Conflicting Needles in a Haystack: How LLMs Behave When Faced with Contradictory Information

Murathan Kurfalı
RISE Research Institutes of Sweden
Sweden
murathan.kurfali@ri.se

Robert Östling
Stockholm University
Sweden
robert@ling.su.se

Abstract

Large Language Models (LLMs) have demonstrated an impressive ability to retrieve and summarize complex information, but their reliability in conflicting contexts remains poorly understood. We introduce an adversarial extension of the Needle-in-a-Haystack framework in which three mutually exclusive "needles" are embedded within long documents. By systematically manipulating factors such as position, repetition, layout, and domain relevance, we evaluate how LLMs handle contradictions. We find that models almost always fail to signal uncertainty and instead confidently select a single answer, exhibiting strong and consistent biases toward repetition, recency, and particular surface forms. We further analyze whether these patterns persist across model families and sizes, and we evaluate both probability-based and generation-based retrieval. Our framework highlights critical limitations in the robustness of current LLMs—including commercial systems—to contradiction. These limitations reveal potential shortcomings in RAG systems' ability to handle noisy or manipulated inputs and exposes risks for deployment in high-stakes applications.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in processing and retrieving information from long documents (Tay et al., 2020; Xu et al., 2023; Liu et al., 2024). The ability to effectively process information within extended contexts has been crucial for a wide range of applications, e.g. question answering, and document summarization, often via retrieval-augmented generation (RAG) (Lewis et al., 2020; Gao et al., 2023).

However, a critical challenge arises when LLMs are faced with inconsistencies or contradictions within their input context (Chen et al., 2024). Real-world documents often contain conflicting information due to updates, errors, differing perspectives, or adversarial attacks. The ability of an LLM to identify and resolve such conflicts, or at least to indicate uncertainty, is vital for its reliability and trustworthiness. As shown in Figure 1, even a mid-sized model like Qwen2.5-32B outputs opposing reasoning strategies depending on how its input is structured.

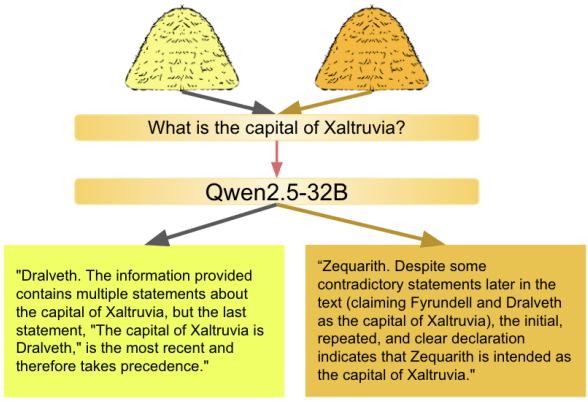

Figure 1: Actual examples from our experiments illustrating how Qwen2.5-32B employs contradictory reasoning strategies. The haystacks represent the same document containing conflicting information presented in different configurations (e.g. see Figure 2). Qwen favors recency in one instance and repetition in the other. Note that this example is atypical in that most LLMs we tried do not identify the conflict at all.

In the current paper, we investigate how LLMs behave when presented with explicit contradictory knowledge in their context. We introduce an evaluation framework that extends the classic Needle-in-a-Haystack (NIAH) test (Kamradt, 2023) where a single relevant statement ("needle") is hidden in a long distractor text ("haystack"), to a more adversarial setting with multiple conflicting needles. Specifically, three mutually exclusive candidate statements are embedded in haystacks constructed with different configurations. We set out to answer the following research questions:

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 34350–34365
November 4-9, 2025

Domain	Question	Candidate Needles
Geography	What is the capital of Xaltruvia?	
History	Which treaty resolved the Xaltruvian conflict?	*Dralveth*, *Fyrundell*, *Zequarith*
Biology	In which genus is Xaltruvia classified?	

Table 1: Domains and the questions used in our experiments. Each question is paired with three mutually exclusive pseudo-answers (needles).

- RQ1: To what extent are current LLMs able to identify inconsistent information in an information retrieval setting?
- RQ2: When models fail to identify inconsistent information, which factors influence their choice of which information to rely on?
- RQ3: If we model one LLM's behavior, to what extent does it predict the behaviors of other LLMs—within and across families/sizes?

By systematically manipulating factors such as the position and frequency of conflicting statements, the layout and the domain of the document, we aim to provide a comprehensive analysis of how LLMs prioritize and select information when confronted with contradictions. Our contributions are threefold: (i) we introduce a novel contradictory "multi-needle" evaluation framework for stress-testing LLMs with conflicting in-context information, (ii) we quantify the factors that affect which of the multiple conflicting pieces of information the LLMs pick, and (iii) we analyze the extent to which selection behavior generalizes across models, highlighting both shared patterns and important model-specific biases.

Our findings reveal that most current LLMs, including smaller commercial models, tend to confidently resolve contradictions by arbitrarily favoring one alternative, with systematic biases towards position, repetition, and surface-form features, while often failing to indicate uncertainty. These results point to fundamental limitations in current LLMs' abilities to reason under conflicting information—a critical consideration for reliable deployment in real-world applications.

2 Literature Review

Early benchmarks like the Long Range Arena (Tay et al., 2020) and NIAH (Kamradt, 2023) tested retrieval of a single known item from lengthy distractors, revealing that model performance drops when relevant information is not located at the input's boundaries. The "Lost in the Middle" study by Liu et al. (2024) showed that many models exhibit a distinctive U-shaped accuracy curve when the relevant passage's position is varied—performance is highest when information appears at the beginning (primacy effect) or end (recency effect), and lowest in the middle.

More recent benchmarks like LongBench (Bai et al., 2024) and L-Eval (An et al., 2024) systematically evaluated long-context understanding across diverse tasks, confirming that model accuracy consistently degrades as input length increases.

NoCha (Karpinska et al., 2024) introduced a challenging claim verification benchmark over book-length narratives, where each false claim minimally contradicts a true one. Results reveal that even the commercial models like GPT-4 struggle to maintain consistency under contradiction. Pham et al. (2024) proposed WhoQA, a benchmark that tests LLMs under entity-based ambiguities, showing that models often produce confident but incorrect answers when facing conflicting entity references. Similarly, Neeman et al. (2023) introduced DisentQA, a framework that evaluates models' ability to disentangle parametric (internal) knowledge from contextual (retrieved) information using a counterfactual QA setup, enhancing robustness to knowledge conflicts.

3 Framework

We construct *haystacks* from Wikipedia articles, with *needles* consisting of statements inserted at different locations. Figure 2 shows an example. In all cases, we provide three contradictory needles giving the answer to the question as **Fyrundell**, **Zequarith** and **Dralveth** (Table 1).[1]

We would like to note that, unlike the other NIAH-based evaluation frameworks, there is no "correct" answer in our setup. Instead, we are inter-

[1]Pseudo-words that are confirmed not to have any hits on Google at the time of experiments.

Code	Hypothesis
H1	Repetition increases selection likelihood.
H2	Position bias—models favor earlier/later needles.
H3	Needle identity affects selection likelihood.
H4	Needle identity effects vary by domain.
H5	Layout modulates repetition effects.
H6	Layout modulates position effects.
H7	Position and repetition interact.
H8	Semantic relatedness modulates repetition effects.

Table 2: Summary of hypotheses tested in the study.

ested in (a) whether the LLM identifies that there is no single answer to the question, or (b) if not, how does the LLM select between the different options?

3.1 Haystack Construction

Unlike the earlier extensions of NIAH, we use substantially more variables to systematically assess how different factors influence the probability assigned by the model to each of the possible answers. We define two sets of experimental factors—needle-level variables that vary across the three inserted needles within a document, and haystack-level moderators that define the configurations of the haystacks.

Needle-level variables When inserting needles, we manipulate the following variables:

- **Repetition count**: The number of times a given needle appears in the haystack. The values in our experiments are 1, 2 or 5 times.
- **Position index**: The order in which a particular needle appears in the document (as number 1, 2 or 3). Note that repeated needles are placed in sequence, so there is always a well-defined order between needles.
- **Needle identity**: The form of the needle. One of *Fyrundell*, *Zequarith* and *Dralveth*.

Haystack-level moderators These variables are constant for all three needles in a document:

- **Layout strategy**: The spatial arrangement of needle insertions, defined by a combination of two factors: i) **position** which refers to the general region of the haystack where needles are inserted: at the *beginning*, *middle*, or *end* of the document; ii) **grouping** which defines how close the needles are: in *sequential*, they appear consecutively; in *5%*, each is separated by 5% of the document length.

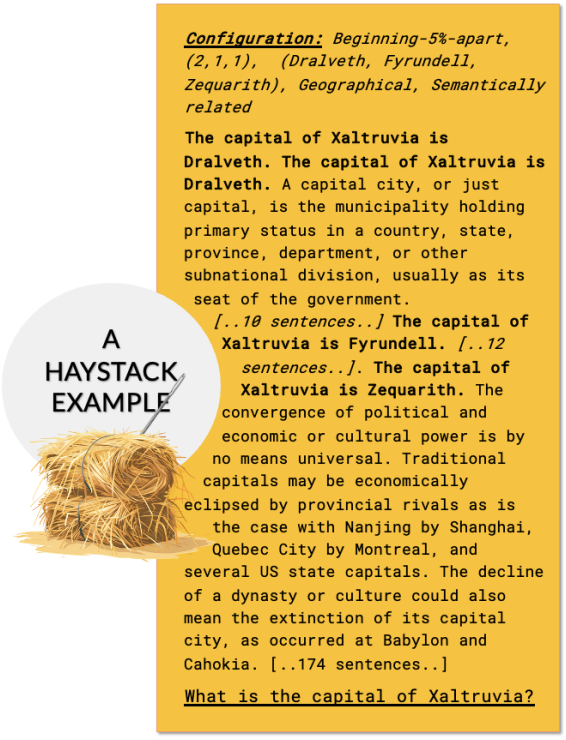

Figure 2: A sample haystack

Additionally, the **even** layout distributes needles at 25%, 50%, and 75% of the document, serving as a neutral baseline for comparing clustered versus spread-out configurations.

- **Question type**: The domain of the question associated with the needle set. Each question belongs to one of three categories: *Geography*, *History*, or *Biology*.
- **Semantic relatedness**: A binary variable indicating whether the haystack domain matches the needle's question domain. See Appendix A for the list of Wikipedia articles used.

In order to fit within the context windows of all models we study, we limit the haystack size to 5,000 words (amounting to about 7,000 tokens) [2].

3.2 Hypotheses

Using the needle- and haystack-level variables defined above, we test the following hypotheses about LLM behavior in conflicting-information scenarios, summarized in Table 2. These hypotheses are designed to systematically probe the influence of information presentation (repetition, position, layout)

[2]We truncated the Wikipedia excerpts at the nearest sentence boundary to preserve fluency.

Variable	Values (reference value)	Interpretation
H1: Repetition	2x, 5x (ref = 1x)	How the likelihood of selection changes when a needle is repeated more than once.
H2: Position	2nd, 3rd (ref = 1st)	How the likelihood of selection changes depending on a needle's position in the document.
H3: Needle ID	Fyrundell, Zequarith (ref = Dralveth)	How the likelihood of selection affects across different fabricated needle names.
H4: Needle × Domain	All needle × question type combinations (ref = Dralveth × Geography)	How the effect of needle identity changes under different question domains.
H5: Repetition × Layout	All repetition × layout combinations (ref= 1x × beginning-sequential)	How the effect of repetition changes under different layouts.
H6: Position × Layout	All position × layout combinations (ref= 1st × beginning-sequential)	How the effect of position changes under different layouts.
H7: Position × Repetition	All position × repetition combinations (ref= 1x × 1st)	How the effect of repetition changes depending on a needle's position.
H8: Semantic × Repetition	All repetition × semantic closeness combinations (ref = 1x × unrelated)	How the effect of repetition changes under different semantic relatedness conditions.

Table 3: Summary of variables, reference levels, and interpretations in the conditional logistic regression models.

and content characteristics (identity, domain, relatedness) on model choice in conflicting-information scenarios.

3.3 Evaluation Approaches

Since our framework sits at the intersection of long-context retrieval and multiple-choice selection – with three mutually exclusive candidate answers, none of which is objectively correct – we leverage this structure to evaluate model behavior using both probability-based and generation-based approaches.

- **Probability-based approach** Our primary evaluation method is based on the log-probabilities that the model assigns to each candidate needle when prompted with a question. Specifically, for each haystack, we compute the likelihood of each needle being the answer and normalize across the three options:

$$p(x|c) = p_{LLM}(x|c) / \sum_{x' \in X} p_{LLM}(x'|c)$$

  given prompt c and the needle set X. In this approach, the needle with the highest conditional probability is considered as the model's selected answer.

- **Generation-based approach** In addition to probability-based modeling, we also conduct generation-based experiments. For each haystack, we prompt the model with the corresponding question and record its generated output (via greedy decoding). We then label the generated answer as: (i) *single* if exactly one needle is mentioned in the answer; (ii) *mixed* if more than one needle is mentioned as plausible responses; and (iii) *refused* if the model declines to answer, expresses uncertainty, or indicates that the information is unavailable.

There is yet to be a consensus on the best method for evaluating LLM understanding (Lyu et al., 2024; Wang et al., 2024a,b). Generation-based evaluation has the advantage of closely mirroring real-world user scenarios; hence, it is practically appealing. However, it has been repeatedly shown that language models can produce different answers depending on whether they are prompted to provide the answer through "introspection," or whether the answer is computed directly from token probabilities during generation (Song et al., 2025).

Probability-based evaluation, in contrast, exposes the model's internal knowledge and preferences (Hu and Levy, 2023). Moreover, since there is no objectively correct needle in our experimental setup, computing the probability distribution over all needles allows for comparisons of model behaviors under different configurations. For these reasons, our main results are based on the probability-based selection of needles. However, we conducted the generation-based experiments and reported the results in the Appendix E. We further discuss the differences and implications of these evaluation approaches in detail in Section 6.

4 Experimental Setting

We quantify the effect of each factor using conditional logistic regression, applied separately to each

Hyp.	Variable	Qwen 2.5-3B	Qwen 2.5-7B	Qwen 2.5-32B	Gemma 2-9b	Gemma 2-27b	Llama 3.2-3B	Llama 3.1-8B	Llama 3.3-70B
H1	2x	17.86**	5.04**	10.48**	14.83**	3.33**	5.79**	26.42**	3.63**
	5x	21.19**	5.64**	8.00**	25.89**	7.66**	12.29**	63.05**	2.34**
H2	2nd	3.63**	0.78	7.86**	0.83	2.20**	4.22**	1.45	0.51**
	3rd	0.77	1.02	6.23**	0.12**	1.45	0.97	0.12**	0.22**
H3	Fyrundell	0.42**	0.79	0.41**	0.94	1.48**	0.62**	0.42**	2.20**
	Zequarith	1.70**	0.57**	0.36**	0.96	0.99	0.24**	0.08**	1.05
H4	Fyrundell × HIST	5.46**	0.58**	0.22**	0.20**	0.42**	1.15	1.16	1.11
	Fyrundell × BIO	0.38**	0.24**	0.23**	1.37	0.71	0.73	5.67**	1.88**
	Zequarith × HIST	0.70*	0.22**	2.67**	0.15**	0.36**	0.46**	2.77**	2.15**
	Zequarith × BIO	0.11**	0.41**	4.25**	1.76**	0.95	0.17**	1.51	2.28**
H5	2x × Clust@45-55	0.19**	0.67	0.26**	1.00	1.13	0.93	1.41	0.94
	2x × Clust@50	0.23**	0.68	1.08	0.97	1.05	0.58	1.44	0.44*
	2x × End-Seq	0.34**	1.02	1.29	1.19	1.30	2.30*	2.31*	1.17
	2x × Even	0.34**	1.44	0.33**	1.80	0.71	0.87	0.61	0.66
	2x × beg-5%-apart	0.46*	1.18	0.48	0.91	1.22	0.95	0.95	0.89
	2x × end-5%-apart	0.19**	0.97	0.41*	1.44	0.91	1.17	0.94	0.80
	5x × Clust@45-55	0.19**	1.14	0.46*	0.51	0.66	0.64	0.75	1.74
	5x × Clust@50	0.33**	0.90	2.06	1.59	1.11	1.13	8.80**	0.83
	5x × End-Seq	1.18	1.18	3.42**	6.57**	0.77	0.92	1.15	1.76
	5x × Even	0.42*	1.89	0.59	1.20	0.43*	0.54	0.31**	0.67
	5x × beg-5%-apart	0.67	1.40	0.79	0.47	0.57	0.70	0.89	0.54
	5x × end-5%-apart	0.34**	0.75	0.91	0.75	0.18*	0.57	0.34*	0.76
H6	2nd × Clust@45-55	1.65	1.53	0.91	3.04**	6.15**	0.56	0.36**	0.95
	2nd × Clust@50	0.67	0.88	1.25	1.05	0.62	1.08	0.70	0.70
	2nd × End-Seq	1.93	3.12**	0.36**	0.37**	0.20**	1.12	0.27**	0.34**
	2nd × Even	1.17	1.23	0.43*	7.59**	6.53**	1.21	1.55	0.81
	2nd × beg-5%-apart	1.16	2.79**	0.93	1.57	1.42	2.26*	3.72**	0.35**
	2nd × end-5%-apart	0.51*	1.79*	0.30**	0.59	0.23*	1.25	0.45*	0.19**
	3rd × Clust@45-55	6.96**	0.55*	1.82	12.53**	1.18	5.81**	5.70**	0.32**
	3rd × Clust@50	2.72**	1.02	2.41*	2.07	0.27**	4.56**	2.07	1.14
	3rd × End-Seq	31.88**	2.20**	0.72	6.19**	4.30**	15.86**	12.73**	1.21
	3rd × Even	7.74**	1.49	0.56	67.08**	3.76**	7.55**	48.10**	1.73*
	3rd × beg-5%-apart	0.77	2.32**	1.24	1.20	4.36**	6.42**	18.58**	0.52*
	3rd × end-5%-apart	13.01**	2.62**	1.66	64.92**	28.04**	79.24**	114.27**	6.04**
H7	2nd × 2x	0.33**	1.20	0.94	0.37**	1.54	0.53*	0.20**	0.71
	2nd × 5x	0.36**	1.32	1.22	0.63	0.76	0.77	0.38**	1.10
	3rd × 2x	0.81	0.92	2.92**	0.77	2.19*	1.34	0.32**	1.16
	3rd × 5x	0.76	0.99	2.07*	1.38	1.68	1.00	0.61	1.79*
H8	2x × Related	0.64*	0.93	0.54**	0.33**	0.79	1.21	0.67*	1.88**
	5x × Related	0.81	1.51*	0.53**	0.32**	0.88	1.18	0.45**	2.12**

Table 4: Odds ratios from conditional logistic regression for probability-based needle selection.

model's outputs. Following the grouped-data setup in `statsmodels`, each haystack forms a group with three data points—one per needle—labeled to indicate the selected candidate. Coefficients are exponentiated and reported as odds ratios: values above 1 indicate increased selection likelihood relative to the reference; values below 1 indicate decreased likelihood. Statistical significance follows standard thresholds (* $p < 0.05$, ** $p < 0.01$).

To assess model fit, we compared conditional logistic regression models with only main effects to full models with the interaction terms, using Akaike (AIC) and Bayesian (BIC) Information Criteria. In all cases, the full models showed substantially lower AIC and BIC scores (see Appendix C), supporting the inclusion of interactions and indicating improved fit without overfitting. The variables and their interpretations of the logistic regression model is provided in Table 3.

We evaluate eight open-source and four commercial OpenAI models that are listed in Appendix B. All experiments with the open-source models were run on two H100 GPUs using a local cluster and the total computational budget was approximately 120 GPU hours. On the other hand, the OpenAI experiments cost around $3. In the probability-based approach, we compute the log probability of the exact token sequence for each needle. For generation-based evaluation, we use the same prompt and generate answers via greedy decoding, with annotations as described in Section 3.3. The prompts' configuration was as follows: "<haystack> <question>

Answer:". Due to API limitations, the commercial models were evaluated using only the generation-based approach.

Generalization Furthermore, to verify the robustness of our findings , we conducted an additional set of experiments on three other needle configurations, e.g. different needle sets including real-world entities, different domains and questions, evaluated on a subset of the models. Due to space constraints, we present a summary of these results in our main discussion, with detailed results provided in Appendix F.

5 Results and Discussion

5.1 Evaluation of the Hypotheses

The Impact of Repetition (H1, H5, H7, H8) Repetition reliably raises the chance that a needle is selected, confirming **H1**. Even with conflicts in the context, models lean toward what they see more often: in the probability-based results, the 5× condition is strongest for *Llama-3.1-8B* (OR=63.05) and remains large for *Gemma-2-9b* (25.89) and *Qwen 2.5-3B* (21.19), but is more modest for *Qwen2.5-7B* (5.64) and *Llama-3.3-70B* (2.34), indicating that this bias does not grow monotonically with size (e.g., *Qwen2.5-32B* = 8.00 vs. *Qwen 2.5-3B* = 21.19). Layout modulates this tendency to various degrees (**H5**): end-oriented clustering can amplify it (e.g., 5××End-Seq in *Qwen2.5-32B* and *Gemma-2-9b*), whereas neutral "Even" placements often dampen it. Position interacts with repetition as well (**H7**): repeating a needle in the second slot typically lowers its odds, while the third slot can either help or hurt depending on the model. Finally, semantic relatedness (**H8**) usually narrows the repetition advantage (e.g., *Gemma-2-9b*, *Qwen2.5-32B*), though a few models show the opposite pattern (notably *Llama-3.3-70B* and *Qwen2.5-7B*). Taken together, repetition has a strong effect, yet not uniform, that is shaped by where and how needles appear, and by how closely they match the question's topic.

The Impact of Position (H2, H6, H7) Position matters, but not in a single direction across models, partially supporting **H2**. Some models clearly favor later placements (e.g., *Qwen2.5-32B* prefers the 2nd and 3rd positions), while others penalize the last item (e.g., *Gemma-2-9b*, *Llama-3.1-8B*, *Llama-3.3-70B*). These patterns do not suggest a general "recency" effect. If anything, the second position is often advantaged over the first for several models. Layout helps explain when recency emerges (**H6**): when the last needle truly sits at the end (e.g., end-sequential or end–5%-apart layouts), third-position odds can spike dramatically; by contrast, the same position can be suppressed in other layouts (e.g., second×End-Seq). As noted above, repetition and position interact (**H7**), with repetition at the second slot frequently weakened and the third slot showing model-specific boosts or drops. In short, position exerts consistent influence, but its direction is conditional on layout and its interplay with repetition.

The Impact of Needle Identity (H3 and H4) Surface form also shapes selection, confirming **H3**. The fabricated names are not neutral: for instance, "Fyrundell" is down-weighted in several models (*Qwen 2.5-3B*, *Qwen2.5-32B*, *Llama-3.2-3B*, *Llama-3.1-8B*), while "Zequarith" is especially disfavored in the Llama family, yet preferred in *Qwen 2.5-3B*. Domain further modulates these identity effects (**H4**): the same surface form can flip from penalty to boost depending on whether the question is historical, biological, or geographic (e.g., History reduces "Zequarith" in *Gemma-2-9b* but raises it in *Qwen2.5-32B*). Together, these results suggest that beyond repetition and position, models carry stable—but model-specific—preferences for particular strings, and that those preferences interact with topic.

We further illustrate these findings by plotting needle selection rates across different domains for each model in Figure 4. The observed dependence on surface form, together with domain-specific modulation, may present challenges in specialized fields where terminology is complex and highly precise. In such contexts, LLMs may overlook critical information simply due to a bias against less familiar or uncommon terms.

5.2 Performance of the Commercial Models

We further evaluate four commercial models from OpenAI: GPT-4o, GPT-4.1, GPT-4o-mini, and GPT-4.1-nano. Because the API does not give access to token-level probabilities, these models are evaluated using the generation-based approach only (same prompt as open-source models; greedy decoding) and we, again, fit conditional logistic regression to the resulting selections.

In a preliminary experiment with the first 50 haystacks, the larger models (GPT-4o and GPT-4.1) consistently detected the presence of contradictory

Model	Consist.	Single	Mixed	Refused
Qwen 2.5-3B	53.1%	94%	5.4%	0.6%
Qwen2.5-7B	54.8%	76.4%	14.6%	9.0%
Qwen2.5-32B	53.3%	64.6%	22.2%	13.2%
Llama-3.2-3B	54.8%	69.6%	24.6%	5.8%
Llama-3.1-8B	85.7%	76.9%	22.2%	0.9%
Llama-3.3-70B	68.1%	69.0%	30.7%	0.3%
Gemma-2-9b	84.1%	98.2%	0.7%	1.1%
Gemma-2-27b	89.0%	92.5%	7.3%	0.2%
GPT-4o-mini	-	88.3%	11.7%	0%
GPT-4.1-nano	-	97.0%	3.0%	0%
Overall	67.9%	82.6%	14.2%	3.1%

Table 5: Comparison of probability and generation-based predictions, including response types in the latter approach. Consistency reflects cases where both methods selected the same needle; mixed outputs count as consistent if the first-mentioned needle matches.

information and typically refused to commit to a single answer. Given this consistent behavior and budget limits, we did not experiment with these models further.

On the other hand, the smaller variants (GPT-4o-mini and GPT-4.1-nano) rarely signaled conflict and predominantly selected a single needle. Their odds ratios (Table 6) show a strong frequency effect (H1; e.g., GPT-4o-mini: 2× = 35.38^{**}, 5× = 28.39^{**}; GPT-4.1-nano: 5× = 9.96^{**}), model-specific position patterns rather than a uniform recency effect (H2; GPT-4.1-nano penalizes later positions, while GPT-4o-mini penalizes 2nd but boosts 3rd), and substantial modulation by layout (H5), often in opposite directions across the two models (e.g., 2××Clust@50: 6.65^{**} vs. 0.10^{**}). Relatedness can weaken repetition for GPT-4.1-nano (H8; 5××Related = 0.53^{**}), with weaker trends for GPT-4o-mini.

5.3 Inter-Model Analysis

To assess similarity in needle selection across models, we compute the pairwise symmetrized Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) between their predicted answer distributions for each haystack. Each model assigns a probability distribution over the three candidate needles per document, reflecting its response to the contradictory input. This allows us to examine how consistently different models exhibit biases. Let $\mathbf{p} = [p_1, p_2, p_3]$ and $\mathbf{q} = [q_1, q_2, q_3]$ denote the probability vectors for two models on the same document.

Hyp.	Variable	GPT-4.1-nano	GPT-4o-mini
H1	2x	1.51	35.38**
	5x	9.96**	28.39**
H2	2nd	0.03**	0.07**
	3rd	0.31**	3.43**
H3	Fyrundell	0.79	1.03
	Zequarith	0.89	0.71*
H4	Fyrundell × HIST	0.88	1.53*
	Fyrundell × BIO	0.82	1.08
	Zequarith × HIST	0.65*	1.32
	Zequarith × BIO	0.98	1.66*
H5	2x × Clust@45-55	6.73**	0.31
	2x × Clust@50	6.65**	0.10**
	2x × End-Seq	6.38**	0.31
	2x × Even	3.15**	0.11**
	2x × beg-5%-apart	2.88**	0.17**
	2x × end-5%-apart	3.44**	0.09**
	5x × Clust@45-55	2.18	0.42
	5x × Clust@50	2.44	0.19**
	5x × End-Seq	1.35	0.30
	5x × Even	1.27	0.12**
	5x × beg-5%-apart	0.72	0.42
	5x × end-5%-apart	1.88	0.13**
H6	2nd × Clust@45-55	0.69	2.09
	2nd × Clust@50	8.17**	1.74
	2nd × End-Seq	2.71*	1.90
	2nd × Even	2.25	0.56
	2nd × beg-5%-apart	6.28**	0.63
	2nd × end-5%-apart	9.17**	6.49**
	3rd × Clust@45-55	0.13**	0.61
	3rd × Clust@50	0.87	0.21**
	3rd × End-Seq	1.05	0.11**
	3rd × Even	0.49*	0.01**
	3rd × beg-5%-apart	5.20**	0.01**
	3rd × end-5%-apart	1.11	0.66
H7	2nd × 2x	2.93**	2.96**
	2nd × 5x	8.42**	3.88**
	3rd × 2x	1.80	4.02**
	3rd × 5x	2.53*	8.87**
H8	2x × Related	1.12	0.96
	5x × Related	0.53**	1.24

Table 6: Odds ratios from conditional logistic regression for generation-based needle selection for OpenAI models.

The (asymmetric) KL divergence is defined as:

$$D_{\mathrm{KL}}(\mathbf{p}\|\mathbf{q}) = \sum_{i=1}^{3} p_i \log \frac{p_i}{q_i}$$

We use the **symmetrized** KL divergence to measure the dissimilarity between models (though it's important to note that KL divergence is not a true metric as it doesn't satisfy the triangle inequality):

$$D_{\mathrm{SKL}}(\mathbf{p}, \mathbf{q}) = D_{\mathrm{KL}}(\mathbf{p}\|\mathbf{q}) + D_{\mathrm{KL}}(\mathbf{q}\|\mathbf{p})$$

We compute D_{SKL} for each model–haystack pair and report the mean across configurations as a summary of pairwise model similarity (Figure 3) where lower values indicate more similar selection behavior; higher values reflect greater differences.

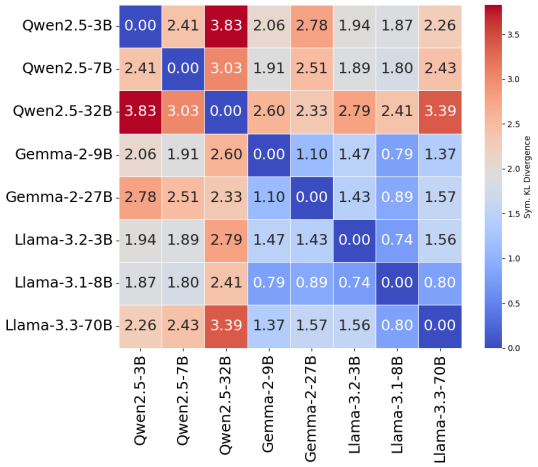

Figure 3: Inter-model similarity via symmetrized KL across all configurations; lower values indicate higher similarity.

The analysis of model similarity reveals distinct patterns related to both model family and size. We present two focused analyses: intra-family agreement and agreement among similar-sized models. The model families include Llama (3 models), Gemma (2 models), and Qwen2.5 (3 models), and the model sizes are grouped into tiny (2B and 3B - 2 models), small (7B, 8B and 9B - 3 models), and medium (27-32B - 2 models).

Within-family agreement: Model behavior demonstrates a notable dependence on the LLM family. Llama models show the highest consistency in predicted distributions (mean SKL = 1.03), with Gemma models also exhibiting relatively strong agreement (1.10). In contrast, Qwen2.5 models display greater internal variation (3.09), indicating that these models display considerably different tendencies in their needle selection.

Agreement among similar-sized models: Examining cross-family agreement by model size, we observe the greatest similarity among the small-sized models (7B–9B; mean SKL = 1.72). Tiny models (2B–3B) show moderate agreement (1.94), while medium-sized models (27B–32B) exhibit the least similarity (2.56). Notably, Gemma-2-9B and Llama-3.1-8B present a high degree of similarity (0.79), despite belonging to different families.

Generalization of the results: To further validate our results, we ran a subset of models on three additional setups: (C2) a second synthetic triplet on the same haystacks, (C3) the economy and sports domains, and (C4) a real-world case with known entities (France, Sweden, Italy). Details and results for these configurations appear in Appendix F. Findings are consistent across configurations. Repetition (H1) consistently boosts selection, while Position (H2) effects are stable within model and are modulated by layout and repetition (H6–H7). Surface form (H3) continues to influence preferences, with interactions by domain (H4). Full results are provided in Table 13.

6 Comparing generation and probability-based approaches

So far, we have focused on probability-based needle selection via normalized likelihood. We now turn to a generation-based approach, which reflects model behavior in open-ended text generation.

6.1 Consistency between the approaches

We compared the selected needles in both approaches. Table 5 summarizes the relationship between probability-based and generation-based evaluation outcomes, focusing on two main aspects: the type of generative response (single, mixed, or refused) and the consistency between the model's most probable selection and its generated answer.

The first pattern we observe is that the overwhelming majority of model generations produce a single response. Single-needle outputs account for 98.2% in Gemma-2-9b, and 92.5% even in a mid-sized model Gemma-2-27b, with 80% on average. Mixed responses are much less frequent, ranging from 1% to 30% depending on the model, while refusals are extremely rare (less than 1% for half of the models). Surprisingly, the two commercial models also follow this suit and produced no refusals, with 92.6% single responses on average. This is concerning, as it suggests that even when presented with explicit conflicting information, these models rarely signal any ambiguity or uncertainty in their output. Instead, they overwhelmingly opt for a confident, singular answer, potentially giving a false impression of reliability and obscuring the underlying contradictions within the input (see Figure 1 for an actual example).

On the other hand, the agreement between the probability-based and generation-based outputs is relatively low, averaging 67.9%. This finding supports the previous observation by (Song et al., 2025) that LLMs may exhibit different behaviors when evaluated through distinct methodological approaches. To better understand if the internal

biases observed in the probability-based approach remain, we conducted the same logistic regression analysis on the generated outputs.

6.2 Logistic regression results for generation-based responses

In this logistic regression analysis, we used the instances where the generated text mentioned only one needle. The results are presented in Table 12 (and Table 6 for the commercial models).[3]

Overall, the generation-based results largely mirror the findings from the probability-based analysis. Key trends observed in the probability-based modeling are consistently reflected in the generated outputs, suggesting that the LLMs' underlying biases are expressed similarly in both their probabilistic tendencies and their explicit text generation.

Specifically, in both settings, repetition yields a large and statistically significant effect on selection likelihood across models, confirming a strong repetition bias under open-ended generation. In the probability-based results, 5x repetition is substantial for Llama-3.1-8B (OR=63.05**), Gemma-2-9b (OR=25.89**), and Qwen 2.5-3B (OR=21.19**). In generation-based modeling (single-needle), magnitudes differ by configuration: 5x repetition reaches OR=181.012** for Gemma-2-9b and OR=138.498** for Llama-3.3-70B, with similarly large effects for Qwen 2.5-3B (OR=92.454**) and Llama-3.1-8B (OR=89.37**). Position effects vary across approaches: the probability model shows later-position advantages for some models (e.g., Qwen-2.5-32B: 2nd OR=7.86**; 3rd OR=6.23**), while the generation-based results highlight strong boosts for later positions in Qwen 2.5-3B (2nd OR=10.85**; 3rd OR=24.754**) and Gemma-2-9b (2nd OR=13.353**; 3rd OR=51.359**), with others penalizing later slots (e.g., Llama-3.2-3B: 3rd OR=0.049*; Qwen-2.5-32B: 2nd OR=0.168**). Surface-form preferences show parallel tendencies: "Fyrundell" and "Zequarith" are down-weighted in Llama-3.1-8B (OR=0.391** and OR=0.107**, respectively), and "Zequarith" is reduced in Llama-3.2-3B (OR=0.457**), while patterns for other families vary by model and domain.

While the direction of effects generally agrees across methods, the magnitudes can diverge—sometimes larger under generation (as in certain repetition–layout combinations for Llama-3.1-8B), sometimes comparable or smaller (as with Gemma-2-9b). These differences likely reflect configuration-dependent dynamics in generation and the restriction to single-needle mentions in this analysis. Despite such magnitude shifts, the consistent presence and direction of repetition, position, and identity effects across both approaches reinforce the robustness of these biases in LLMs.

[3]We also tested a setting that considers the first-mentioned needle in mixed responses. Results were similar, with consistent main effects.

7 Conclusion

We introduced a framework for evaluating how LLMs handle conflicting information in long contexts by extending the NIAH setup with mutually exclusive "needles." Across four configurations (synthetic/real entities; multiple domains), models rarely express uncertainty and instead commit to a single answer; selection is strongly driven by repetition, while position effects are model-specific and shaped by layout, and surface form further shifts preferences.

This suggests that LLMs still lack mechanisms for recognizing and signaling ambiguity—an important limitation for RAG and other long-context applications. Future work should explore training objectives or prompting strategies that better equip LLMs to detect and express uncertainty.

Limitations

Our study has several limitations. First, the open-source models we evaluated are only up to 70B parameters, all in English. This might not capture the full range of behaviors seen in larger models or those trained on different languages. Second, our haystacks are built from a fixed set of Wikipedia articles and use fabricated statements to simulate contradictions. While this design allows for controlled comparisons, a more detailed analysis, for instance, one with more levels of repetition or varied types of contradictory information, would offer a more comprehensive view. Broader evaluation, however, was limited by resource constraints, as the number of configurations expands drastically with each addition. Finally, this paper serves as a diagnostic tool and does not attempt to provide solutions for the observed biases.

Acknowledgements

We would like to thank ***Alena Línková*** for her work on an earlier version of this idea, in which she compiled the questions and needles in the initial configuration. This research is partially funded by

the Swedish Research Council through grant agreement no. 2024-01506, and by the Swedish national research infrastructure Språkbanken, jointly financially supported by the Swedish Research Council (2018–2028; grants 2017-00626 and 2023-00161) and the 10 participating partner institutions. The experiments with the open-source LLMs were partially enabled by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) (project: 2025/22-601).

References

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2024. L-eval: Instituting standardized evaluation for long context language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14388–14411.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, and 1 others. 2024. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

Roger Fletcher. 2000. *Practical methods of optimization*. John Wiley & Sons.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2:1.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060.

Gregory Kamradt. 2023. Needle in a haystack - pressure testing llms. https://github.com/gkamradt/needle-in-a-haystack. GitHub repository.

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. One thousand and one pairs: A "novel" challenge for long-context language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17048–17085.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

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

Chenyang Lyu, Minghao Wu, and Alham Aji. 2024. Beyond probabilities: Unveiling the misalignment in evaluating large language models. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 109–131.

Meta. 2024. Llama 3.3 Model Card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md. Accessed: February 2025.

Ella Neeman, Roee Aharoni, Or Honnovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. Disentqa: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Annual Meeting of the Association for Computational Linguistics*.

Quang Pham, Hoang Ngo, Luu Anh Tuan, and Dat Quoc Nguyen. 2024. Who's who: Large language models meet knowledge conflicts in practice. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10142–10151.

Siyuan Song, Jennifer Hu, and Kyle Mahowald. 2025. Language models fail to introspect about their knowledge of language. *arXiv e-prints*, pages arXiv–2503.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2020. Long range arena: A benchmark for efficient transformers. *arXiv preprint arXiv:2011.04006*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving Open Language Models at a Practical Size. *Preprint*, arXiv:2408.00118.

Xinpeng Wang, Chengzhi Hu, Bolei Ma, Paul Rottger, and Barbara Plank. 2024a. Look at the text: Instruction-tuned language models are more robust multiple choice selectors than you think. In *First Conference on Language Modeling*.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024b. "my answer is c": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7407–7416.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Retrieval meets long context large language models. In *The Twelfth International Conference on Learning Representations*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

A Haystack texts

For constructing the haystack documents in the matching condition, we used Wikipedia articles from three relevant domains: capital cities (Geography), peace treaties (History), and botanical taxonomy (Biology). For the mismatching condition, we selected articles from unrelated topics: Formula 1, rock music, and economic crises. Table 7 lists the specific Wikipedia articles and their URLs used in each domain. The selected articles are among the longest Wikipedia articles and exceed the haystack length used in our experiments. Wikipedia text is available under the Creative Commons Attribution–ShareAlike license (CC BY-SA 4.0).

Domain	Condition	Wikipedia Article
Geography	Related	Capital city
History	Related	Treaty
Biology	Related	Species
Geography	Unrelated	2012 Formula One
History	Unrelated	Rock music
Biology	Unrelated	Euro area crisis

Table 7: Wikipedia articles used to construct haystack texts for each domain under matching and mismatching conditions.

B Models

Table 8 lists the baseline models used in our paper, along with their corresponding repository names on Hugging Face's model hub[4]. Overall, we evaluated three LLM families Qwen2.5 (Yang et al., 2025), Gemma-2 (Team et al., 2024), and Llama 3.* (Meta, 2024).

Model Name	HuggingFace Repository
Gemma-2-9B	google/Gemma-2-9b-it
Gemma-2-27B	google/Gemma-2-27b-it
Qwen 2.5-3B	Qwen/Qwen2.5-3B
Qwen-2.5-7B	Qwen/Qwen2.5-7B
Qwen-2.5-32B	Qwen/Qwen2.5-32B
Llama-3.2-3B	meta-llama/Llama-3.2-3B-Instruct
Llama-3.1-8B	meta-llama/Llama-3.1-8B-Instruct
Llama-3.3-70B	meta-llama/Llama-3.3-70B-Instruct
GPT-4o-mini	gpt-4o-mini-2025-04-14
GPT-4.1-nano	gpt-4.1-nano-2025-04-14

Table 8: HuggingFace/OpenAI identifiers of the models used in our evaluation.

[4] https://huggingface.co/models

Model	Probability-based				Generation-based			
	AIC_{main}	BIC_{main}	AIC_{full}	BIC_{full}	AIC_{main}	BIC_{main}	AIC_{full}	BIC_{full}
Gemma-2-9B	2974.54	3007.40	1935.65	2154.67	2784.06	2816.80	1799.14	2017.42
Gemma-2-27B	3342.27	3375.12	2385.39	2604.40	2870.04	2902.43	1888.33	2104.23
Qwen 2.5-3B	3393.24	3426.09	2130.51	2349.52	1981.81	2014.29	1597.56	1814.10
Qwen-2.5-7B	2736.05	2768.90	2367.87	2586.89	1756.78	1788.02	1534.07	1742.32
Qwen-2.5-32B	2831.70	2864.55	1920.50	2139.52	1328.05	1358.28	1131.39	1332.94
Llama-3.2-3B	2690.79	2723.65	1979.98	2198.99	1571.17	1601.85	1288.31	1492.84
Llama-3.1-8B	2543.53	2576.38	1998.75	2217.77	1743.61	1774.89	1424.57	1633.06
Llama-3.3-70B	3159.92	3192.77	2630.01	2849.03	1857.30	1887.92	1213.54	1417.71
GPT-4.1-nano	-	-	-	-	2159.29	2192.15	1896.10	2115.11
GPT-4o-mini	-	-	-	-	2371.17	2404.02	1897.97	2116.98

Table 9: Model selection criteria (AIC/BIC) for main effects and full conditional logistic regression models, shown separately for probability-based and generation-based evaluations. Lower values indicate better model fit.

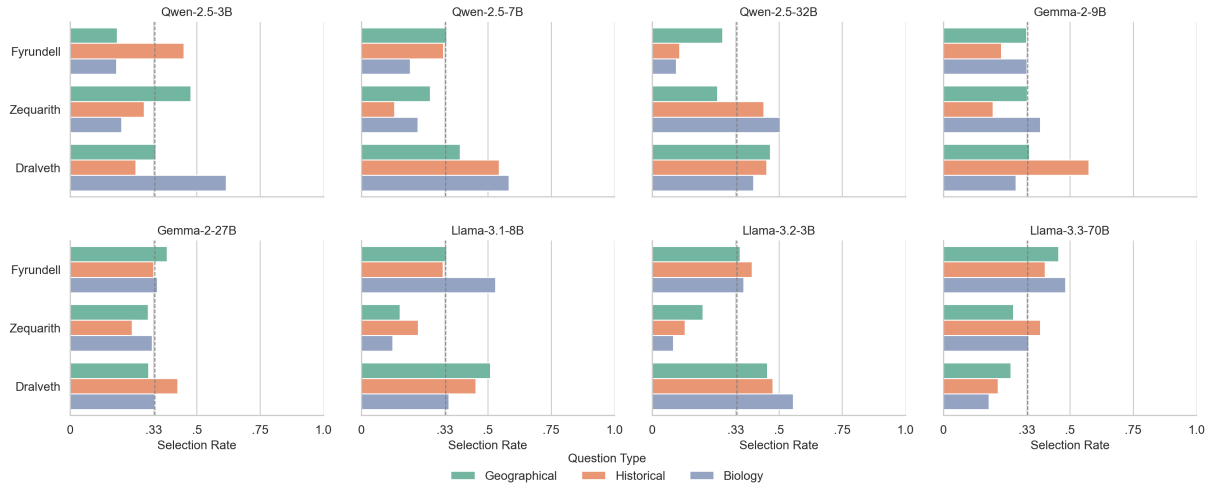

Figure 4: Needle selection rates by question type across models. The vertical line indicates the expected selection rate under a uniform distribution (33%).

The Wikipedia articles are processed using the WIKIPEDIAAPI[5] and Spacy's (Honnibal and Montani, 2017) EN_CORE_WEB_SM is used as the sentence-tokenizer to determine the sentence boundaries.

C Details of the Conditional Logistic Regression experiments

Model fitting was performed using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm (Fletcher, 2000), with a maximum of 2000 iterations to ensure convergence.

To evaluate model fit and justify model complexity, we compared conditional logistic regression models including only main effects to full models including all main effects and interaction terms. Model fit was assessed using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), calculated as:

$$\text{AIC} = 2k - 2\log L$$

$$\text{BIC} = k\ln(n) - 2\log L$$

where k is the number of model parameters, L is the maximized likelihood, and n is the number of choice sets.

As can be seen in Table 9, including interaction terms resulted in substantially lower AIC and BIC values, indicating that the added complexity of the full model improves fit to the data and is statistically justified. These results support the inclusion of interaction terms in our primary analyses.

D Descriptive Plots

To complement the statistical analysis presented in Section 5, here we present complementary descriptive plots of the selection behavior observed in the LLMs when the probability-based approach was employed. These visualizations aim to provide a

[5] https://pypi.org/project/Wikipedia-API/

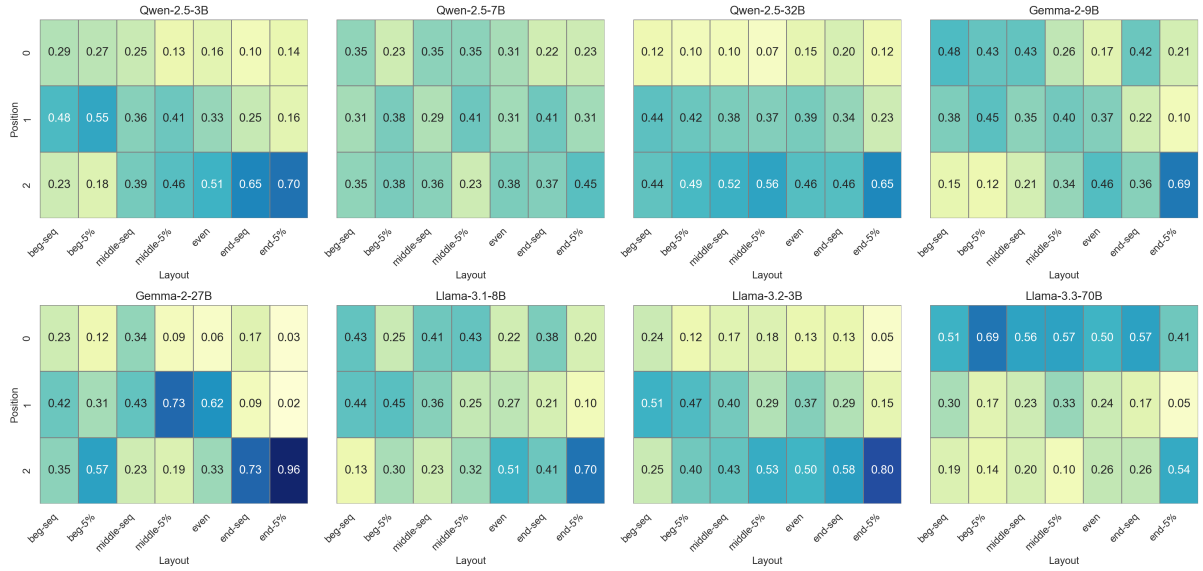

Figure 5: Heatmaps show the selection likelihood of each needle position (1st, 2nd, 3rd) across different layout strategies, for each evaluated model.

more intuitive understanding of how LLMs handle contradictory information and exhibit biases under various conditions.

Figure 4 illustrates the observed needle selection rates for different question types across the various language models. As clearly can be seen in the figure, the selection rates are non-uniform, indicating a clear preference towards certain needles in certain domains. This potentially highlights a concerning bias towards specific surface forms of information, which could lead to undesired behaviors, especially in specialized domains.

On the other hand, Figure 5 presents a series of heatmaps that visualize the selection likelihood of needles based on their specific position (1st, 2nd, or 3rd), modulated by different layouts. Different models react distinctly to the various layouts. For example, Qwen2.5-7B exhibits minimal positional bias, with selection likelihoods for each position remaining relatively uniform across layouts. Conversely, there are extreme cases, such as Gemma-2-27B-Instruct, which almost exclusively selects the last needle in the end-5% layout (96% likelihood for the last position). Other models also show strong preferences in this layout, though less prominent; for instance, Llama-3.2-3B selects the last needle in 80.2% of the time, and Llama 3.1 8B does so 69.7% of the time. This highlights significant model-specific behaviors and the impact of layout on positional biases.

E Statistical results based on generation experiments

This appendix presents detailed statistical results from our generation-based experiments. These analyses mirror those in the main paper (Section 6) but are based on model outputs generated via greedy decoding. Specifically, we focus on *single* responses—cases where the respective LLM mentioned only one needle in its generated text. The number of examples used per model in this condition is listed in Table 10.

Table 12 reports odds ratios from conditional logistic regression for each hypothesis. The results largely confirm the trends observed in the probability-based setting. In particular, repetition and recency effects (H1 and H2) are strong and consistent across models. Notably, some models exhibit extremely large odds ratios (e.g., for 5x repetition in the Gemma and Llama families), indicating even a more pronounced bias toward repeated content.

F Extended Configurations

Beyond the haystack configuration discussed in the main text, we evaluate three additional configurations that vary entity sets and domains to validate the robustness of the results:

- **C2** uses the synthetic needles *Fenvirox*, *Vrenzalik*, and *Qorandel* with the same three domains as C1: *Geography* (*What is the capi-*

Model	Data Size (N)
Gemma-2-9B	1732
Gemma-2-27B	1632
Qwen 2.5-3B	1658
Qwen-2.5-7B	1348
Qwen-2.5-32B	1140
Llama-3.2-3B	1228
Llama-3.1-8B	1356
Llama-3.3-70B	1217
GPT-4.1-nano	1703
GPT-4o-mini	1557

Table 10: Number of examples used per model for conditional logistic regression in the generation setting.

Domain	Condition	Wikipedia Article
Geography	Related	Capital city
History	Related	Treaty
Biology	Related	Species
Economy	Related	Currency
Sports	Related	Association football
Eurovision	Related	Eurovision Song Contest
Geography	Unrelated	2012 F1 Championship
History	Unrelated	Rock music
Biology	Unrelated	Euro area crisis
Economy	Unrelated	Guitar
Sports	Unrelated	The Avengers (2012 film)
Eurovision	Unrelated	Soybean

Table 11: Wikipedia articles used to construct haystack texts for the extended configurations (C2–C4) under matching and mismatching conditions.

tal of Xaltruvia?), *History* (*Which treaty resolved the Xaltruvian conflict?*), and *Biology/Scientific* (*In which genus is Xaltruvia classified?*)

- **C3** reuses the C2 needles in two new domains: *Economy* (*What is the official currency of Ortazea?*) and *Sports* (*Which team won the 2025 Ortazean Premier League?*)
- **C4** employs real entities—*France*, *Sweden*, and *Italy*—within a *Eurovision* prompt (*Who won the Eurovision Song Contest 2025?*).

Representative haystack sources for related and unrelated conditions used across C2–C4 are listed in Table 11.

Hyp.	Variable	Qwen 2.5-3B	Qwen 2.5-7B	Qwen 2.5-32B	Gemma 2-9b	Gemma 2-27b	Llama 3.2-3B	Llama 3.1-8B	Llama 3.3-70B
H1	2x	24.351**	19.785**	30.246**	7.19**	8.921**	57.71**	24.502**	69.007**
	5x	92.454**	18.892**	5.875**	181.012**	9.626**	62.62**	89.37**	138.498**
H2	2nd	10.85**	1.783	0.168**	13.353**	1.737*	1.738	2.008	5.643*
	3rd	24.754**	1.641	1.856	51.359**	2.072**	0.049*	1.247	2.446
H3	Fyrundell	0.567**	2.019**	1.3	0.522**	0.998	0.89	0.391**	0.59**
	Zequarith	0.854	1.009	0.69	0.556**	0.776	0.457**	0.107**	0.74
H4	Fyrundell × HIST	2.303**	0.332**	0.711	1.188	1.639*	4.487**	4.431**	2.609**
	Fyrundell × BIO	1.765*	0.351**	0.62	2.294**	1.699*	2.517**	5.825**	1.002
	Zequarith × HIST	5.762**	0.598*	1.379	1.024	0.527**	13.092**	3.9**	1.847*
	Zequarith × BIO	5.442**	0.415**	1.186	2.782**	1.383	12.219**	1.116	0.642
H5	2x × middle-5%	0.459	1.192	0.624	0.877	1.347	0.201	0.844	0.136
	2x × middle-seq	0.316*	0.946	2.237	2.416	6.577**	0.133	2.427	0.2
	2x × end-seq	1.341	3.231*	1.042	31.458**	2.413*	0.109	1.109	0.075**
	2x × Even	0.142**	1.168	0.119**	2.617*	0.519*	0.127	0.61	0.039**
	2x × beg-5%	1.781	0.615	0.442	1.188	0.888	0.096*	0.386	12.693
	2x × end-5%	0.355*	0.428	0.105**	1.002	2.374	0.103*	0.632	0.027**
	5x × middle-5%	0.178	1.764	5.528**	0.508	5.452**	0.001	1.658	0.264
	5x × middle-seq	0.055**	2.851*	4.378*	1.278	20.937**	0.001	5.337*	0.991
	5x × end-seq	82.482	2.371	6.758**	2675.696	15.007**	0.001	1.609	0.136
	5x × Even	0.02**	4.277**	1.358	0.29*	0.226**	0.0*	0.412	0.023**
	5x × beg-5%	0.417	1.86	1.393	0.258*	0.378**	0.0	0.297*	8.599
	5x × end-5%	0.018**	1.882	2.111	0.112**	0.199*	0.0*	0.322*	0.013**
H6	2nd × middle-5%	0.277*	0.512	1.339	0.326*	2.492**	1.933	2.996	0.769
	2nd × middle-seq	0.251*	0.341*	1.35	0.488	11.473**	2.617	5.533*	0.591
	2nd × end-seq	0.083**	0.099**	6.787**	0.33	0.808	0.492	1.38	0.51
	2nd × Even	0.101**	1.695	1.514	0.409	6.031**	0.288*	0.693	0.017**
	2nd × beg-5%	0.188**	1.597	21.29**	0.045**	0.899	0.731	0.607	0.0
	2nd × end-5%	0.016**	0.052**	1.271	0.003**	0.001**	0.389	0.057**	0.01**
	3rd × middle-5%	0.219*	0.433	0.385	0.126**	0.549	7.838	6.916**	0.791
	3rd × middle-seq	0.093**	0.725	0.274*	0.136**	0.754	29.529**	3.421	2.735
	3rd × end-seq	0.035**	0.201**	4.38*	2.652	1.095	1.961	0.861	2.567
	3rd × Even	0.014**	1.392	1.549	0.129**	3.358**	77.176**	2.173	0.074**
	3rd × beg-5%	0.004**	6.292**	13.635**	0.001**	1.403	16.758*	0.967	0.0
	3rd × end-5%	0.014**	0.628	0.576	0.003**	0.013**	19.654*	0.129**	0.412
H7	2nd × 2x	0.927	0.597	3.003*	0.718	0.582	1.149	0.314**	1.258
	2nd × 5x	1.15	0.401*	3.541**	0.229**	1.021	1.497	0.253**	2.24
	3rd × 2x	1.286	0.454*	0.449	1.265	0.332**	0.92	0.395*	2.053
	3rd × 5x	1.288	0.688	0.207**	0.493	0.639	0.738	0.539	1.275
H8	2x × Related	0.969	0.725	1.806*	0.458**	0.953	0.86	1.059	1.063
	5x × Related	0.485*	0.607*	2.845**	0.353**	1.137	1.085	0.543*	1.597

Table 12: Odds ratios from conditional logistic regression based on generated outputs with single needle mentions. Higher values indicate increased likelihood of being selected relative to the reference condition.

Table 13: Odds Ratios of the selected models across four configurations

		Llama 3.2-3B				Gemma 2-9b				Qwen 32B				GPT-4.1-nano			
Hyp.	Variable	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
H1	2x	5.79**	13.36**	7.83**	13.10**	14.83**	11.15**	2.20	12.55**	10.48**	9.54**	9.82**	18.56**	8.57**	3.02**	3.18**	5.65**
	5x	12.29**	27.55**	21.98**	56.09**	25.89**	21.19**	4.67**	23.10**	8.00**	10.54**	12.51**	45.52**	10.85**	9.08**	7.65**	9.72**
H2	2nd	4.22**	8.00**	3.62**	4.90**	0.83	1.05	0.12**	0.91	7.86**	3.53**	5.48**	68.78**	0.53**	0.67*	0.73*	1.54
	3rd	0.97	1.33	1.46	0.92	0.12**	0.19**	0.08**	0.14**	6.23**	3.82**	4.93**	37.35**	0.36**	0.44**	0.23**	1.18
H3	Fyrundell	0.62**	-	-	-	0.94	-	-	-	0.41**	-	-	-	1.68**	-	-	-
	Zequarith	0.24**	-	-	-	0.96	-	-	-	0.36**	-	-	-	0.94	-	-	-
	Qorandel	-	0.12**	0.07**	-	-	0.92	2.33**	-	-	2.54**	2.81**	-	-	0.79	1.89**	-
	Vrenzalik	-	0.86	0.32**	-	-	0.83	3.18**	-	-	2.19**	2.37**	-	-	1.78**	2.50**	-
	Italy	-	-	-	0.20**	-	-	-	1.15	-	-	-	0.69*	-	-	-	0.76
	Sweden	-	-	-	1.27	-	-	-	0.98	-	-	-	0.75	-	-	-	1.68**
H4	Fyrundell × historical	1.15	-	-	-	0.20**	-	-	-	0.22**	-	-	-	3.03**	-	-	-
	Fyrundell × biology	0.73	-	-	-	1.37	-	-	-	0.23**	-	-	-	0.93	-	-	-
	Zequarith × historical	0.46**	-	-	-	0.15**	-	-	-	2.67**	-	-	-	2.25**	-	-	-
	Zequarith × biology	0.17**	-	-	-	1.76**	-	-	-	4.25**	-	-	-	1.21	-	-	-
	Qorandel × historical	-	3.59**	-	-	-	5.17**	-	-	-	1.09	-	-	-	1.10	-	-
	Qorandel × biology	-	5.96**	-	-	-	2.08**	-	-	-	0.76	-	-	-	1.08	-	-
	Vrenzalik × historical	-	1.62**	-	-	-	5.27**	-	-	-	2.83**	-	-	-	0.79	-	-
	Vrenzalik × biology	-	5.39**	-	-	-	2.14**	-	-	-	0.81	-	-	-	0.86	-	-
	Qorandel × sports	-	-	1.58	-	-	-	0.56**	-	-	-	0.95	-	-	-	0.39**	-
	Vrenzalik × sports	-	-	2.78**	-	-	-	0.34**	-	-	-	0.61*	-	-	-	0.66*	-
H5	2x × Clust@45-55	0.93	0.42*	0.42	0.47	1.00	1.11	1.48	1.18	0.26**	0.29**	0.45*	7.30	0.59	0.74	1.03	2.29
	2x × Clust@50	0.58	0.47	0.46	0.65	0.97	1.03	2.60	1.55	1.08	1.35	1.15	3.02	0.18**	0.41**	0.41*	2.88
	2x × End-Seq	2.30*	1.37	0.72	0.99	1.19	1.77	6.21**	3.35**	1.29	0.92	1.88	1.51	1.56	3.01**	0.92	4.63*
	2x × Even	0.87	0.31**	0.56	0.46	1.80	2.26*	4.18**	2.75*	0.33**	0.37**	0.52*	1.54	0.61	0.82	2.22	0.97
	2x × beg-5%-apart	0.95	0.36*	0.39*	0.27*	0.91	1.38	0.88	1.05	0.48	0.58	0.78	0.62	0.44*	0.59	1.95	1.14
	2x × end-5%-apart	1.17	0.49	0.70	0.76	1.44	1.28	4.56**	2.76*	0.41*	0.42*	0.65	1.67	2.35*	2.02*	1.99	0.90
	5x × Clust@45-55	0.64	0.37*	0.31*	2.40	0.51	0.95	1.36	0.73	0.46*	0.48	0.61	1.97	0.51	0.59	0.92	6.45
	5x × Clust@50	1.13	0.54	1.98	1.17	1.59	1.48	5.88**	2.98*	2.06	2.35	2.11*	0.47	0.30**	0.78	0.55	3.25
	5x × End-Seq	0.92	1.12	0.69	3.94	6.57**	7.19**	16.72**	10.50**	3.42**	2.66*	3.15**	*1.28	1.49	1.98	0.79	1.79
	5x × Even	0.54	0.23**	0.25**	1.24	1.20	1.19	2.35	1.58	0.59	0.58	0.71	0.44	0.83	0.75	2.04	0.74
	5x × beg-5%-apart	0.70	0.41*	0.45	0.10**	0.47	0.78	0.60	0.61	0.79	0.77	0.88	0.13*	0.44*	0.47*	1.27	0.21*
	5x × end-5%-apart	0.57	0.26*	0.36	1.16	0.75	1.00	3.92*	1.88	0.91	0.53	0.79	1.38	0.87	1.58	4.79*	0.30
H6	2nd × Clust@45-55	0.56	0.36**	0.46	0.02**	3.04**	3.36**	6.90**	4.50**	0.91	1.20	1.05	0.04*	1.00	0.68	0.20**	0.51
	2nd × Clust@50	1.08	0.56	1.02	0.44	1.05	1.17	1.07	1.11	1.25	2.17	1.55	0.06*	0.46**	0.34**	1.22	0.62**
	2nd × End-Seq	1.12	0.57	1.61	0.13**	0.37**	0.36**	0.63	0.45*	0.36**	0.52	0.48*	0.02**	0.78	1.02	0.73	0.46
	2nd × Even	1.21	0.46*	2.25	0.07**	7.59**	25.23**	38.98**	27.25**	0.43*	0.65	0.81	0.36	0.72	1.47	0.42*	0.36
	2nd × beg-5%-apart	2.26*	1.97	1.76	0.49	1.57	1.67*	5.58**	2.95**	0.93	0.94	1.10	0.16	2.60**	1.23	0.12**	0.06**
	2nd × end-5%-apart	1.25	0.68	1.70	0.33	0.59	0.57	0.38	0.51	0.30**	0.31**	0.41**	0.01**	0.51*	0.55*	0.05**	0.02**
	3rd × Clust@45-55	5.81**	5.62**	5.95**	0.03**	12.53**	10.90**	4.97**	8.15**	1.82	2.81**	2.15*	0.07	0.82	0.79	0.49	0.09**
	3rd × Clust@50	4.56**	3.90**	1.99	1.36	2.07	2.04*	0.38	1.50	2.41*	5.93**	4.71**	*2.04*	0.45**	0.51*	5.65**	0.75
	3rd × End-Seq	15.86**	52.46**	26.10**	3.44	6.19**	5.53**	25.60**	12.43**	0.72	1.72	1.21	0.02**	0.43*	0.48*	5.98**	1.48
	3rd × Even	7.55**	4.59**	15.51**	0.55	67.08**	137.17**	180.99**	128.05**	0.56	0.56	0.89	0.19	1.62	3.02**	6.38**	0.08**
	3rd × beg-5%-apart	6.42**	6.07**	0.43*	3.04	1.20	2.10*	2.24*	1.85*	1.24	1.49	1.72	0.09	5.32**	4.57**	2.14**	3.11**
	3rd × end-5%-apart	79.24**	75.31**	55.15**	7.28**	64.92**	52.02**	58.24**	58.40**	1.66	1.86	1.95*	0.02**	0.51	0.36**	1.07	0.47
H7	2nd × 2x	0.53*	0.71	0.45*	0.42	0.37**	0.32**	1.46	0.71	0.94	2.06*	1.55	0.76	1.23	1.89*	2.14*	6.13**
	2nd × 5x	0.77	0.72	0.37*	0.30	0.63	0.42**	1.45	0.83	1.22	3.09**	2.18*	1.30	2.44**	2.34**	2.36*	4.66**
	3rd × 2x	1.34	1.24	1.93	0.60	0.77	0.50*	1.19	0.82	2.92**	1.98*	1.88*	12.22*	0.97	1.74*	5.45**	0.70
	3rd × 5x	1.00	1.11	1.26	0.21*	1.38	0.54	1.71	1.21	2.07*	1.60	1.75	2.47	1.64**	1.44**	7.38**	2.55**
H8	2x × Related	1.21	0.91	1.13	0.90	0.33**	0.56**	0.59*	0.51**	0.54**	0.56**	0.62**	0.04**	0.82	1.08	0.66	0.54
	5x × Related	1.18	1.16	1.37	2.05	0.32**	0.61*	0.48**	0.47**	0.53**	0.43**	0.51**	0.17**	0.59**	0.62**	0.54*	0.24**