all parameters on the QA-SFT objective collapses model performance entirely, with Qwen3-8B's average accuracy across all four question types falling to 1.2%.

Model	Method	Fictional, Future Events				Real, Historical Events			
		Single	Unanswerable	Open	Multi	Single	Unanswerable	Open	Multi
Qwen3-32B	RAG (full)	68.8	59.2	70.4	46.8	81.1	36.1	84.2	85.9
	DPO on real train (top $k = 1$ article)	70.6	81.3	81.4	47.5	80.2	74.9	88.8	93.1
	DPO on real train	72.8	66.5	82.7	49.6	78.5	54.3	79.4	89.7
Qwen3.5-27B	RAG (full)	76.8	97.8	65.5	46.8	80.9	87.0	86.3	94.1

Table 10: Results for Qwen3-32B and Qwen3.5-27B on fictional and real events. SE is single-event accuracy and Unans. Abst. is the abstention rate on unanswerable questions. Row color indicates method family: base/RAG and reasoning-training methods (DPO). For Qwen3-32B, DPO improves fictional performance over RAG across all question types, and restricting retrieval to the top-ranked article ($k = 1$) yields the largest gains in abstention.

Qwen3-32B and Qwen3.5-27B. Table 10 reports results for Qwen3-32B and Qwen3.5-27B. Under standard RAG with articles, Qwen3.5-27B is a clear step up from Qwen3-32B on fictional events: its single-event accuracy (0.768 vs. 0.688) and, most strikingly, its unanswerable-abstention

score (0.978 vs. 0.592) both improve substantially, while its open-ended score is slightly lower (0.655 vs. 0.704) and multi-event accuracy is unchanged (0.468 for both models). This suggests that Qwen3.5-27B has learned much better calibration on when *not* to answer. On real events, both models are broadly comparable in single-event accuracy (0.809 vs. 0.811), but Qwen3.5-27B again shows markedly better abstention (0.870 vs. 0.361) and better open-ended and multi-event accuracy (0.863 vs. 0.842 and 0.941 vs. 0.859, respectively). Compared to the proprietary models, Qwen3.5-27B (RAG full) outperforms GPT-5.2 and Gemini-2.5 Pro on single-event accuracy (0.768 vs. 0.734/0.640) and unanswerable abstention (0.978 vs. 0.765/0.891), and it outperforms Gemini-2.5 Pro on multi-event accuracy (0.468 vs. 0.297) while matching GPT-5.2 (0.468 vs. 0.468).

We trained Qwen3-32B with DPO on real training data under two conditions: providing a single retrieved article in context (intended to simulate noisy or imperfect retrieval-and-reasoning), and regular DPO with no such retrieval. On fictional events, single-article DPO is not clearly better than plain RAG (full) on single-event accuracy (0.706 vs. 0.688) or multi-event accuracy (0.475 vs. 0.468), though it does improve open-ended generation (0.814 vs. 0.704) and abstention (0.813 vs. 0.592). Regular DPO improves upon the base model on all question types, with a single-event accuracy of 0.728, open-ended score of 0.827, multi-event accuracy of 0.496, and unanswerable abstention of 0.665. On real events, both DPO variants improve substantially over the base model on abstention: single-article DPO more than doubles it (0.749 vs. 0.361), and regular DPO also improves it clearly (0.543 vs. 0.361). Both variants also improve real multi-event accuracy over RAG (0.931 and 0.897 vs. 0.859). Overall, training with a retrieved article in context is costlier, since every training example carries a full article, and the added context appears to introduce noise on fictional events: single-article DPO trails regular DPO on fictional single-event, open-ended, and multi-event accuracy, improving only abstention. On real events, however, it is the stronger variant, outperforming regular DPO on all four question types (0.802 vs. 0.785 single-event, 0.749 vs. 0.543 abstention, 0.888 vs. 0.794 open-ended, 0.931 vs. 0.897 multi-event).

RAG Noise-free Ablation. We conduct an ablation to minimize retrieval noise, pruning the provided context for each single-event question to only the sentences necessary to answer it. Using this minimal, curated evidence rather than full retrieved passages, GPT-5.2 achieves nearly the same accuracy as the full-context setup (74.5% vs. 73.2%), and the vast majority of remaining errors stem from mishandling the relevant facts rather than failing to locate them. Accuracy instead diverges sharply by question type: single-hop and perspectival questions, which require locating a single fact, reach near-ceiling accuracy (97%+), while aggregative (37%) and multihop (61%) questions, which require assembling or chaining multiple facts into one complete answer, remain low. These results indicate that retrieval noise is not the primary cause for errors in RAG. Instead, the dominant failure mode is a reasoning error in which models are unable to synthesize multiple retrieved facts into a single coherent answer.

	Single Event		Unanswerable		Multi Event	
Config	Comb.	Rel.	Comb.	Rel.	Comb.	Rel.
GPT 5.2	70.8	**73.4**	**90.9**	76.5	22.4	**46.8**
GPT 5.2 Agentic	65.0	**74.8**	36.3	**36.8**	56.6	**57.1**
Qwen	60.5	**62.5**	**61.6**	61.4	37.7	**41.9**
Qwen w/ Paragraphs	**60.1**	59.1	53.7	**59.3**	42.6	**47.8**
Qwen Agentic	**44.4**	41.7	**66.7**	64.5	29.6	**31.8**
Qwen DPO	63.9	**65.3**	**77.7**	77.0	37.5	**39.4**
Qwen OPD	62.4	**64.5**	**54.0**	53.8	40.9	**43.8**

Table 11: Fictional-domain performance comparing retrieval over the full **Combined** pool of real and fictional Wikipedia articles versus the **Relevant**-only oracle pool, across Single Event, Unanswerable, and Multi Event questions. Bold marks the better of the two pooling strategies.

Combined Retrieval System. Since 31% of questions require joint reasoning over real and fictional knowledge, we test merging real Wikipedia documents into a single retrieval pool. Tables 11 and 12 show that this combined pool generally hurts performance, especially on fictional events. Fictional single-event accuracy drops modestly (e.g., GPT 5.2 Full K10: 73.4%→70.8%; Qwen Gold RAG: 62.5%→60.5%; Qwen OPD Gold + RAG Full K10: 64.5%→62.4%), while fictional multi-event accuracy suffers far more as distractor real-event articles compete for a fixed retrieval budget

	Single Event			Unanswerable			Multi Event		
Config	Param.	Comb.	Rel.	Param.	Comb.	Rel.	Param.	Comb.	Rel.
GPT 5.2	**83.0**	81.4	80.9	59.1	**82.9**	82.0	81.2	82.7	**83.7**
GPT 5.2 Agentic	83.0	**85.8**	85.1	**59.1**	36.7	35.4	81.2	**90.9**	90.2
Qwen	72.6	78.2	**78.3**	39.7	**69.5**	**69.5**	**71.9**	70.5	71.0
Qwen w/ Paragraphs	72.6	**79.7**	79.4	39.7	64.0	**64.4**	71.9	72.2	**73.0**
Qwen Agentic	**72.6**	72.1	71.8	39.7	65.3	**66.2**	71.9	**76.8**	76.2
Qwen DPO	64.2	66.6	**67.6**	64.4	81.7	**81.9**	76.7	**77.1**	77.0
Qwen OPD	74.1	77.2	**79.1**	34.5	**61.5**	61.1	84.1	85.6	**85.7**

Table 12: Real-event performance comparing the closed-book **Parametric** baseline against retrieval over the full **Combined** pool of fictional and real events and the **Relevant**-only oracle pool, across Single Event, Unanswerable, and Multi Event questions. Bold marks the best of the three per metric.

(GPT 5.2 Full K10: 46.8%→22.4%; Qwen Gold RAG: 41.9%→37.7%), consistent with the change of retriever database in Figure 4. Real-event performance is comparatively insulated and occasionally improves (e.g., GPT 5.2 Agent Full K10: 85.1%→85.8% single-, 90.2%→90.9% multi-event), since real-event questions can fall back on parametric knowledge, making fictional distractors less disruptive than the reverse. On hybrid-knowledge questions (31.2% of the benchmark, 2,340 questions), GPT-5.2 averages 54.1% (relevant) vs. 48.0% (combined), and Qwen averages 32.7% vs. 32.6%. Across all seven configs, combined retrieval yields an average Δ of -1.8pp, indicating the unified pool adds retrieval noise rather than useful cross-domain evidence.

Retrieval on real events generally outperforms parametric-only answering, even for untrained base models, and its benefit is most pronounced on multi-event questions. In Table 2, GPT-5.2 reaches 83.0% real single-event accuracy, Gemini-2.5 Pro 85.6%, and Qwen 3 8B 72.6% closed-book. With real-event articles made into a retrievable database (Table 12), real multi-event accuracy improves over the parametric baseline in six of seven configurations, and the largest gains in the table occur here, driven by agentic systems that can issue separate queries for each event (GPT 5.2 Agentic: 81.2%→90.9%, +9.7pp; Qwen Agentic: 71.9%→76.8%, +4.9pp). Notably, GPT 5.2 gains on multi-event questions (81.2%→83.7%) even though retrieval slightly lowers its single-event accuracy (83.0%→80.9%), suggesting that retrieval matters most when a question requires recalling and linking several events at once. Single-event gains are less uniform: Qwen 3 8B w/ Paragraphs improves to 79.4% (+6.8pp) and GPT 5.2 Agent Full K10 reaches 85.1% (+2.1pp), but GPT 5.2 and Qwen Agentic decline slightly.

H.1 Calibration

We measure calibration using verbalized confidence $c_i \in [0, 1]$ and correctness $y_i \in \{0, 1\}$ (for unanswerable questions, $y_i = 1$ if the model correctly abstains). The **Brier score** (Brier, 1950), $\frac{1}{N}\sum_i (c_i - y_i)^2$, is the mean squared error between confidence and outcome; it heavily penalizes confident errors and rewards confidence that separates correct from incorrect answers. **Adaptive Calibration Error (ACE)** (Nixon et al., 2019) sorts predictions by confidence into $B = 8$ equal-size bins and averages the gap between accuracy and mean confidence, $\frac{1}{B}\sum_b |\mathrm{acc}(b) - \mathrm{conf}(b)|$; equal-size bins avoid empty bins, since verbalized confidences cluster at high values. For both metrics, lower is better.

Figure 7 shows calibration curves for each model. Single-event curves (blue) generally rise with confidence and track the diagonal best for GPT-5.2 on real events. Unanswerable curves (orange) are erratic, showing little link between stated confidence and correct abstention. Single-event questions are generally better calibrated than unanswerable ones, and real single-event questions are better calibrated than fictional ones. Real unanswerable questions are the worst calibrated, consistent with models' overconfidence on familiar events. DPO gives the most consistent improvement, especially on single-event questions, with the gains highest on fictional single-event questions. OPD and CPT-based methods show no benefit, and CPT-based models remain confident but mostly wrong.

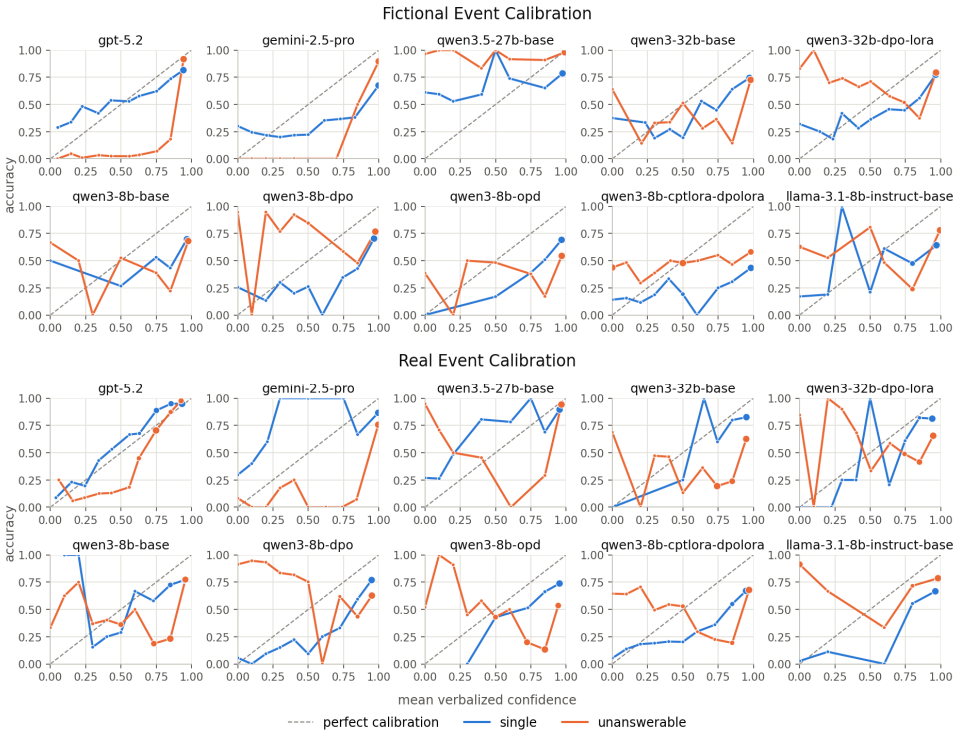


Figure 7: Calibration plots comparing mean verbalized confidence against accuracy for fictional (top) and real (bottom) events. Blue lines show single-event questions and orange lines show unanswerable questions, where accuracy is the rate of correct abstention. The dashed diagonal marks perfect calibration. GPT-5.2 tracks the diagonal most closely, particularly on real single-event questions. Most other models are poorly calibrated, with erratic curves on unanswerable questions that show little relationship between stated confidence and correct abstention.

	Fictional			Real		
Method	Single Event	Unanswerable	Multi Event	Single Event	Unanswerable	Multi Event
			Qwen3-8B			
Qwen3 8B DPO	65.3	77.0	39.4	64.2	64.4	76.7
Qwen3 8B Flipped Preference	4.4	18.0	8.5	0.2	12.3	0.1

Table 13: Preference-flipping ablation on Qwen3-8B. Standard DPO is compared with DPO trained for one epoch on flipped preference pairs, where the reasoning-error answers are preferred over correct answers. Values are accuracy, except for unanswerable questions, which report abstention rate. Rather than producing a fluent but systematically wrong model, flipping the preference collapses performance across all question types on both fictional and real events.

I DPO Preference-Flipping Ablation

To test whether event based training gains from DPO and OPD stem from genuine improvements in event-based reasoning rather than the model simply aligning to the distributional patterns of our benchmark, we construct a control condition in which the chosen and rejected responses in the DPO preference pairs are swapped: the model is trained to prefer the reasoning-error answers over the correct one. Critically, the rejected answers were constructed to reflect specific reasoning failures (hallucinations, factual retrieval errors) rather than incoherent text, so this manipulation isolates preference direction without altering the surface form of the training data.

Training on this flipped objective for a single epoch does not simply invert accuracy in the expected direction; instead, it largely destroys output coherence. Table 13 reports the accuracy of the standard DPO model for each category. Most notably, single-event and multi-event question accuracy collapses for both fictional and real event subsets. Additionally, as the preferred answers in the

DPO set contain a large portion of abstentions for hallucination guardrails, we find that this model hallucinates more and abstains less on the unanswerable question subsets (18.0% fictional, 12.3% real). Overall, optimizing against the correct-answer signal collapses the model's ability to generate valid output altogether. This suggests that standard DPO's gains are not merely an artifact of distributional alignment to benchmark-specific answer patterns, since inverting that same distributional signal does not produce a fluent, systematically-wrong model, but instead breaks the model outright.

J HISTORICAL VS. FUTURE FRAMING

We investigate a design question underlying CASCADE's use of future events: does presenting a fabricated item as set in the past (pre-training-cutoff) versus the future change how the model treats it, independent of item content? We construct a historical version of CASCADE events by systematically shifting all fictional dates back 40 years, covering the decades of the 1990s to the 2010s. We run analysis on event likelihoods with Qwen3-8B.

Gold-answer log-probability is uniformly higher under historical framing. Across four context sizes (186–3,151 tokens), Qwen3-8B assigns higher probability to the correct answer when the surrounding text is framed historically rather than in the future. The direction is consistent and significant at every context size we test, and the delta is largest, not smallest, at the two longest contexts (50 and 100 atomic facts), as Qwen3 8B views the historical events as more plausible because it does not contradict a model's given knowledge cutoff date.

Context	Prompt tokens	Gold logprob Δ	p
Single statement	186	0.0260	1.4×10^{-5}
First paragraph	307	0.0160	1.6×10^{-2}
50 atomic facts	1752	0.0421	1.4×10^{-10}
100 atomic facts	3151	0.0359	5.4×10^{-14}

Table 14: Difference in Qwen3-8B's gold-answer log-probability between historical and future framing (historical − future) across context sizes. Positive values mean the model assigns higher probability to the event text when events are dated in the past, seen at every context size.

	Historical	Future
Mean $P(\text{Yes})$	28.4	3.0
Endorsed as real ($p > 0.5$)	28.0	2.9
Confidently endorsed ($p > 0.9$)	21.8	2.4
Specific checkable fabrications endorsed	22.2	1.5

Table 15: Qwen3-8B endorsement of fabricated events under historical versus future framing, measured on 8,184 dated single-sentence claims. $P(\text{Yes})$ is the probability placed on answering that the event occurred; "No" is correct under both framings. Backdating the same claims increases endorsement roughly tenfold, including confident endorsement ($p > 0.9$) of specific, checkable fabrications.

Historical framing also raises endorsement of outright fabrications. To test this outside the QA setting, we construct a separate dated-fabrication probe: 8,184 single-sentence claims describing our fictional benchmark events that occurred in the past versus in the future, with the item's factual content held fixed. The model is prompted with each dated claim and asked to answer Yes or No to whether the event occurred. "No" is the correct response in both framings, since the event is fabricated regardless of which date is attached to it. We score the model's calibration via $P(\text{Yes})$, the probability mass it places on the single "Yes" token. On this task, backdating sharply raises $P(\text{Yes})$: 25.5% of items (2,085/8,184) flip from correctly rejected to wrongly endorsed purely as a result of the date shift, versus only 668 items moving in the opposite direction. Taken together with the log-probability result above, this is the same underlying effect viewed from two sides: backdating raises the model's blanket plausibility prior for a claim, which both inflates the probability assigned to true statements and erodes rejection of false ones.

When comparing our benchmark questions between historical and future dates, we find that the base models of both GPT-5.2 and Qwen3-8B score higher accuracy and abstain less under historical

framing than under future framing. For single-event questions, GPT-5.2 reaches 76.4% accuracy with 5.5% abstention historically, versus 73.2% accuracy with 6.7% abstention under future framing; Qwen3-8B shows the same direction at a larger magnitude, reaching 68.2% accuracy with 6.6% abstention historically, versus 62.5% accuracy with 8.8% abstention under future framing. In both cases the accuracy gain and the abstention drop move together, indicating that historical framing makes models more willing to commit to an answer at all, which is consistent with the analysis that LLMs view historical fictional events as more plausible than future fictional events.

Ultimately, we construct the benchmark around future rather than historical events for two reasons, both supported by the results above. First, future framing is the harder benchmark: models are trained with an explicit knowledge cutoff, so events dated after it fall outside anything the model could have seen, and a well-calibrated model should treat them as correspondingly less likely by default. Historical framing removes this signal entirely, as a fabrication dated within the training window may be indistinguishable to a model from a true historical fact – consistent with our results, where historical fabrication endorsement rises to 28.0% from just 2.9% under future framing. Second, future framing is safer by construction: fabricated future events cannot yet be checked against any real record and become self-evidently false once the relevant date passes, whereas plausible, detailed historical fabrications are closer in kind to disinformation and, per our results, are exactly the kind of content models are more willing to endorse as fact. Future-event framing therefore gives us a benchmark that is both more diagnostic of real calibration failures and lower-risk to construct.

K Error Taxonomy Analysis

When models abstain on single- or multi-event questions, they may do so for one of two distinct reasons: (1) they judge the question unanswerable given the provided context, or (2) they view the underlying event as wholly fictional and decline to engage with it. We classify both, but count only the first as incorrect: fictionality-based skepticism is a reasonable response to a synthetic dataset, not an error. This unanswerable-abstention rate is 0–0.1% across all methods.

To understand why models fail rather than only how often, we manually categorize every incorrect response into one of five error types: **recall** (a wrong or misremembered fact), **hop** (correct facts identified but incorrectly combined or compared), **partial** (an incomplete or over-inclusive enumeration), **abstain** (declining a question with a valid answer), and **artifact** (formatting or parsing failures unrelated to reasoning). Missing (%) denotes questions the model's response failed to address at all. Tables 16 and 17 report per-reasoning-mode breakdowns for GPT-5.2 and Gemini-2.5-Pro on real and fictional event sets.

Real and fictional errors look qualitatively different. On real events, single-event error is dominated by recall (75% for GPT-5.2, 72% for Gemini), with every other category marginal. On fictional events, recall's share drops to 25–34%, and the remainder splits across abstention, partial, and hop errors. Supplying supporting documents does not simply lower the overall error rate — it spreads errors across several competing failure modes rather than concentrating them in one.

This split can be obscured by looking at error *shares* rather than rates. Hop share falls in every comparable cell moving from real to fictional (e.g., 22%→21%, 23%→16%, 22%→12%, 21%→9%), suggesting hop reasoning becomes relatively less of a problem. But the underlying hop error *rate* (errors per 100 questions asked) rises in each of these same cells; the share falls only because abstention and partial errors grow even faster.

Task shape, rather than difficulty, determines which error type appears. `multihop` on real events is 82–88% recall error rather than hop error, despite involving multi-step inference. Fictional `single_hop` questions have a higher hop-error share (45%) than fictional `multihop` questions (34%). `aggregative` and `cross_event_aggregative`, two of the hardest categories in our benchmark, are 0–3% hop error. What matters is whether a question asks the model to enumerate (partial errors), compute or compare (hop errors), or identify (recall errors).

Temporal reasoning illustrates this recall-versus-hop split within a single category: on real events, `temporal_reasoning` errors are 71–85% recall and 11–29% hop; on fictional events this reverses to 28–32% recall and 46–64% hop. Closed-book, models tend to get the dates themselves wrong; handed the dates in context, they instead get the date arithmetic wrong.

Table 16: Error taxonomy for GPT-5.2 on real events (left) vs. fictional events (right). All values except the number of questions asked are percentages of answer errors. Abbreviations: Number of Questions, **Err**or rate, **Abs**tain, **Mis**sing, **Rec**all, **Hop** error, **Par**tial.

	Real events							Fictional events						
Reasoning mode	# Questions	Err.	Abs.	Mis.	Rec.	Hop.	Par.	# Questions	Err.	Abs.	Mis.	Rec.	Hop.	Par.
Single-event														
All	1983	17	1	0	75	22	1	7507	27	22	6	34	21	17
Single hop	191	6	0	0	100	0	0	909	4	15	0	27	45	12
Multi-source	287	10	0	0	57	43	0	982	13	1	6	38	14	39
Multihop	304	45	2	0	82	15	1	1118	40	11	14	40	34	1
Aggregative	42	26	0	0	91	0	9	884	67	14	3	42	2	38
Causal	571	8	4	0	57	37	2	1392	26	55	1	18	17	9
Temporal	148	51	0	0	71	29	0	948	22	3	0	32	64	0
Comparative	356	7	0	0	92	0	8	774	27	49	10	28	7	7
Perspectival	84	0	0	0	0	0	0	500	1	0	0	29	0	57
Multi-event														
All	2859	19	7	4	43	22	22	4722	53	22	2	34	12	30
Shared entity	500	0	50	0	50	0	0	743	56	0	1	94	1	2
Inferential	500	6	3	22	50	3	22	858	41	73	0	7	10	9
Causal chain	430	32	13	9	16	26	36	602	81	0	3	45	36	15
Aggregative	500	36	5	0	68	3	23	920	89	3	0	20	0	76
Temporal	500	22	0	0	41	52	0	920	38	61	9	12	17	1
Chain multihop	429	3	27	0	13	13	47	679	13	61	5	11	16	7

Gemini's larger real-to-fictional gap on single-event questions is a refusal effect more than a difference in error rate. Its fictional single-event error rate is 36% versus GPT-5.2's 27%, but Gemini abstains on 34% of these questions versus GPT-5.2's 22%, correspondingly lowering its recall and hop shares. This is concentrated in `cross_event_temporal` (86% of Gemini's errors there are abstentions), `causal_reasoning` (66%), and `comparative_reasoning` (58%), suggesting Gemini tends to convert likely-wrong answers into refusals rather than committing to mistakes.

The single-event fictional error increase is also not driven by more wrong facts: recall errors per 100 questions asked stay flat or drop slightly from real to fictional single-event questions (GPT-5.2: 13→9; Gemini: 10→9), so the overall rise in error rate (17%→27% for GPT-5.2; 14%→36% for Gemini) comes almost entirely from abstention, partial, and hop errors. Multi-event questions behave differently: recall errors roughly double to triple (GPT-5.2: 7→18; Gemini: 7→22), indicating that multi-event fictional questions introduce genuine new factual-recall difficulty that single-event questions do not.

One consequence of this taxonomy is that our current grading likely understates model differences on the hardest categories. `cross_event_aggregative` has an 89% (GPT-5.2) to 98% (Gemini) error rate, and roughly three-quarters of those errors are partial — the model recovers some but not all members of a set. Because grading is binary, a response recovering four of five correct members scores identically to one recovering zero, making the hardest category also the least discriminating between models; set-level partial credit would likely separate model capability here without changing question difficulty. The artifact rate stays at or below 2% in every cell (0% for real single-event questions), so grading artifacts do not meaningfully distort this taxonomy.

L Answerable Abstention Rates

Since `ABSTAIN_REFUSAL` is negligible everywhere ($\leq$ 0.24%, and 0% for both proprietary models and nearly all fine-tuned checkpoints), we analyze *answerable* abstention, `ABSTAIN_UNANSWERABLE` (AU): the model declining because it believes it lacks the information, not because of a safety/policy refusal. Tables 18 and 19 report AU% only.

On real events, AU collapses to near-zero ($\leq 1\%$) for almost every method on both backbones. The exceptions are the base models and, more notably, the reasoning-finetuning runs (OPD, OPD+gold, CPT→OPD) on Llama, which sit at 3–14% — the highest real-event AU values in the table, concentrated in multi-event questions. Methods trained to consult retrieved evidence become more willing

Table 17: Error taxonomy for Gemini-2.5-Pro on real events (left) vs. fictional events (right). All values except the number of questions asked are percentages of answer errors. Abbreviations: Number of Questions, **Err**or rate, **Abs**tain, **Mis**sing, **Rec**all, **Hop** error, **Par**tial.

	Real events							Fictional events						
Reasoning mode	# Questions	Err.	Abs.	Mis.	Rec.	Hop.	Par.	# Questions	Err.	Abs.	Mis.	Rec.	Hop.	Par.
Single-event														
All	1983	14	2	0	72	23	3	7507	36	34	4	25	16	21
Single hop	191	6	0	0	92	0	8	909	10	15	0	17	34	31
Multi-source	287	11	9	0	67	21	3	982	30	4	3	21	14	56
Multihop	304	29	0	0	88	11	1	1118	40	33	9	28	27	2
Aggregative	42	17	0	0	100	0	0	884	67	11	3	50	1	35
Causal	571	12	4	0	37	58	1	1392	43	66	0	10	14	10
Temporal	148	36	0	0	85	11	4	948	26	18	0	28	46	8
Comparative	356	6	0	0	80	10	10	774	52	58	6	11	5	19
Perspectival	84	1	0	0	0	0	100	500	3	6	0	35	0	53
Multi-event														
All	2859	22	6	3	38	21	32	4722	70	27	1	31	9	31
Shared entity	500	2	30	0	50	10	10	743	69	4	1	82	6	7
Inferential	500	10	8	6	37	2	47	858	65	58	0	8	6	29
Causal chain	430	45	5	7	11	38	39	602	88	6	2	52	24	14
Aggregative	500	40	2	1	60	4	33	920	98	2	0	23	0	75
Temporal	500	12	7	0	52	40	0	920	50	86	0	8	6	1
Chain multihop	429	6	35	0	19	31	15	679	53	32	5	13	22	27

	Llama-3.1-8B		Qwen3-8B	
Method	Single	Multi	Single	Multi
Base	4.5	8.9	0.0	1.7
CPT	0.0	0.6	0.3	0.6
CPT→DPO	0.0	0.4	0.0	0.7
CPT→OPD	3.0	13.9	0.7	6.8
DPO	0.2	0.1	0.1	1.2
GRPO	0.4	2.8	0.1	1.4
OPD	2.9	7.5	0.1	0.6
DiSC	0.5	0.8	0.2	3.1

Table 18: Answerable abstention (AU) rates (%) on real events for Llama-3.1-8B and Qwen3-8B. AU is the fraction of answerable questions on which the model declines because it believes it lacks the information. AU is near zero for most methods, but it rises for the Llama-3.1-8B base model and OPD-based methods, especially on multi-event questions.

to say "I don't know" about real events they otherwise have parametric knowledge of, apparently because a failed retrieval signal leads the model to defer to that absence rather than fall back on its own knowledge. Qwen shows the same directional pattern (OPD-family and CPT→OPD elevated relative to DPO/GRPO/KL) at much smaller magnitude, consistent with answering more confidently overall. Multi-event AU is also reliably higher than single-event AU for nearly every method on both models, suggesting abstention scales with question difficulty rather than being calibrated purely to the model's real-world knowledge boundary. The proprietary models follow the same pattern, with both GPT-5.2 and Gemini-2.5-Pro abstaining on only 0.1% of real single-event and 1.0% of real multi-event questions.

On fictional events — where no method can have true parametric knowledge — AU is uniformly higher, but the spread is enormous (2.6% to 44.5%) and tracks training recipe cleanly. Raw CPT-family checkpoints not further exposed to RAG or preference data (CPT full, CPT→DPO, CPT→OPD on Llama) are the worst offenders at 30–45% AU, with CPT→DPO the single highest cell in the appendix. Combined with the main-doc finding that these same checkpoints answer real events at ∼0% AU (confidently but often incorrectly), this looks like a genuine calibration failure rather than principled abstention: pushed to answer fictional-distribution content fluently, the model swings to heavy refusal once that fluency runs out, rather than degrading gracefully. By contrast, every method giving the model retrieved or gold evidence (Gold docs, RAG chunk/full, IRCoT,

	Llama-3.1-8B		Qwen3-8B	
Method	Single	Multi	Single	Multi
Gold docs (oracle)	6.7	–	10.2	–
RAG (paragraphs)	11.1	2.8	11.5	1.6
RAG (articles)	6.9	8.4	8.8	7.5
CPT	32.3	19.1	17.2	8.2
CPT→DPO	44.5	26.7	11.6	2.6
CPT→OPD	29.9	33.5	2.6	3.9
DPO	5.2	6.2	8.1	9.3
GRPO	8.8	7.0	5.8	5.6
OPD	4.5	2.6	5.1	5.4
DiSC	13.8	15.5	3.3	1.3

Table 19: Answerable abstention (AU) rates (%) on fictional events for Llama-3.1-8B and Qwen3-8B. CPT-based methods abstain far more often than methods given retrieved or gold evidence, particularly for Llama-3.1-8B. Dashes indicate settings that were not evaluated.

OPD, OPD+gold) lands in a tighter 2.6–14.5% band regardless of backbone — grounding evidence, not preference optimization alone, appears to suppress fictional abstention. The proprietary models also fall within this band: with retrieved or gold evidence, GPT-5.2 abstains on 5.8–7.7% of fictional single-event and 2.0–8.4% of multi-event questions, while Gemini-2.5-Pro abstains more often (11.6–14.5% single-event, 3.6–13.2% multi-event).

Comparing backbones directly, Qwen is consistently lower-AU (more willing to answer) than Llama at matched method on fictional events — e.g., CPT (full) is 32.3%/19.1% for Llama vs. 17.2%/8.2% for Qwen, and CPT→DPO is 44.5%/26.7% vs. 11.6%/2.6% — while on real events the two converge once past the base model. Whether this reflects better calibration in Qwen or simply a lower bar for guessing is the natural follow-up question (checked against accuracy, not shown here), since low AU alone is not evidence of good behavior: a model that never abstains and is frequently wrong scores identically to one that never abstains because it's usually right.

M Chain-of-Thought Exhibition as a Recall Heuristic

As a heuristic probe of whether correct answers on real events are being directly recalled from parametric memory through factual shortcuts versus derived through explicit multi-step reasoning, we measure the rate at which each model's response exhibits visible chain-of-thought (CoT) structure from stepwise intermediate reasoning. If a model has the answer memorized as a retrievable fact, it has little need to reason its way there, so a high CoT-exhibition rate on a given question type is evidence that the answer is not memorized from pre-training data.

Type	GPT-5.2	Gemini-2.5	Union
Single-hop	21.5	25.7	30.9
Multi-source synthesis	56.8	57.5	65.5
Multi-hop	54.3	20.1	61.8
Aggregative	64.7	64.7	82.4
Causal	60.4	72.5	77.4
Temporal	46.6	51.4	64.9
Comparative	68.8	67.7	80.3
Perspectival	10.7	27.4	28.6
Total	**52.8**	**52.4**	**65.4**

Table 20: Rate of visible (%) chain-of-thought (CoT) reasoning in GPT-5.2 and Gemini-2.5-Pro responses to real single-event questions, by question type. Union is the fraction of questions where at least one model shows CoT. Higher rates indicate answers derived through multi-step reasoning rather than recalled directly. Single-hop and perspectival questions show the least CoT, while comparative, causal, and multi-source synthesis questions show the most.

Type	GPT-5.2	Gemini-2.5	Union
Shared entity	86.8	88.2	95.4
Causal chain	77.4	86.3	91.6
Long-chain multi-hop	98.4	95.8	99.3
Cross-event inferential	41.2	53.4	63.2
Cross-event temporal	55.4	49.6	67.2
Cross-event aggregative	25.6	23.0	34.8

Table 21: Rate (%) of visible chain-of-thought (CoT) reasoning in GPT-5.2 and Gemini-2.5-Pro responses to real multi-event questions, by question type. Union is the fraction of questions where at least one model shows CoT. CoT rates are higher than in the single-event setting and nearly universal for long-chain multi-hop questions. Cross-event aggregative questions are the main exception.

Table 20 reports CoT-exhibition rates on the single-event evaluation. Single-hop questions show the lowest rate of any type (21.5–25.7%), consistent with these answers being retrieved directly from memory rather than derived, whereas comparative (67.7–68.8%), causal (60.4–72.5%), and multi-source synthesis (56.8–57.5%) questions all exceed 55% for both models, indicating that the answer is typically being assembled through explicit reasoning rather than recalled outright. The two models diverge sharply on the multi-hop category within this evaluation (54.3% GPT-5.2 vs. 20.1% Gemini-2.5-Pro), the largest single gap in the table, though the union rate of 61.8% shows that between the two models most multi-hop items receive explicit reasoning from at least one.

Table 21 reports the same measurement on the multi-event evaluation, where every subtype requires combining information across events by construction. Rates here are substantially higher across the board than in the single-event evaluation: shared-entity, explicit-causation, causal-chain, and long-chain multi-hop items all exceed 77% CoT exhibition for both models, reaching 98.4% (GPT-5.2) and 99.3% (union) on long-chain multi-hop, indicating that these answers are being derived step-by-step rather than recalled as a unit. Cross-event aggregative is the clear exception, remaining low for both models (23.0–25.6%), which suggests these items are effectively answered as several independent single-hop lookups rather than through multi-step derivation. Across both tables, the union rate exceeds either individual model's rate in almost every category, meaning the two models frequently reason explicitly on different subsets of the same question type. Thus, a nontrivial share of correct answers on both single-event and multi-event questions are derived through explicit reasoning at inference time rather than retrieved as memorized facts, which limits the extent to which reasoning difficulty is confounded with direct factual recall due to factual shortcuts existing in pre-training.

N Why Abstention Is Low on Unanswerable Questions

Abstention on the unanswerable questions in the real event subset is markedly low for the base models of GPT-5.2, Gemini-2.5-Pro, and Qwen3-32B. However, accuracy on single event, multi event, and open ended questions is very high, indicating that there is not a domain-knowledge gap that results in higher hallucinations. We decompose unanswerable questions into three broad reasoning types: rejecting a false premise, in which the named entity, event, or date does not exist as stated (55% of questions); recognizing a knowledge gap, in which the referent is real but the specific fact was never recorded (29% of questions); and refusing a baseless inference between two real entities (16% of questions). Table 22 shows that false-premise rejection is where the models differ most from one another, and that every model does best on spurious-link refusal.

We label a response as containing an epistemic marker if it uses source-checking or uncertainty language (e.g., "no public record," "cannot be determined") prior to a claim. Fewer than 5% of incorrect responses contain such a marker, versus 18–40% of correct ones. Interestingly, for hallucinations, models behave differently, with gpt-5.2 building an uncited, plausible narrative around the fabricated answer; gemini-2.5-pro providing a fabricated citation to 16.7% of its wrong answers; and qwen3-32b reasoning on false premise with conditional language ("would have," "can be inferred") and presenting the result as a given fact. Qwen3 32b also provides a reasoning trace for each question. On false-premise items where the trace questions whether the entity is real, it still abstains only 48.2% of the time, versus 12.9% when the trace never raises the question.

Model	Invalid Premise	Info Gap	Baseless Inference
GPT-5.2	47.0	69.2	81.6
Gemini-2.5	77.4	61.1	82.4
Qwen3-32b	24.3	49.7	51.0

Table 22: Abstention rates of base models on real-event unanswerable questions, grouped by the reasoning needed to abstain correctly. *Invalid Premise* questions name an entity, event, or date that does not exist as stated (55% of questions). *Explicit Info Gap* questions refer to something real but ask for a fact that was never recorded (29%). *Baseless Inference* questions assert a causal connection between two unrelated real entities (16%). Models differ most on rejecting false premises, and all models abstain most reliably on spurious links.

Additionally, 14% of questions request an *exact*, *precise*, or *verbatim* value, and this phrasing is concentrated almost entirely in the epistemic-gap subtypes. It appears in 360/420 (86%) of missing-information questions, where it raises GPT-5.2's abstain rate from 0.350 to 0.867 (+0.517), Qwen3-32B's from 0.150 to 0.581 (+0.431), and Gemini-2.5-Pro's from 0.517 to 0.633 (+0.117); it also appears in 15% of indeterminate questions, raising abstention by 0.294, 0.172, and 0.235 for GPT-5.2, Qwen3-32B, and Gemini-2.5-Pro, respectively. Removing all cue-bearing questions therefore shifts the overall abstention picture (Table 23): GPT-5.2 and Qwen3-32B abstain less often, while Gemini-2.5-Pro abstains slightly more.

Model	Total	Cue-only	Cue-free	Δ
GPT-5.2	59.1	86.5	54.6	+4.5
Gemini-2.5	73.4	66.4	74.5	-1.1
Qwen3-32b	36.1	58.9	32.4	+3.7

Table 23: Effect of precision cues on abstention for real-event unanswerable questions. *Cue-only* covers the 14% of questions that explicitly request an exact, precise, or verbatim value, and *cue-free* covers the remainder. *Total* is the abstention rate over all questions, and Δ is headline minus cue-free. Precision cues sharply raise abstention for GPT-5.2 and Qwen3-32B, so their headline rates slightly overstate how often they abstain without such a cue.

Finally, we check whether abstention failures are systematic across questions or specific to individual models, and find that 83.7% of unanswerable questions are correctly abstained on by at least one model. Thus, the questions themselves are answerable in principle by refusal. Since this evaluation is parametric and no evidence is provided at inference time, this cannot be attributed to a failure to weigh external evidence. Instead, despite being heavily familiar with the underlying real events, models are poorly calibrated when a question rests on a false premise or asks for information that was never recorded: rather than recognizing the gap and abstaining, they default to fabricating a plausible answer.

O DPO VS. OPD COMPARISON

Breaking down single-event fictional accuracy by question type (n=7,507) shows that DPO and OPD reach similar overall correctness gains over the base RAG-10 policy (+2.8pt and +1.7pt, respectively) through very different mechanisms. DPO's gain is concentrated in genuine correctness improvements on multihop (+11.2pt) and multi-source-synthesis (+5.3pt), at the cost of regressions on comparative-reasoning (-5.8pt) and single-hop (-1.2pt). OPD's smaller aggregate gain conceals a more troubling pattern: on aggregative, causal-reasoning, and comparative-reasoning, correctness barely moves or slips slightly (+3.1, +2.5, +0.8pt) despite a large simultaneous rise in incorrect-answer rate (+8.8, +1.1, +5.9pt), indicating OPD is not converting its reduced abstention into correct answers on these categories. Adding the gold-answer KL term (OPD+gold) partially corrects this: it posts the largest multihop gain of any method ($+7.5$pt) and the highest overall correct rate ($+2.0$pt over base, $+2.8$pt over OPD alone), but still shows a comparative-reasoning correctness gain smaller than DPO's multihop gain, suggesting the underlying instability is dampened rather than eliminated.

On multi-event fictional questions, this instability is largely absent for OPD: it improves or holds steady on every subtype (fictional total $+1.9$pt over base, e.g., long-chain multihop $+6.0$pt, cross-

event-inferential $+5.1$pt), and its real-event gains are both larger and more consistent (real total $+14.2$pt, driven by causal-chain $+38.3$pt and cross-event-aggregative $+23.2$pt). DPO's multi-event picture is more mixed: it posts a net fictional decline (-2.5pt total), driven almost entirely by a sharp regression on cross-event-inferential (-9.7pt), even as it matches OPD's real-event gains on causal-chain ($+31.6$pt) and cross-event-aggregative ($+17.0$pt). DPO's advantage on single-event questions came from targeted, largely genuine correctness gains, but on multi-event reasoning – particularly under fictional-to-real domain shift – OPD is the more stable and broadly positive method, despite its weaker calibration on unanswerable questions.

O.1 Abstention Behavior

DPO and OPD are trained on the same 11,430 real_train records, yet for unanswerable questions, DPO's abstention rate increases by +15.6pt, while OPD lowers it by as much as -17.4pt to -26.4pt depending on model. To test whether this divergence simply mirrors teacher behavior, we measured abstention rate at $T = 1$ with 4 rollouts on the unanswerable questions that make up 27.6% of the 11,430-prompt OPD training distribution. Table 24 reports the full distribution of abstentions-per-4-rollouts for both teacher and student in each model family. Qwen3 32B never abstains on 46.8% of questions, and the 8B student is actually *more* abstention-prone than its own teacher at 1.742/4, which may partly explain the degradation in abstentions. However, for Llama-3.1-70B, we find that it abstains on all 4 rollouts for 61.2% of questions, and the 8B student is nearly identical at 61.0% / 3.279. A likelier culprit is the objective itself: OPD's per-token imitation loss reproduces teacher behavior that is not consistent between models, and the KL term toward the gold answer is weighted at only $\lambda = 0.1$. DPO instead weights pairs at the sequence level, so refusals retain their full share of the training signal, and its rejected output (confident hallucination) is stylistically indistinguishable from a correct answer, resulting in gradient updates that downweight confident answers on unanswerable questions.

Model	0/4	1/4	2/4	3/4	4/4	Expectation
Qwen3-8B-base	40.0	11.9	9.8	10.3	28.0	1.742 / 4
Qwen3-32B-base	46.8	12.7	10.6	10.0	19.9	1.435 / 4
Llama-3.1-8B-Instruct	3.8	6.0	9.7	19.4	61.0	3.279 / 4
Llama-3.1-70B-Instruct	3.8	7.3	10.0	17.7	61.2	3.254 / 4

Table 24: Number of abstentions out of 4 sampled rollouts on the unanswerable questions in the OPD training set, for each teacher and student model. Each column gives the percentage of questions on which the model abstained that many times, and the last column gives the expected number of abstentions per question. The Llama-3.1-8B student abstains almost exactly as often as its teacher, while the Qwen3-8B student abstains more often than its teacher. Distilling toward the teacher therefore pushes Qwen3-8B toward less abstention.

Because Llama already abstains frequently in its base behavior, we further investigate its hallucination behaviors. We label a response as containing verify-language if it uses source-checking phrasing before committing to a claim (e.g., "let me check if the sources support this" or explicit qualification of what the retrieved text does or does not establish), as opposed to a flat, confident assertion straight out of the gate. Table 25 shows the behavioral signature of this marker. DPO nearly doubles the base rate of verify-language (34.6%→60.1%), which is most of why its overall hallucination rate rises only mildly (40.2%→45.0%). OPD actively suppresses verify-language below even the untrained base rate (34.6%→25.7%), and its overall hallucination rate climbs much further, to 63.8%. Verify-language presence cuts the hallucination rate by roughly 2–3$\times$ in every model, so the divergence between methods can also be seen as a downweighting of verify-language that emerges specifically when relying on teacher outputs through distillation.

Beyond emitted abstention rate, we probe whether each training objective installs a genuine learned preference over refusal versus fabrication. We calculate the within-model contrast $\log p(\text{refusal}) - \log p(\text{hallucination})$, scored on 1500 held-out fictional pairs never seen in training; these candidate continuations are generated by GPT-4.1, so there is no contamination from any student or teacher model. We run this probe on OPD both with and without the λ-weighted gold-answer KL term, to separate what the gold term itself contributes from the base OPD objective. On Qwen, the log weighting moves from -0.175 to -0.139 under OPD ($+0.036$) and to -0.114 under OPD+gold

	Verify Lang. %	Halluc. % (present)	Halluc. % (absent)	Overall Halluc. %
LLaMA-3.1 Base	34.6	15.3	53.4	40.2
LLaMA-3.1 DPO	60.1	33.3	62.7	45.0
LLaMA-3.1 OPD	25.7	32.5	74.7	63.8

Table 25: Verify-language usage and hallucination rates (%) for Llama-3.1-8B under each training method. Verify-language is source-checking phrasing that appears before the model commits to a claim. The present and absent columns give hallucination rates for responses with and without this phrasing. Responses with verify-language hallucinate far less often in every model. DPO increases verify-language use, while OPD reduces it below the base rate.

($+0.061$), versus -0.175 to $+4.063$ under DPO ($+4.238$); on Llama, it moves from $+0.162$ to $+0.116$ under OPD (-0.046) and to $+0.089$ under OPD+gold (-0.073), versus $+0.162$ to $+2.167$ under DPO ($+2.005$). DPO's weighting shift to favor abstentions is 30–100$\times$ larger than either OPD variant's in both models, while OPD raises log-probability on *both* refusal and hallucination continuations by similar amounts, indicating an undifferentiated teacher-matching objective that does not upweight refusals on unanswerable questions. This confirms that DPO results in a learned preference for refusal over fabrication compared to OPD.

Model	Method	Base	After	Shift	Fictional abstention Δ
Qwen3-8B	DPO	-0.175	$+4.063$	$+4.238$	$+15.6$
	OPD+gold answer	-0.175	-0.114	$+0.061$	-7.6
	OPD only	-0.175	-0.139	$+0.036$	-13.6
Llama-3.1-8B	DPO	$+0.162$	$+2.167$	$+2.005$	-4.8
	OPD+gold answer	$+0.162$	$+0.089$	-0.073	-23.6
	OPD only	$+0.162$	$+0.116$	-0.046	-17.4

Table 26: Learned preference for refusal over fabrication, measured as $\log p(\text{refusal}) - \log p(\text{hallucination})$ on 1,500 held-out fictional unanswerable questions with GPT-4.1-generated candidate continuations. Positive values favor refusal. The table reports this preference before (Base) and after training, the resulting shift, and the change in fictional abstention rate (percentage points). DPO shifts the preference toward refusal by orders of magnitude more than OPD.

As a final ablation, we remove all unanswerable (refusal-labeled) examples from DPO training for Qwen3 8B, isolating the effect of that 27.6% slice on the resulting policy. Abstention rate on unanswerable questions falls sharply as a result – from 77.00% to 50.68% on fictional events and from 64.40% to 44.61% on real events, confirming that DPO's abstention gain is driven directly by exposure to refusal preference pairs during training.