Beyond Semantics: How Temporal Biases Shape Retrieval in Transformer and State-Space Models

Anooshka Bajaj* **Deven Mahesh Mistry*** **Sahaj Singh Maini**
Yash Aggarwal **Zoran Tiganj**
Department of Computer Science
Indiana University Bloomington
{anobajaj,demistry,sahmaini,yaggarw,ztiganj}@iu.edu

Abstract

In-context learning depends not only on *what* appears in the prompt but also on *when* it appears. To isolate this temporal component from semantic confounds, we construct prompts with repeated anchor tokens and average the model's predictions over hundreds of random permutations of the intervening context. This approach ensures that any observed position-dependent effects are driven purely by temporal structure rather than token identity or local semantics. Across four transformer LLMs and three state-space/recurrent models, we observe a robust serial recall signature: models allocate disproportionate probability mass to the tokens that previously followed the anchor, but the strength of this signal is modulated by serial position, yielding model-specific primacy/recency profiles. We then introduce an overlapping-episode probe in which only a short cue from one episode is re-presented; retrieval is reliably weakest for episodes embedded in the middle of the prompt, consistent with "lost-in-the-middle" behavior. Mechanistically, ablating high-induction-score attention heads in transformers reduces serial recall and episodic separation. For state-space models, ablating a small fraction of high-attribution channels produces analogous degradations, suggesting a sparse subspace supporting induction-style copying. Together, these results clarify how temporal biases shape retrieval across architectures and provide controlled probes for studying long-context behavior.

1 Introduction

The remarkable ability of Large Language Models (LLMs) for in-context learning (ICL) allows them to adapt to new tasks using only the information provided within the input prompt, without explicit parameter updates (Brown et al., 2020). However, while much research has focused on their semantic processing and reasoning capabilities, the mechanisms governing how LLMs retrieve and utilize contextual information, particularly concerning its temporal structure, remain less understood. It is increasingly recognized that the temporal position of information within the context significantly influences retrieval (Liu et al., 2024). In particular, models often exhibit better recall for information presented at the very beginning or end of the input context, a phenomenon termed the "lost in the middle" effect (Liu et al., 2024), mirroring the well-documented primacy and recency effects observed in human memory studies (Murdock, 1962; Ebbinghaus, 1913).

This parallel extends to the principles of human episodic memory, where the temporal organization of experiences is fundamental for segregating and retrieving specific past events, even when they share semantic content. This ability relies on encoding not just *what* happened, but *when* it happened relative to other events, enabling the formation and recall of distinct episodes (Howard and Kahana, 1999; Kahana, 1996). Recent work has begun exploring analogous processes in LLMs, investigating how they might implement memory-like functions to track and utilize temporal context (Ji-An et al., 2024; Mistry et al., 2025; Pink et al., 2024).

Mechanistically, specific components within transformer architectures, known as "induction heads", have been identified as crucial for ICL (Olsson et al., 2022; Elhage et al., 2021). These heads operate by finding previous occurrences of a current token and attending to the token that followed it, effectively learning and reproducing sequences based on temporal association (Olsson et al., 2022; Singh et al., 2024). Ablation studies confirm their essential role, demonstrating that removing induction heads significantly impairs ICL capabilities and temporal dependency processing, such as the tendency for serial recall (predicting the token immediately following a repeated token) (Mistry et al.,

*These authors contributed equally to this work.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 7562–7581

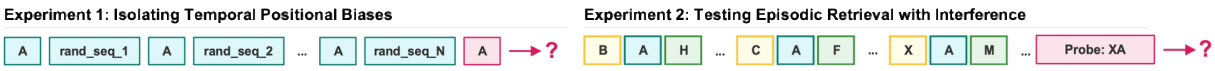

Figure 1: **Schematic of the experiments. Left:** Experiment 1 (serial recall). Prompts contain repeated occurrences of an anchor token A separated by random gap tokens. Across many permutations of the gap tokens, we measure the probability assigned to each successor (+1) token to isolate position-dependent retrieval biases (primacy/recency) from semantic confounds. **Right:** Experiment 2 (episodic retrieval with overlap). Five episodes (C_i, A, T_i) are embedded in long random fillers; the prompt ends with a probe pair (C_k, A). Successful retrieval corresponds to assigning high probability to the matching target token T_k while suppressing targets from the non-probed episodes.

2025; Crosbie and Shutova, 2024). Furthermore, the behavior of these induction heads exhibits characteristics reminiscent of human episodic memory recall, including temporal contiguity (recalling items presented close together in time) and forward asymmetry (preferential recall in the forward direction) (Ji-An et al., 2024; Mistry et al., 2025; Kahana, 1996). This connection has been formalized by linking induction head mechanisms to computational models of human memory like the Context Maintenance and Retrieval (CMR) model (Ji-An et al., 2024; Polyn et al., 2009; Howard and Kahana, 1999).

While semantic relationships in LLM retrieval are critical, temporal separation is essential for effective context use because semantics alone cannot disambiguate repeated content. Recent work has begun to map these dynamics, drawing connections between LLM behavior and serial-position effects like primacy and recency in both transformers and state-space models (Janik, 2023; Morita, 2025; Airlangga et al., 2025). However, differentiating pure temporal bias from semantic confounds remains a challenge. Our contribution addresses this by introducing a permutation-based probe that controls the marginal token distribution at each position, ensuring that observed modulations cannot be attributed to token identity or local semantic regularity. We further connect these biases to causal mechanisms via ablations, grounding broader phenomena of position bias in specific architectural ingredients (Wu et al., 2025). We apply this methodology to a diverse range of architectures, including transformers and state-space models (SSMs), prompting them with sequences containing multiple, temporally distinct presentations of the same tokens. By fixing the positions of these repeated tokens and analyzing next-token predictions across permutations of intervening tokens, we isolate the model's ability to use pure temporal cues for retrieval. Our findings reveal consistent temporal biases across all tested models, characterized by a distinct preference for information near the beginning or end of the context. In transformers, we mechanistically link this behavior to induction heads; remarkably, SSMs exhibit comparable biases despite their architectural differences, suggesting these are fundamental properties of sequential processing. These results expand our understanding of ICL dynamics, illustrating how temporal biases—analogous to principles of human memory—enable episodic separation within LLMs.

2 Experiments

To characterize temporal biases and the ability to retrieve temporal context, we conducted two experiments evaluating seven open-weight LLMs using distinct prompt structures. The selected models represent both the transformer architecture (Vaswani et al., 2017) and the SSM architecture (Gu and Dao, 2023).

Models based on the transformer architecture include Llama-3.1-8B-Instruct (Dubey et al., 2024), Gemma-2-9b-it (Team et al., 2024), Mistral-7B-Instruct-v0.1 (Jiang et al., 2023), and Qwen2.5-7B-Instruct (Yang et al., 2024). The SSM-based models include mamba-130m-hf (Gu and Dao, 2023), Falcon3-Mamba-7B-Instruct (Zuo et al., 2024), and Recurrent-Gemma-9b-it (Botev et al., 2024).

2.1 Permutation-controlled prompt family and scoring

Why permutations Token identity and frequency priors do influence next-token probabilities, even without context. Our goal is *not* to eliminate these priors; rather, we design prompts so that such priors are *constant across positions* in expectation. Under this construction, systematic differences between early and late retrieval signals are attributable to temporal structure (e.g., position bias, recency/primacy) rather than particular tokens occurring at particular positions.

Sampling scheme For each prompt instance, we sample a single anchor token A uniformly from

Algorithm 1 Experiment 1: Serial recall probe

Require: model p_θ, repeats N, gap length L, permutations M

for $m = 1$ to M **do**
  sample anchor A uniformly from allowed tokens
  for $i = 1$ to N **do**
    sample gap tokens $r_{i,1:L}$ uniformly without replacement from allowed tokens excluding A
  end for
  $x \leftarrow (A, r_{1,1:L}, A, r_{2,1:L}, \ldots, A, r_{N,1:L}, A)$
  compute $p_\theta(\cdot \mid x)$ and record $p_i^{(+1)} = p_\theta(r_{i,1} \mid x)$
end for
return $\{p_i^{(+1)}\}_{i=1}^{N}$ statistics across permutations

the vocabulary. All non-anchor tokens are sampled uniformly from the remaining pool, excluding A. Within each gap (the tokens between anchor occurrences), we sample without replacement so the marginal distribution of the token at any fixed within-gap position is identical across positions.

Position-indexed probability profile Given a prompt $x = (x_1, \ldots, x_T)$ and the model next-token distribution $p_\theta(\cdot \mid x)$, we define a *position-indexed probability profile* by mapping each position t to the probability assigned to the token that actually appears at that position:

$$\pi_t(x) \;=\; p_\theta(x_t \mid x).$$

Our figures plot $\pi_t(x)$ against token position t. Throughout, vertical gray lines indicate anchor occurrences.

+1 tokens If anchor occurrences are at positions $\{a_1, \ldots, a_N\}$, we call the successor tokens x_{a_i+1} the +1 tokens. We summarize serial recall with

$$p_i^{(+1)} \;=\; p_\theta(x_{a_i+1} \mid x),$$

and report means and standard errors across M independently sampled prompts (typically $M = 5000$).

2.2 Experiment 1: Examining Temporal Positional Preferences

This experiment investigated inherent temporal biases in LLM retrieval, independent of semantic content. We designed prompts containing repetitions of a specific token ('anchor token') interspersed with sequences of random tokens, and quantified how the predicted probability of the next token is influenced by the temporal position of these repetitions (Figure 1, left).

Prompt construction We construct prompts containing N occurrences of a single anchor token A, each followed by a length-L gap of random tokens, and a final anchor A at the end (Algorithm 1). The token immediately following the i-th anchor occurrence is the +1 token for that occurrence.

Quantification For each prompt instance x, we compute the next-token distribution $p_\theta(\cdot \mid x)$ and form the position-indexed profile $\pi_t(x) = p_\theta(x_t \mid x)$ (Section 2.1). Serial recall is summarized by $p_i^{(+1)} = p_\theta(x_{a_i+1} \mid x)$, the probability assigned to the +1 token from the i-th anchor. We report means and SEM across $M = 5000$ independently sampled prompts per condition.

Results Figure 2 shows the resulting probabilities (averaged across permutations) as a function of token position within the prompt (position 0 is the start), varying the number of repetitions (columns) while keeping the spacing fixed at 10 tokens. Figure A3 shows results varying the spacing (columns) with a fixed number of 10 repetitions. Probabilities for the repeated fixed token itself were excluded from visualization due to their artificially high values resulting from repetition frequency.

Across all seven models, probabilities exhibit distinct peaks corresponding to the token immediately following each instance of the fixed token (the +1 token). This preference for the +1 token demonstrates a tendency towards serial recall (repeating sequences in the order presented), consistent with previous findings in transformer models (Mistry et al., 2025).

Our results reveal several novel insights: 1) *SSMs exhibit serial recall:* SSMs demonstrate a similar +1 token preference, indicating a comparable tendency for serial recall despite architectural differences from transformers. 2) *Positional modulation of peaks:* The magnitude of the +1 probability peaks varies depending on the repetition's position within the prompt, indicating temporal biases (primacy/recency). This is highlighted in Figure A1, which plots only the +1 token probabilities against the repetition number (depth in prompt). 3) *Model-specific biases:* These temporal preferences differ across models. Mistral shows a recency bias (higher probabilities towards the end). Falcon-Mamba exhibits a primacy bias (higher probabilities near the beginning). Gemma's preference shifts: with fewer repetitions, the peak is mid-prompt, but with more repetitions, it shifts towards the end. These shifts are also observable

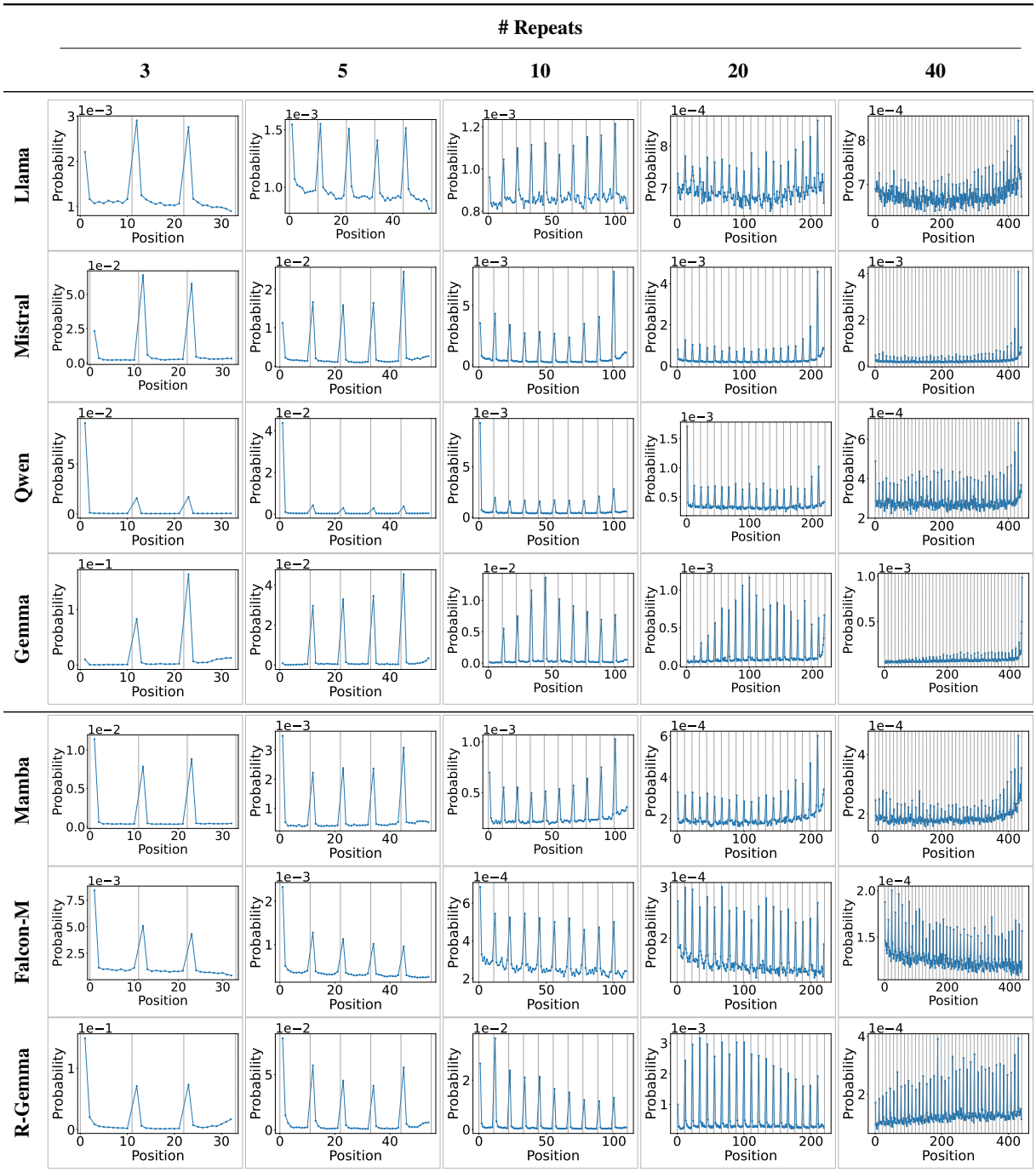

Figure 2: Experiment 1: Next-token probability vs. position for varying number of fixed-token repetitions (columns) across models (rows). Vertical gray lines mark the anchor token. Peaks in the blue lines confirm the presence of +1 recall preference and positional biases.

when varying spacing (Figure A3). No globally distinct temporal patterns consistently separated transformer and state-space architectures.

2.3 Experiment 2: Characterizing Episodic Retrieval as a Function of Temporal Distance and Context Overlap

Experiment 1 probes serial recall for a repeated anchor in a largely context-free setting. Here we test a more episodic setting with partially overlapping structure, in which retrieval requires selecting the correct continuation given a short cue.

Prompt Construction We created prompts containing five distinct episodes. Each episode consisted of a unique 'context' token, followed by the 'anchor' token (same 'A' across all episodes), followed by a unique 'target' token (e.g., 'BAH', 'CAF', 'XAM', 'GAD', 'RAP'). These five

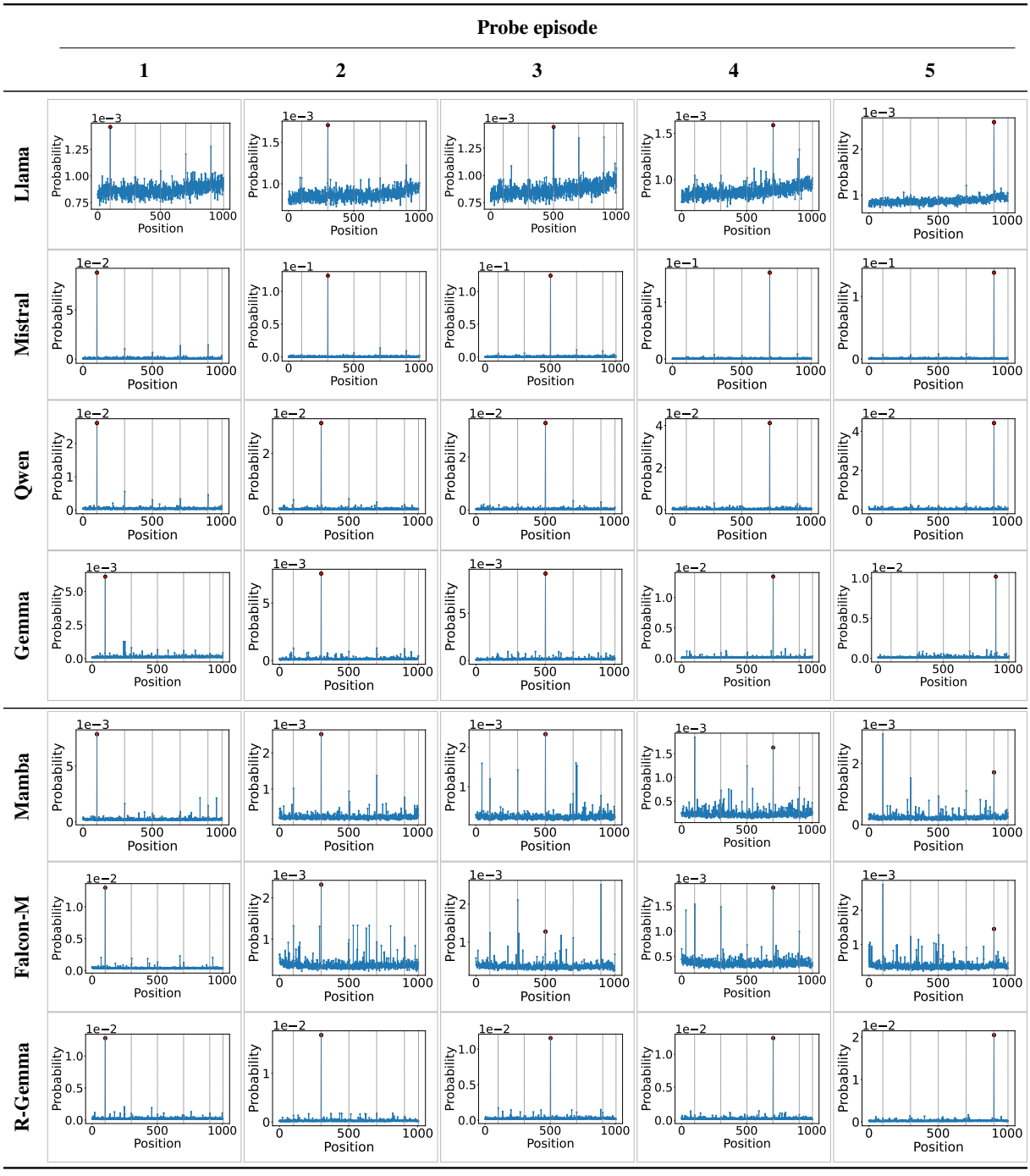

Figure 3: Experiment 2: Episodic retrieval results. Probability of potential target tokens after probing episodes at different temporal positions (columns 1-5). Vertical gray lines mark the anchor token. Red dots mark the target token corresponding to the probed episode. A high probability at the red dot indicates successful retrieval of the episode's target token.

episodes were embedded within longer sequences of random tokens, separated by 200 random tokens each. The final probe consisted of the context token and fixed token from one specific target episode (e.g., ending the prompt with '...XA'), Figure 1, right.

Quantification The models were evaluated on their ability to predict the token following the probe pair (e.g., predict 'M' given 'XA'). Successful retrieval requires distinguishing the target episode ('XAM') from the others sharing the fixed token 'A'. We systematically varied which of the five episodes was used for the probe (testing retrieval from positions 1 through 5). Probabilities were averaged over 500 permutations with shuffled intervening random tokens to isolate temporal effects.

Results Figure 3 displays the next-token probabilities following the probe, plotted against the five possible target episode positions (columns 1-5). Each plot shows the probabilities assigned to the five potential target tokens ('H', 'F', 'M', 'D', 'P' in the example) and other tokens at the relevant positions (the fixed token 'A' probability is omitted).

When probing for the first episode (Column 1, episode near the start), most models correctly assign the highest probability to the target token corresponding to that episode (e.g., 'H' for probe 'BA'), indicating successful retrieval. This pattern holds when probing for subsequent episodes (Columns 2-5). Llama, Mistral, Qwen, Gemma, and RecurrentGemma consistently assign the highest probability to the correct target token across most positions. However, Mamba and Falcon-Mamba show less robust retrieval, particularly when the target episode is closer to the end of the prompt (e.g., column 5).

Beyond the highest peak for the target episode, smaller peaks corresponding to the target tokens of non-probed episodes are often visible, indicating interference or similarity matching based on the shared fixed token 'A'. The magnitude of the correct target peak often varies with position, generally being strongest for episodes nearer the end of the prompt (recency bias) and weaker for earlier episodes in models like Llama, Mistral, Qwen, and RecurrentGemma.

2.4 Ablation Study

2.4.1 Transformers: ablating high-induction-score heads

To investigate the mechanisms underlying these temporal effects, we performed an ablation study focusing on induction heads, known contributors to ICL and temporal processing (Olsson et al., 2022; Elhage et al., 2021; Ji-An et al., 2024). As defined by Olsson et al. (2022), induction heads are attention heads exhibiting pattern completion or copying behavior; they identify previous occurrences of the current token and attend to the subsequent token.

Procedure For each transformer model (Llama, Mistral, Qwen, Gemma), we calculated induction scores for all attention heads (Layer $\times$ Head) using the method described in Ji-An et al. (2024). Heads were ranked by their induction scores. We then progressively ablated (set attention scores to zero, following Crosbie and Shutova (2024)) the top 1, 10, 50, and 100 induction heads. As a control, we ablated the same numbers of randomly selected heads (ensuring they were not among the top 100 induction heads).

Results Figure 4 shows the ablation results for Experiment 1. Ablating top induction heads consistently and significantly degrades the +1 token probability peaks, particularly with 50 or 100 heads ablated. Ablating random heads has a much less systematic effect, sometimes slightly perturbing probabilities but generally preserving the +1 peaks. This confirms the critical role of induction heads in the serial recall behavior. Across 4×4 (models$\times$ablation levels), the drop in the average +1 probability after ablating high–induction heads exceeds the drop after ablating the same number of random heads in $13/16$ settings (median difference 3.52×10^{-5}, IQR 1.12×10^{-5}–6.09×10^{-5}; maximum difference 2.26×10^{-4} for Llama-3.1-8B with 100 heads). The few negative differences are small (at most 1.46×10^{-5}). See Table A1 for complete numerical values. These results, together with the shape changes in Figure 4, indicate that the serial-recall signal is disproportionately carried by high–induction heads, and the effect generally strengthens as more such heads are ablated (most clearly in Llama).

Similar effects were observed in Experiment 2 (Figures A4-A7). Ablating induction heads (top 5 rows in each figure, corresponding to probes 1-5) often disrupted the model's ability to selectively retrieve the single target episode. Instead of a dominant peak for the correct token, probabilities became more distributed across the potential target tokens from different episodes, indicating increased interference. Ablating random heads (bottom 5 rows) generally had a weaker impact, though ablating 50 or 100 random heads sometimes also perturbed the output distributions significantly. These results further underscore the importance of induction heads for temporal context separation and retrieval in transformers.

2.4.2 State-space models: ablating high-attribution channels

State-space models lack attention heads, so we adapt the same causal logic to channels (hidden dimensions) of the selective state-space block.

Procedure For each channel c, we compute an "induction-style" salience score using a gradient-attribution proxy aggregated over prompts:

$$s_c = \mathbb{E}\left[\left| u_c \cdot \frac{\partial \log p_\theta(y \mid x)}{\partial u_c} \right|\right],$$

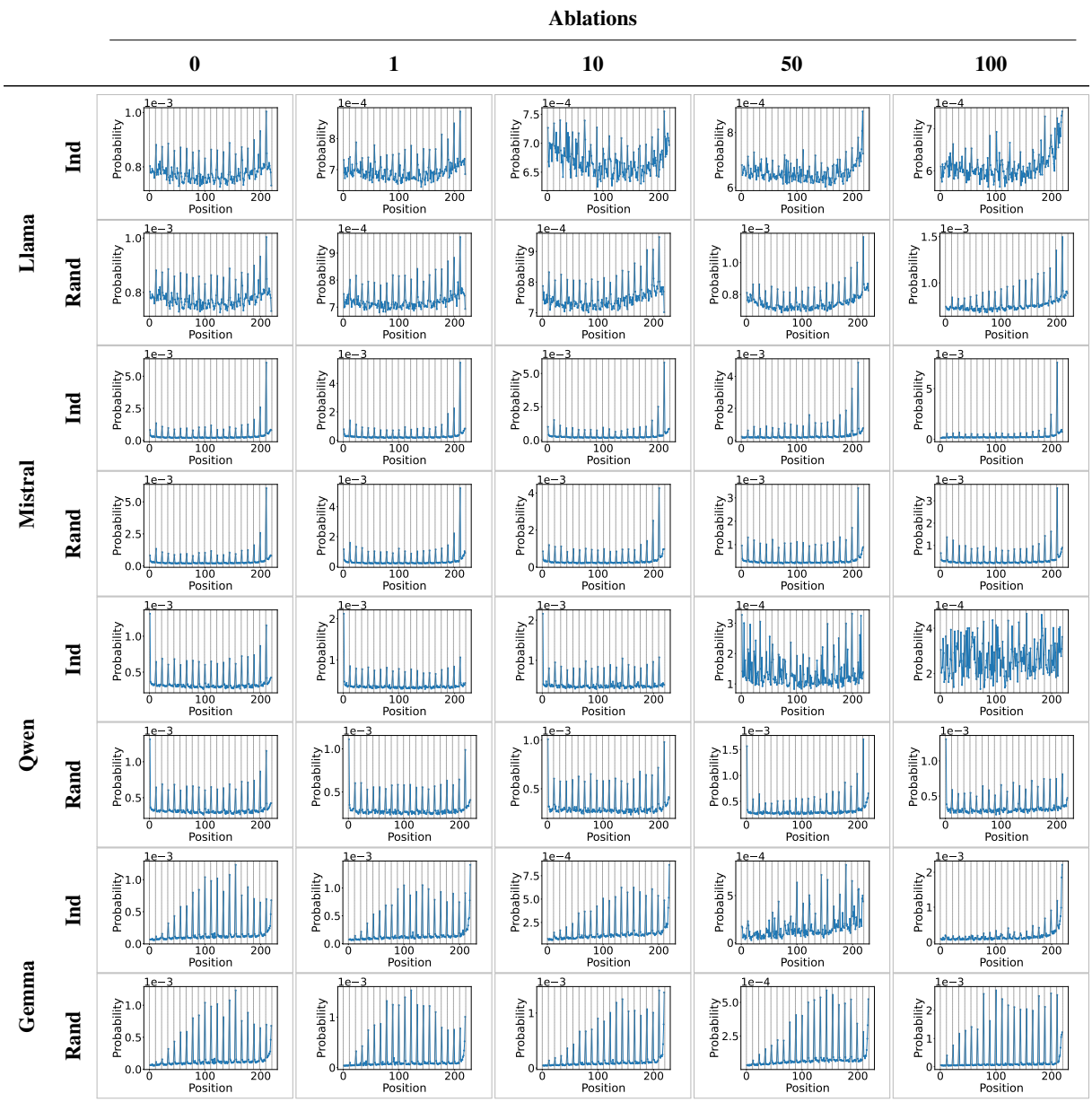

Figure 4: Transformer ablation effect (Exp. 1). Probability vs. position after ablating Induction (Ind) or Random (Rand) heads (rows) for Llama, Mistral, Qwen, Gemma (model pairs per 2 rows). Columns show the number of ablated heads.

where u denotes the residual update produced by the selective SSM block and y is the correct +1 token (Experiment 1) or correct target token (Experiment 2). We ablate the top-ranked channels by zeroing their contribution to the residual update, and compare to ablating a matched number of random channels. To compare fairly across model sizes, we ablate fixed fractions of channels (1%, 5%, and 10%).

Results Figure A2 shows impact of ablation in Experiment 1 and Figures A8 and A9 in Experiment 2 (see also Table A2 for numerical values). In both experiments, for sufficiently large ablations (5–10%), removing high-attribution channels reduces the recall signal more than random ablations across the tested SSM-family models, supporting the presence of a sparse subspace that is disproportionately responsible for the copy-like behavior. Importantly, we do not claim a one-to-one equivalence between transformer induction heads and SSM channels; rather, these channels are *functionally analogous* in the sense that they are causally implicated in the same induction-style behavioral signature.

These findings are consistent with recent analy-

ses arguing that selective SSMs can be interpreted through an implicit-attention lens (Ali et al., 2025; Zimerman et al., 2024), and with reports of sparse memory channels that support primacy/recency effects in Mamba-like models (Airlangga et al., 2025; Morita, 2025).

3 Discussion

This study investigated the role of temporal structure in shaping information retrieval within LLMs. By employing experiments designed to isolate temporal effects, we probed how both transformer and SSM architectures handle repeated information and overlapping temporal contexts. Our findings reveal significant temporal biases influencing retrieval and offer insights into the mechanisms underpinning these effects.

A key finding is a better understanding of positional biases in LLM context utilization, often described as the "lost in the middle" phenomenon (Liu et al., 2024). Our results from Experiment 1 demonstrate that models exhibit strong primacy and recency effects, favoring information associated with tokens at the beginning or end of the context, even when semantic content is neutralized through permutation. This suggests the bias is deeply rooted in the sequential processing capabilities of these models, not merely an artifact of document structure or semantic coherence. The specific nature of this bias (primacy vs. recency dominance) varied across models and context parameters (Figures 2, A1), hinting at complex interactions between model architecture, training data, and the specific structure of the input sequence. In transformers, "attention sinks" can allocate baseline attention to early tokens (Xiao et al., 2023; Gu et al., 2024). These sinks plausibly amplify primacy by biasing which earlier occurrence is selected by induction heads.

Furthermore, Experiment 1 established that the tendency for serial recall, prioritizing the token immediately following a repeated instance (+1 token), is robust across both transformer and SSM architectures. This echoes prior work on induction heads in transformers (Olsson et al., 2022; Mistry et al., 2025) but for the first time demonstrates its presence in SSMs like Mamba and Recurrent-Gemma as well. This shared characteristic suggests that basic sequence copying or pattern completion might be a convergent capability learned by different sequence modeling architectures.

Experiment 2 tests the models' ability to perform episodic-like retrieval by distinguishing between partially overlapping temporal contexts based on unique preceding tokens. Most models demonstrated a capacity for this temporal separation, correctly identifying the target token associated with the probed episode (Figure 3). However, this retrieval was imperfect, often showing interference from competing, temporally adjacent episodes (non-target peaks) and exhibiting strong positional effects, with episodes located near the end of the prompt generally retrieved more reliably (recency bias). This resonates with computational models of human episodic memory, which explicitly account for interference based on contextual similarity and temporal distance (Howard and Kahana, 1999; Polyn et al., 2009). While LLMs showed some ability for temporal separation, their strong reliance on serial position contrasts with the more graded temporal contiguity effects seen in human recall (Kahana, 1996; Ji-An et al., 2024), suggesting potential differences in how temporal context is represented and utilized.

Our ablation studies in both transformer and state space models (Figures 4, A2, A4 –A9) provide mechanistic insights, confirming the crucial role of induction heads and their SSM analogues. Ablation significantly impaired both the +1 serial recall preference in Experiment 1 and the ability to selectively retrieve the correct episode in the presence of interference in Experiment 2. This aligns with the established function of induction heads in pattern matching and ICL (Olsson et al., 2022; Elhage et al., 2021; Crosbie and Shutova, 2024) and supports their proposed link to episodic memory functions (Ji-An et al., 2024).

Despite lacking explicit attention mechanisms, SSMs showed serial-position structure in retrieval. For example, Recurrent-Gemma displayed U-shaped positional preference curves in the serial-recall probe (Figure A1) and a recency advantage in episodic retrieval under overlap (Figure 3), whereas Mamba and Falcon-Mamba exhibited less robust episodic retrieval when the probed episode occurred near the end of the prompt (Figure 3). Together, these patterns suggest that temporal biases are not solely a consequence of attention, but can also arise from more general constraints on representing and querying long sequences. Potential contributing factors include how models encode positional information (explicitly via positional representations in transformers, or implicitly through

recurrent/state dynamics in SSMs) and limitations in maintaining distinct, queryable traces over long temporal distances. SSMs, for instance, compress context history into a fixed-size state (Gu and Dao, 2023; Jelassi et al., 2024); the evolution, saturation, and selective-forgetting dynamics of this state could systematically privilege certain parts of the sequence, yielding primacy/recency-like effects. Consistent with this, our ablations in state-space models show that removing a small fraction of high-attribution channels produces degradations analogous to ablating high-induction heads in transformers, supporting the presence of a sparse subspace that implements induction-style copying and episodic separation. While some studies suggest transformers excel at exact copying tasks (Jelassi et al., 2024), our interference-based probes indicate that when retrieval depends on relative temporal position and resisting overlap, SSMs can exhibit comparable position-dependent limitations and biases.

These findings have several implications. For LLM development, they underscore that addressing the "lost in the middle" problem requires tackling fundamental temporal processing limitations, potentially related to positional information or state management, which may persist even in non-attention architectures. For cognitive science, our methodology provides a controlled paradigm for comparing how different computational architectures handle temporal context and interference, offering insights into the functional constraints shaping memory-like phenomena in artificial systems.

Limitations

We primarily study token-random prompts rather than natural text and analyze next-token probabilities. While methodologically necessary to isolate temporal effects from semantic confounds, random-token sequences represent a simplified information environment. Future work can build upon this understanding of pure temporal biases to explore the more complex interplay between *when* something was said and *what* was said in rich, semantic contexts.

Our SSM "high-attribution channel" identification uses gradient-based salience as a heuristic. Although targeted ablations provide causal evidence that a sparse subset of channels supports the induction-style signal, this is not a one-to-one analog of transformer induction heads, and direct alignment with other proposed SSM memory decompositions remains future work.

References

Muhammad Cendekia Airlangga, Hilal AlQuabeh, Munachiso S Nwadike, and Kentaro Inui. 2025. Emergence of primacy and recency effect in mamba: A mechanistic point of view. *arXiv preprint arXiv:2506.15156*.

Ameen Ali, Itamar Zimerman, and Lior Wolf. 2025. The hidden attention of mamba models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Aleksandar Botev, Soham De, Samuel L Smith, Anushan Fernando, George-Cristian Muraru, Ruba Haroun, Leonard Berrada, Razvan Pascanu, Pier Giuseppe Sessa, Robert Dadashi, and 1 others. 2024. Recurrentgemma: Moving past transformers for efficient open language models. *arXiv preprint arXiv:2404.07839*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, ..., and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.

Joy Crosbie and Ekaterina Shutova. 2024. Induction heads as an essential mechanism for pattern matching in in-context learning. *arXiv preprint arXiv:2407.07011*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hermann Ebbinghaus. 1913. *Memory: A contribution to experimental psychology*. 3. Teachers college, Columbia university.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, and 1 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. 2024. When attention sink emerges in language models: An empirical view. *arXiv preprint arXiv:2410.10781*.

Marc W Howard and Michael J Kahana. 1999. Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(4):923.

Romuald A Janik. 2023. Aspects of human memory and large language models. *arXiv preprint arXiv:2311.03839*.

Samy Jelassi, David Brandfonbrener, Sham M Kakade, and Eran Malach. 2024. Repeat after me: Transformers are better than state space models at copying. *arXiv preprint arXiv:2402.01032*.

Li Ji-An, Corey Y Zhou, Marcus K Benna, and Marcelo G Mattar. 2024. Linking in-context learning in transformers to human episodic memory. *arXiv preprint arXiv:2405.14992*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Michael J Kahana. 1996. Associative retrieval processes in free recall. *Memory & Cognition*, 24(1):103–109.

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.

Deven Mahesh Mistry, Anooshka Bajaj, Yash Aggarwal, Sahaj Singh Maini, and Zoran Tiganj. 2025. Emergence of episodic memory in transformers: Characterizing changes in temporal structure of attention scores during training. *arXiv preprint arXiv:2502.06902*.

Takashi Morita. 2025. Emergence of the primacy effect in structured state-space models. *arXiv preprint arXiv:2502.13729*.

Bennet B. Jr. Murdock. 1962. The serial position effect of free recall. *Journal of Experimental Psychology*, 64(5):482–488.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, and 1 others. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Mathis Pink, Vy A Vo, Qinyuan Wu, Jianing Mu, Javier S Turek, Uri Hasson, Kenneth A Norman, Sebastian Michelmann, Alexander Huth, and Mariya Toneva. 2024. Assessing episodic memory in llms with sequence order recall tasks. *arXiv preprint arXiv:2410.08133*.

Sean M Polyn, Kenneth A Norman, and Michael J Kahana. 2009. A context maintenance and retrieval model of organizational processes in free recall. *Psychological review*, 116(1):129.

Aaditya Singh, Stephanie Chan, Ted Moskovitz, Erin Grant, Andrew Saxe, and Felix Hill. 2024. The transient nature of emergent in-context learning in transformers. *Advances in Neural Information Processing Systems*, 36.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Xinyi Wu, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. 2025. On the emergence of position bias in transformers. *arXiv preprint arXiv:2502.01951*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Itamar Zimerman, Ameen Ali, and Lior Wolf. 2024. Explaining modern gated-linear rnns via a unified implicit attention formulation. *arXiv preprint arXiv:2405.16504*.

Jingwei Zuo, Maksim Velikanov, Dhia Eddine Rhaiem, Ilyas Chahed, Younes Belkada, Guillaume Kunsch, and Hakim Hacid. 2024. Falcon mamba: The first competitive attention-free 7b language model. *arXiv preprint arXiv:2410.05355*.

Model	Ablation Level	Induction Change	Random Change	Difference in Change Δ
Llama	1	9.35×10^{-5}	4.96×10^{-5}	4.39×10^{-5}
	10	1.34×10^{-4}	3.81×10^{-5}	9.63×10^{-5}
	50	1.81×10^{-4}	1.72×10^{-5}	1.64×10^{-4}
	100	2.24×10^{-4}	-1.91×10^{-6}	2.26×10^{-4}
Mistral	1	1.90×10^{-5}	-7.51×10^{-6}	2.65×10^{-5}
	10	2.09×10^{-5}	2.03×10^{-6}	1.88×10^{-5}
	50	6.27×10^{-5}	-4.65×10^{-6}	6.74×10^{-5}
	100	5.85×10^{-5}	-1.19×10^{-7}	5.87×10^{-5}
Qwen	1	-2.62×10^{-6}	8.58×10^{-6}	-1.12×10^{-5}
	10	1.57×10^{-5}	3.46×10^{-6}	1.23×10^{-5}
	50	7.45×10^{-5}	2.49×10^{-5}	4.95×10^{-5}
	100	7.93×10^{-5}	3.17×10^{-5}	4.76×10^{-5}
Gemma	1	-6.59×10^{-6}	6.56×10^{-7}	-7.24×10^{-6}
	10	-1.31×10^{-5}	1.46×10^{-6}	-1.46×10^{-5}
	50	1.72×10^{-5}	-2.80×10^{-6}	2.00×10^{-5}
	100	1.53×10^{-5}	7.27×10^{-6}	8.00×10^{-6}

Table A1: Ablation effects on the average +1 probability in Exp. 1. *Induction Change* and *Random Change* are changes in +1 probability relative to the unablated model. Δ = Induction Change − Random Change; positive Δ means ablating high-induction heads reduces the $(+1)$ signal more than ablating random heads.

Experiment	Model / Ablation	Induction Change	Random Change
Exp. 1	Mamba, 1%	+67.8%	-91.7%
	Mamba, 5%	+98.4%	+6.2%
	Mamba, 10%	+94.2%	+58.3%
	R-Gemma, 1%	-17.2%	+9.6%
	R-Gemma, 5%	+68.2%	+43.2%
	R-Gemma, 10%	+96.9%	+58.9%
	Falcon-M, 1%	+11.2%	+35.7%
	Falcon-M, 5%	+92.6%	+11.0%
	Falcon-M, 10%	+90.4%	+38.5%
Exp. 2	Mamba, 1%	+16.4%	-69.7%
	Mamba, 5%	+64.1%	+27.3%
	Mamba, 10%	+82.3%	+62.1%
	R-Gemma, 1%	-7.3%	+14.7%
	R-Gemma, 5%	+65.2%	+44.0%
	R-Gemma, 10%	+91.6%	+70.1%

Table A2: State-space/recurrent channel ablations. "Change" denotes the relative change in the task signal. Experiment 1: mean +1 recall; Experiment 2: target probability / separation signal after ablating high-attribution channels (Induction) vs. a matched number of random channels. Falcon-M was too computationally demanding for Experiment 2 ablations.

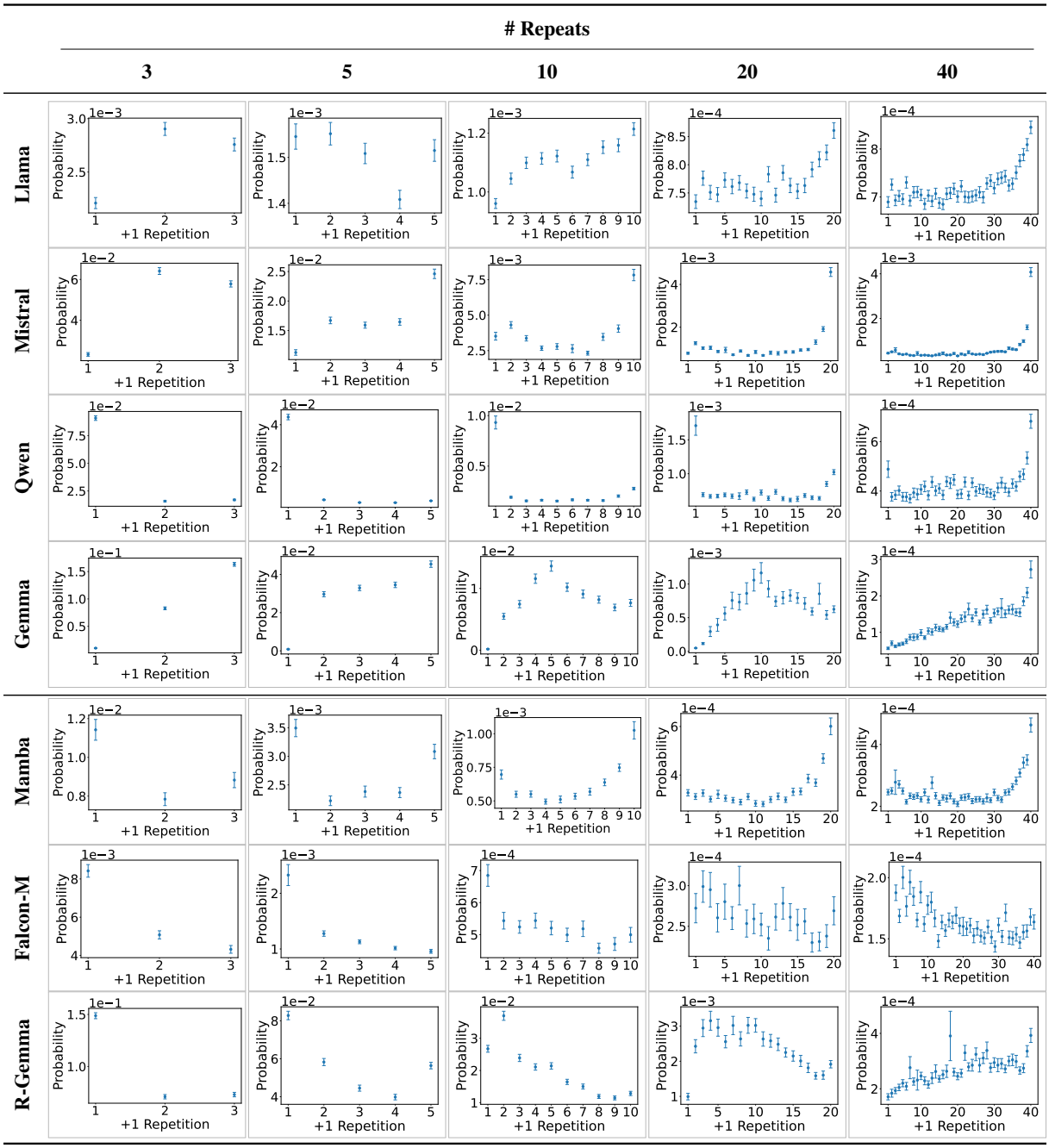

Figure A1: Probability (±SEM) of +1 token retrieval vs. repetition position, for varying number of fixed-token repetitions (columns).

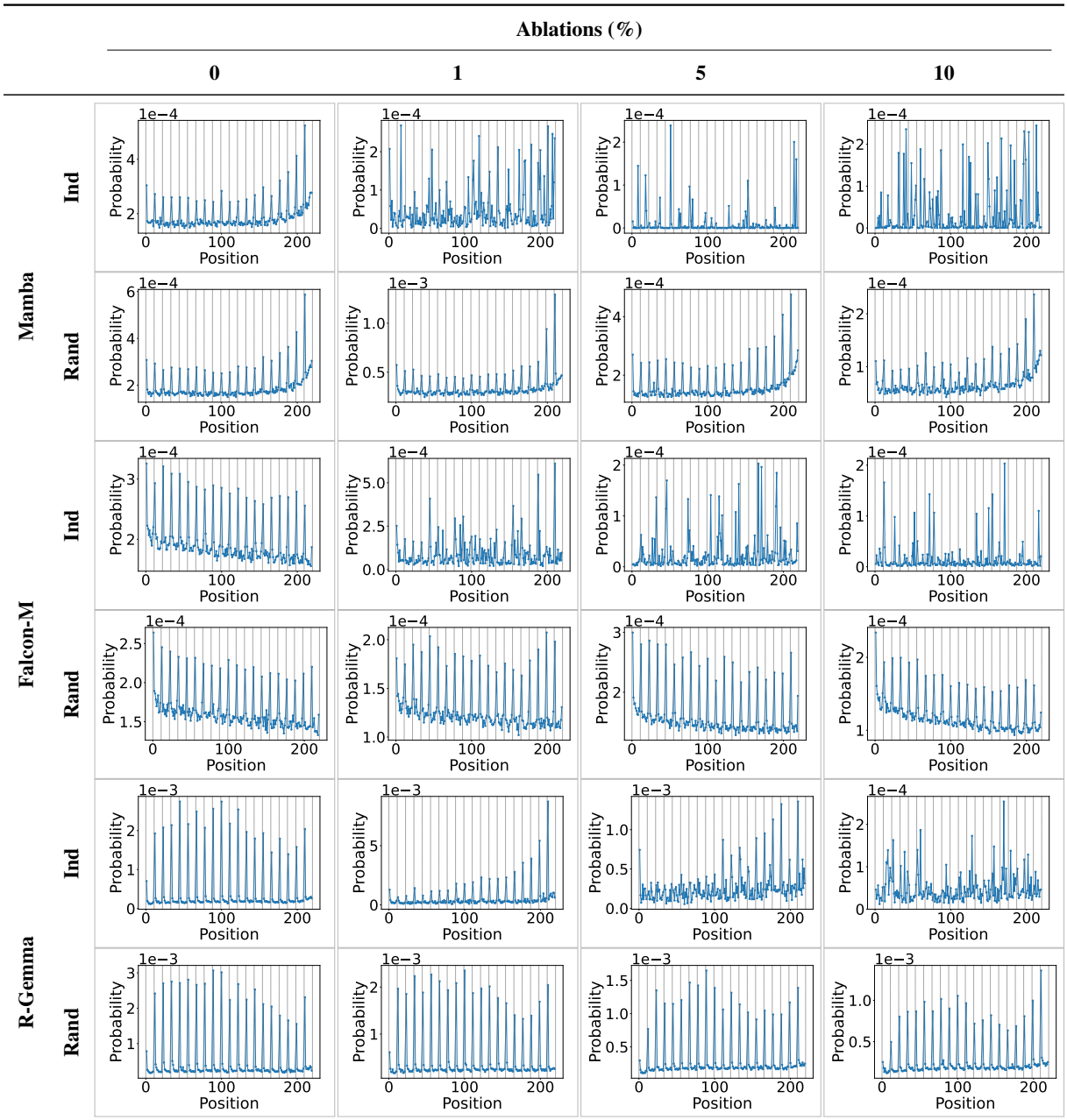

Figure A2: SSM ablation effect (Exp. 1). Probability vs. position after ablating Induction (Ind) or Random (Rand) channels (rows) for Mamba, Falcon-Mamba, Recurrent-Gemma (model pairs per 2 rows). Columns show the percentage of ablated channels.

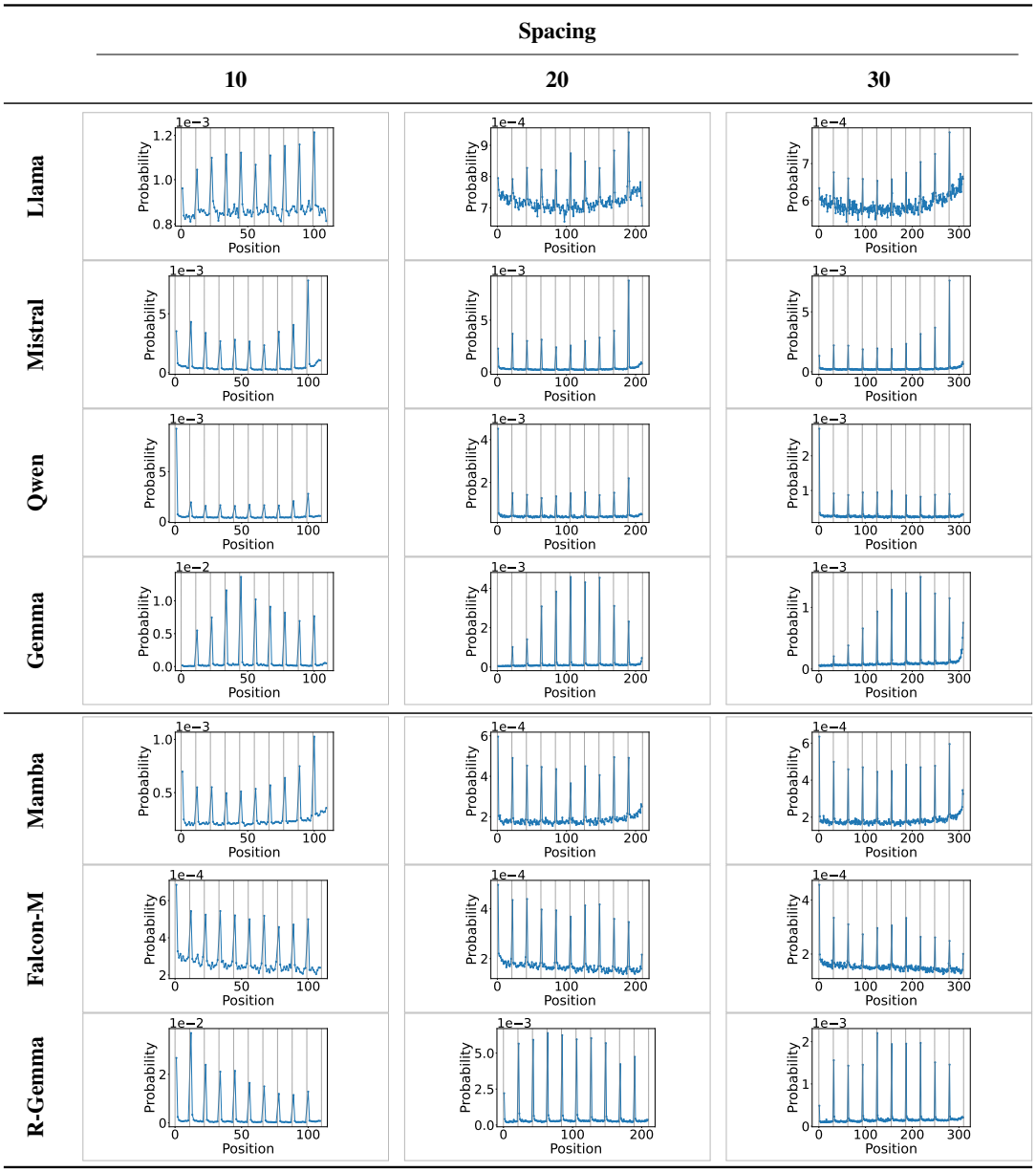

Figure A3: Next-token probability vs. position, varying spacing between fixed tokens (columns, repeats=10). Shows +1 recall preference and positional biases.

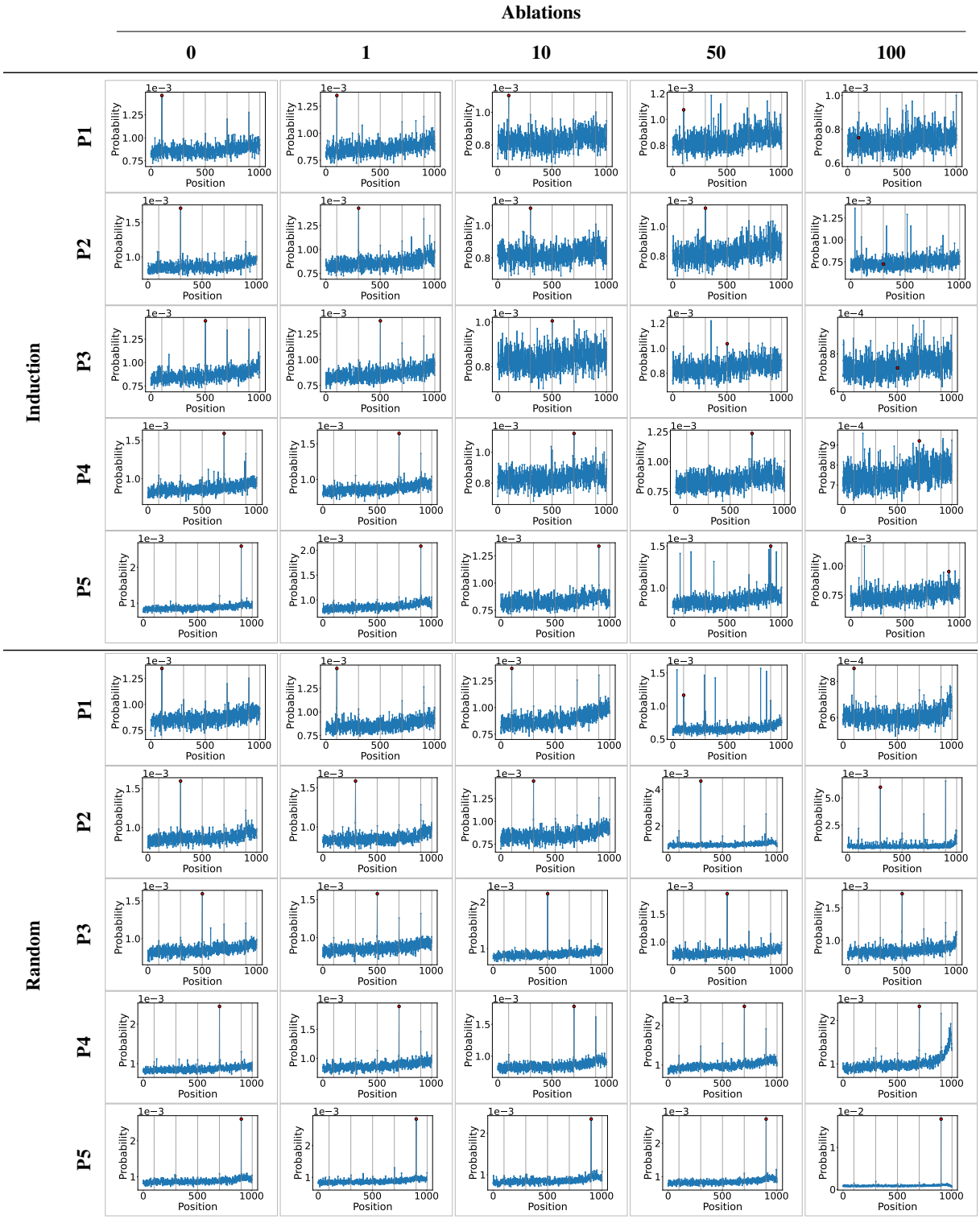

Figure A4: Llama ablation effect (Exp. 2). Episodic retrieval probability after ablating Induction or Random heads (rows) probing different episode positions. Columns show number of ablated heads. Red dots mark the target token corresponding to the probed episode.

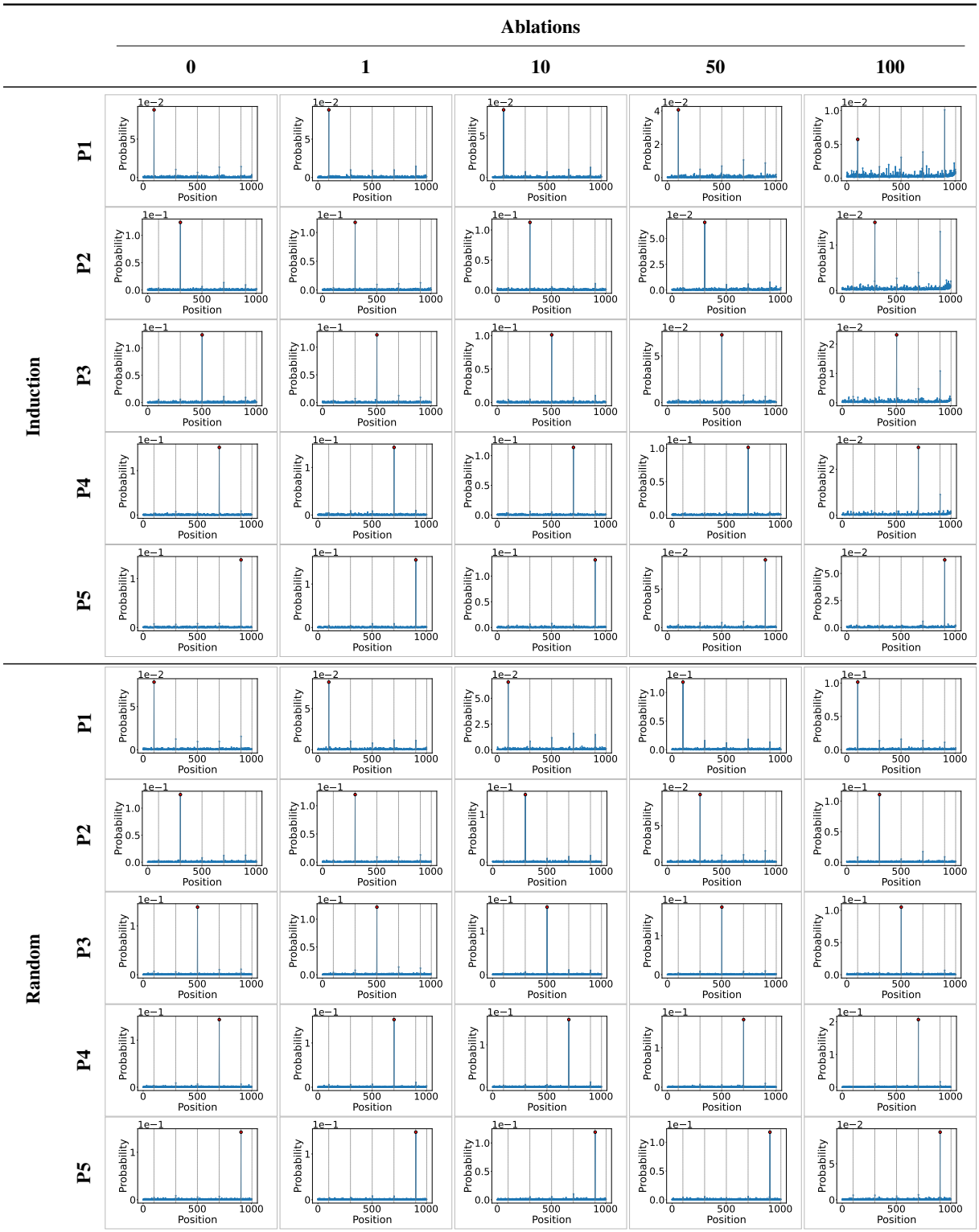

Figure A5: Mistral ablation effect (Exp. 2). Episodic retrieval probability after ablating Induction or Random heads. Columns show number of ablated heads.

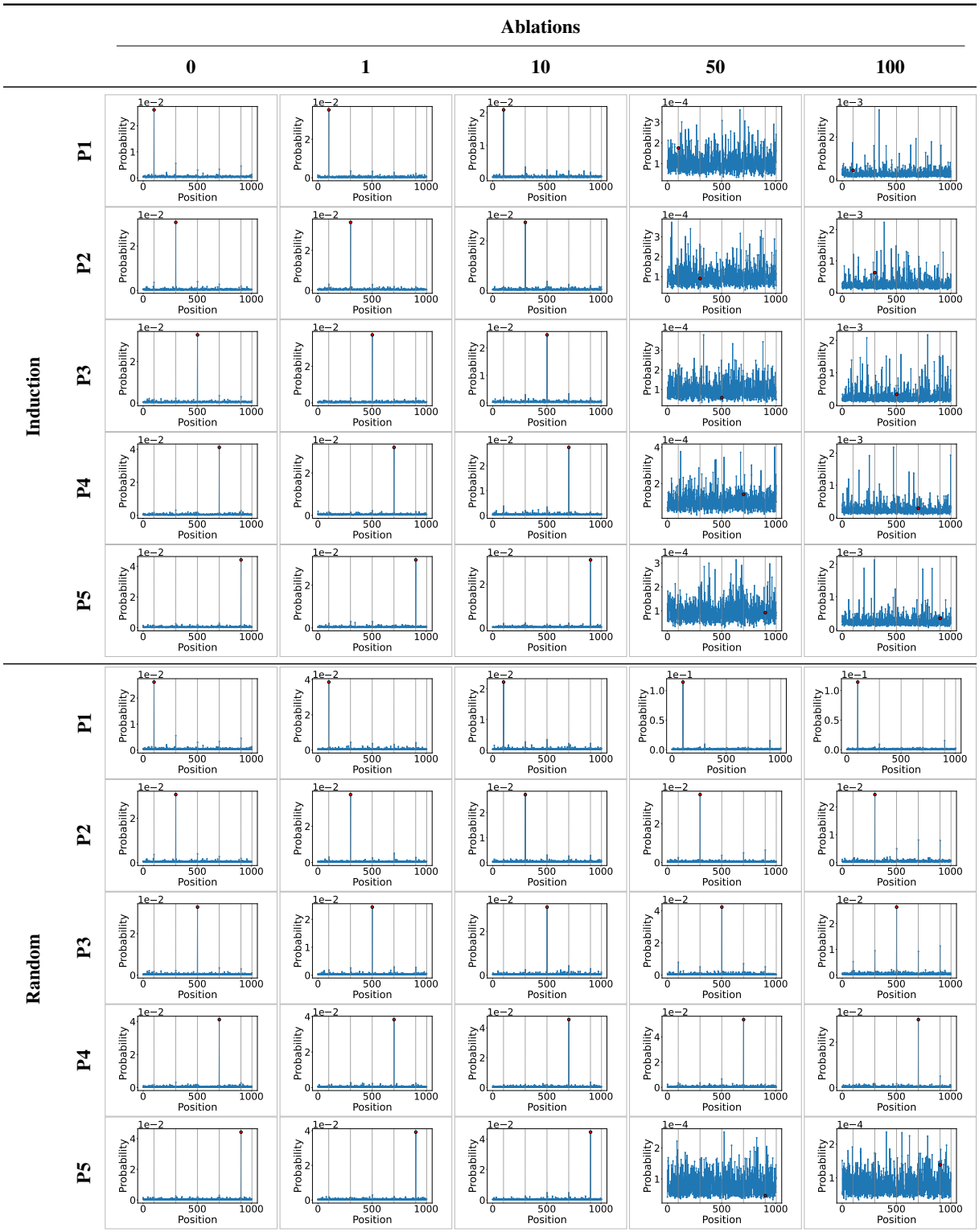

Figure A6: Qwen ablation effect (Exp. 2). Episodic retrieval probability after ablating Induction (Ind P1-P5) or Random (Rand P1-P5) heads. Columns show number of ablated heads.

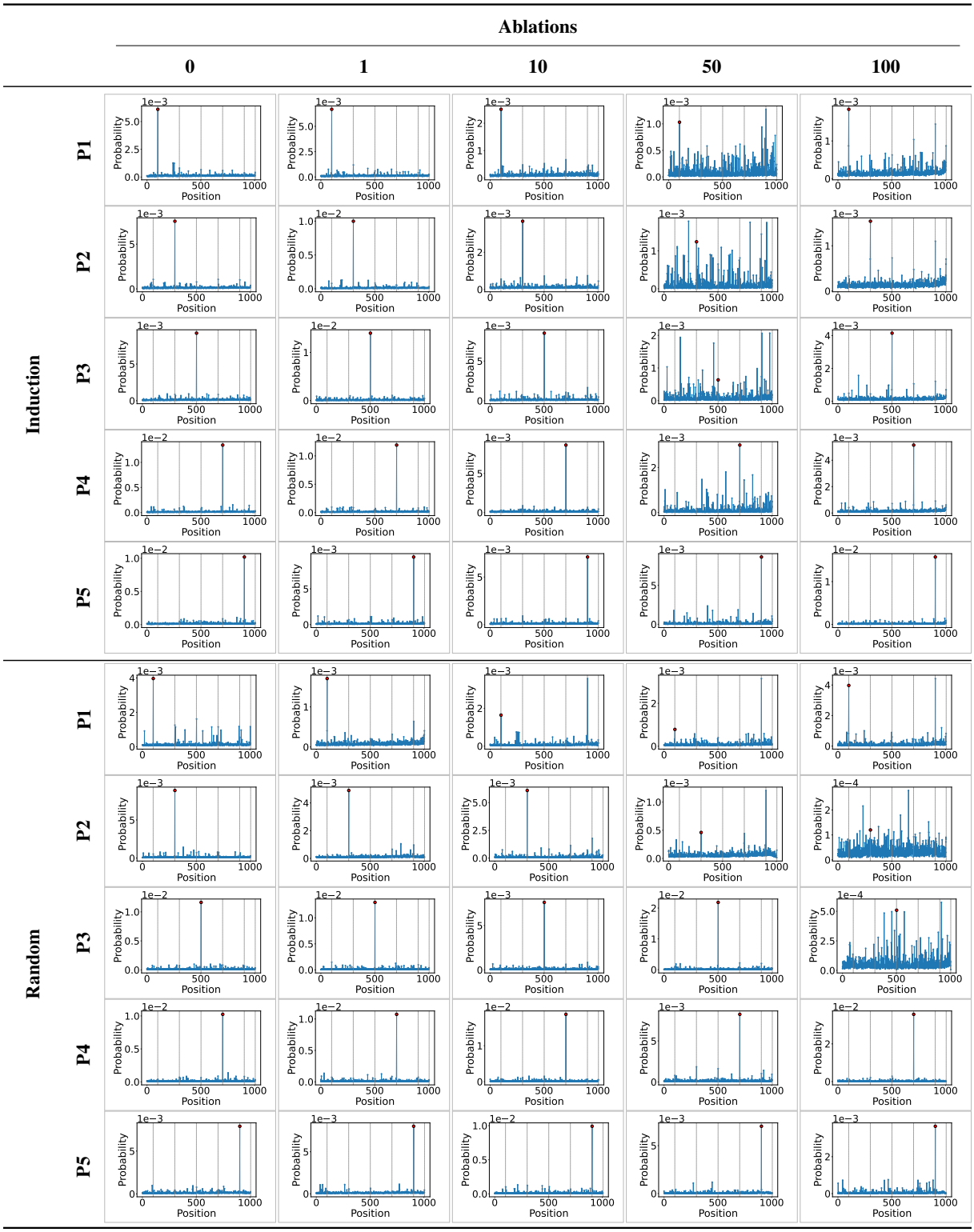

Figure A7: Gemma ablation effect (Exp. 2). Episodic retrieval probability after ablating Induction or Random heads. Columns show number of ablated heads.

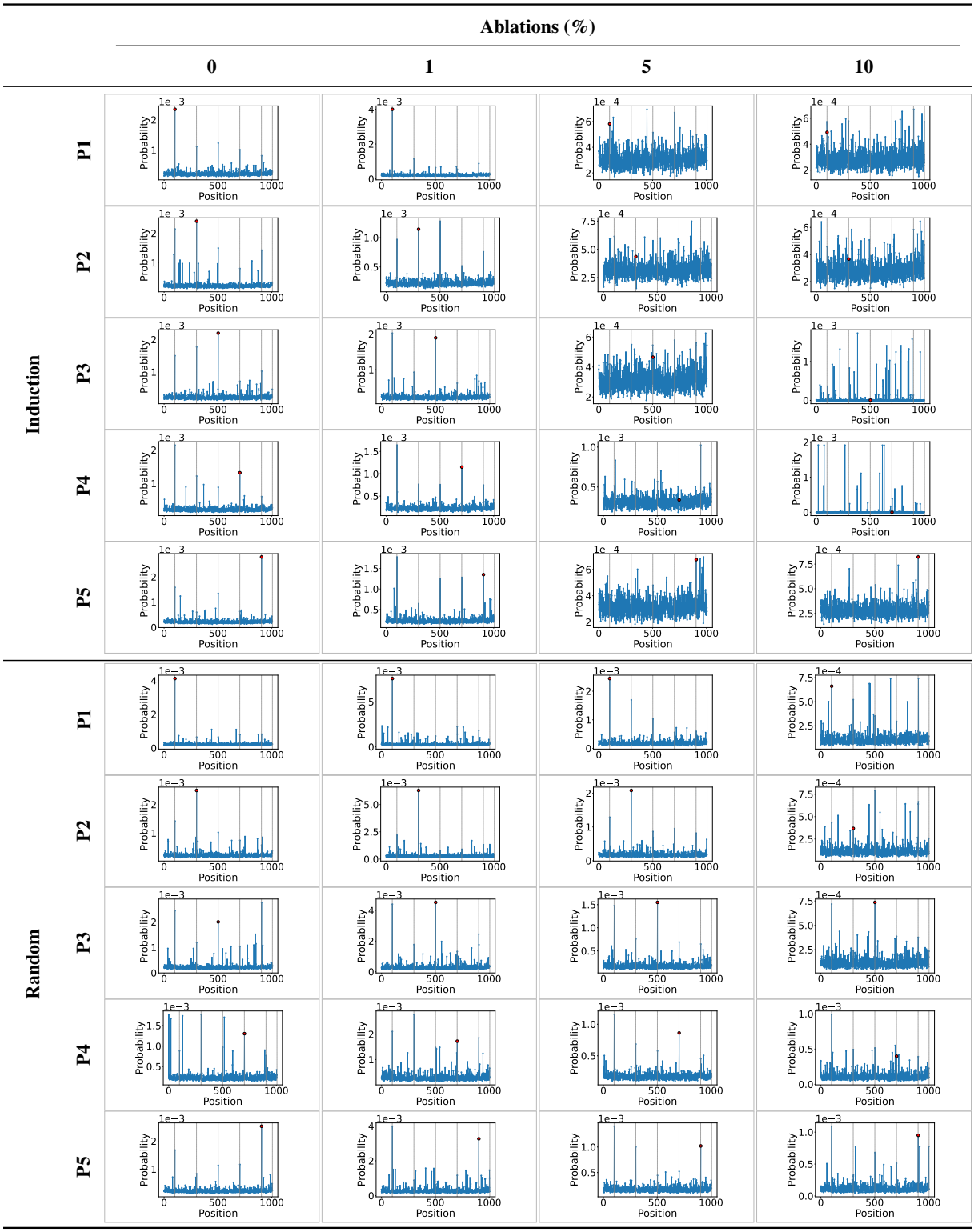

Figure A8: Mamba ablation effect (Exp. 2). Episodic retrieval probability after ablating Induction-related or Random channels. Columns show the percentage of ablated channels.

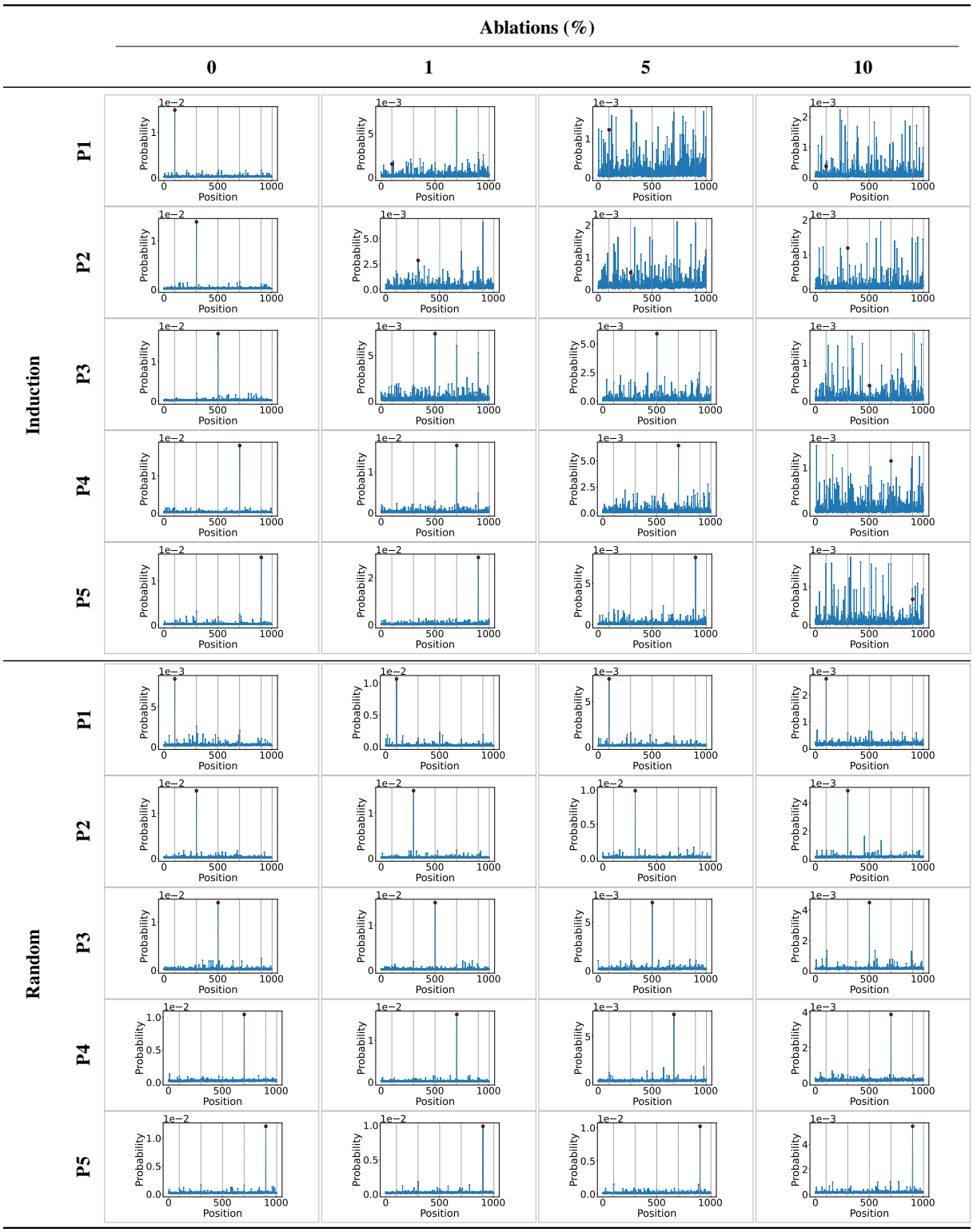

Figure A9: Recurrent-Gemma ablation effect (Exp. 2). Episodic retrieval probability after ablating Induction-related or Random channels. Columns show percentage of ablated channels.