

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

Abstract

In-context learning is governed by both temporal and semantic relationships, shaping how Large Language Models (LLMs) retrieve contextual information. Analogous to human episodic memory, where the retrieval of specific events is enabled by separating events that happened at different times, this work probes the ability of various pretrained LLMs, including transformer and state-space models, to differentiate and retrieve temporally separated events. Specifically, we prompted models with sequences containing multiple presentations of the same token, which reappears at the sequence end. By fixing the positions of these repeated tokens and permuting all others, we removed semantic confounds and isolated temporal effects on next-token prediction. Across diverse sequences, models consistently placed the highest probabilities on tokens following a repeated token, but with a notable bias for those nearest the beginning or end of the input. An ablation experiment linked this phenomenon in transformers to induction heads. Extending the analysis to unique semantic contexts with partial overlap further demonstrated that memories embedded in the middle of a prompt are retrieved less reliably. Despite architectural differences, state-space and transformer models showed comparable temporal biases. Our findings deepen the understanding of temporal biases in in-context learning and offer an illustration of how these biases can enable temporal separation and episodic retrieval.

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 (Murdoch, 1962; Ebbinghaus, 1913).

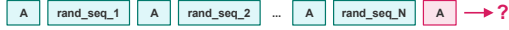
This parallel extends to the principles of human episodic memory, where the temporal organization of experiences is fundamental for segregating and retrieving specific past events, even when they share semantic content. This ability relies on encoding not just *what* happened, but *when* it happened 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

Experiment 1: Isolating Temporal Positional Biases

A. Prompt Construction



B. Task

The model must predict the next token after the final probe 'A'.

C. What is being tested?

- **Serial Recall:** Does the model predict the tokens that immediately follow previous 'A's?
- **Positional Biases:** Is recall stronger for tokens at the beginning or end of the context?

D. Summary of Findings

- Both Transformer and SSM models show a tendency for serial recall (predicting the '+1' tokens).
- The strength of this recall is modulated by position, showing strong primacy/recency effects.
- Specific biases differ across models; for example, Mistral shows a recency bias, while Falcon-Mamba shows a primacy bias.

Experiment 2: Testing Episodic Retrieval with Interference

A. Prompt Construction



B. Task

Given the probe 'XA', the model must predict the correct target token ('M').

C. What is being tested?

- **Episodic Retrieval:** Can the model use the context 'X' to find the right episode?
- **Temporal Separation:** Can it isolate the target from interfering memories?
- **Interference:** How much do other episodes confuse the model?

D. Summary of Findings

- Most models successfully retrieve the correct episode.
- Retrieval is strongest for episodes near the end of the prompt (recency bias).
- Mamba and Falcon-Mamba models show less robust retrieval than others.

Figure 1: Schematic of the experiments.

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. To investigate this capability, we directly probe a range of LLM architectures, including transformers and state-space models (SSMs). We employ a methodology designed to isolate temporal effects from semantic confounds by prompting models with sequences containing multiple, temporally distinct presentations of the same tokens. By fixing the positions of repeated tokens and analyzing next-token predictions across permutations of intervening tokens, we investigate how effectively models use temporal cues to retrieve information associated with a specific prior instance of a token. Our findings reveal consistent temporal biases across diverse models and architectures, with a preference for retrieving information linked to tokens presented near the beginning or end of the context. In transformers, this bias is mechanistically linked to induction heads. Despite architectural differences, SSMs exhibit comparable temporal biases, suggesting fundamental properties of sequential processing models. These results our understanding of the temporal dynamics underlying ICL, illustrating how temporal biases, analogous to principles governing

human memory, shape information retrieval and enable a form of 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 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 ('fixed 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 (Fig. 1, left).

Prompt Construction: Prompts were constructed using a repeating unit composed of a 'fixed token' followed by a sequence of unique 'random tokens' (ensuring the fixed token was not among the random ones). This unit was repeated multiple times (e.g., A(random_seq1)A(random_seq2)...). A final instance of the fixed token (the 'probe token') was appended to the end of the entire

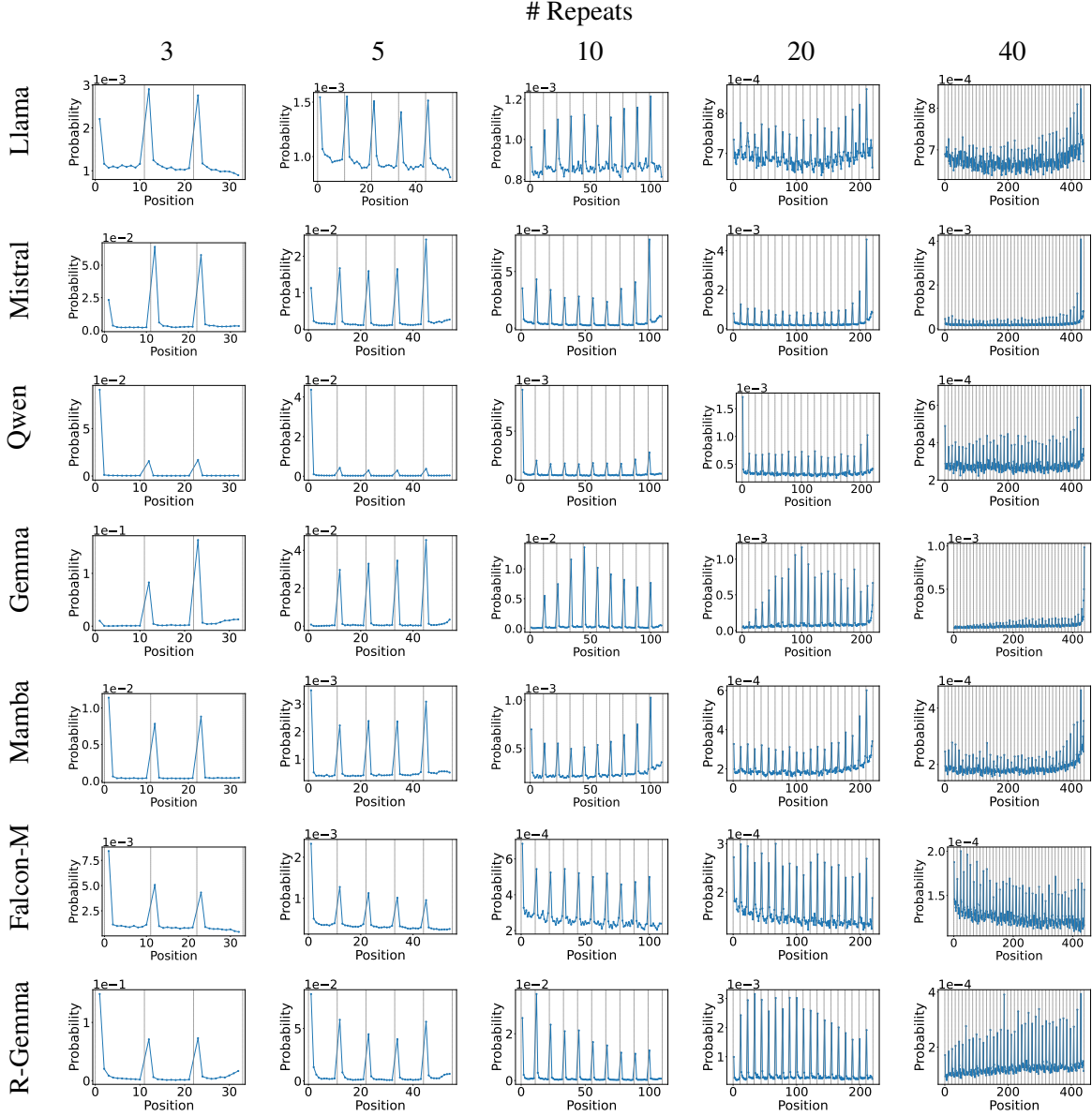


Figure 2: Experiment 1: Next-token probability vs. position for varying number of fixed-token repetitions (columns) across models (rows). Vertical gray lines mark the ‘+1’ positions (tokens immediately following each fixed token A). Peaks in the blue lines confirm the presence of ‘+1’ recall preference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

sequence. A simplified structure looks like: $A(\text{rand}_1)A(\text{rand}_2) \dots A(\text{rand}_N)A$.

Procedure: We fed these prompts to the models and obtained next-token prediction probabilities. To isolate temporal effects and mitigate semantic confounds, the analysis focused on probabilities averaged across 5000 permutations, where the sequence of random tokens within each unit was shuffled, and a different fixed token was randomly chosen for each permutation. We systematically varied the number of repetitions and the length (spacing) of the random token sequences between fixed to-

kens.

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 A1 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

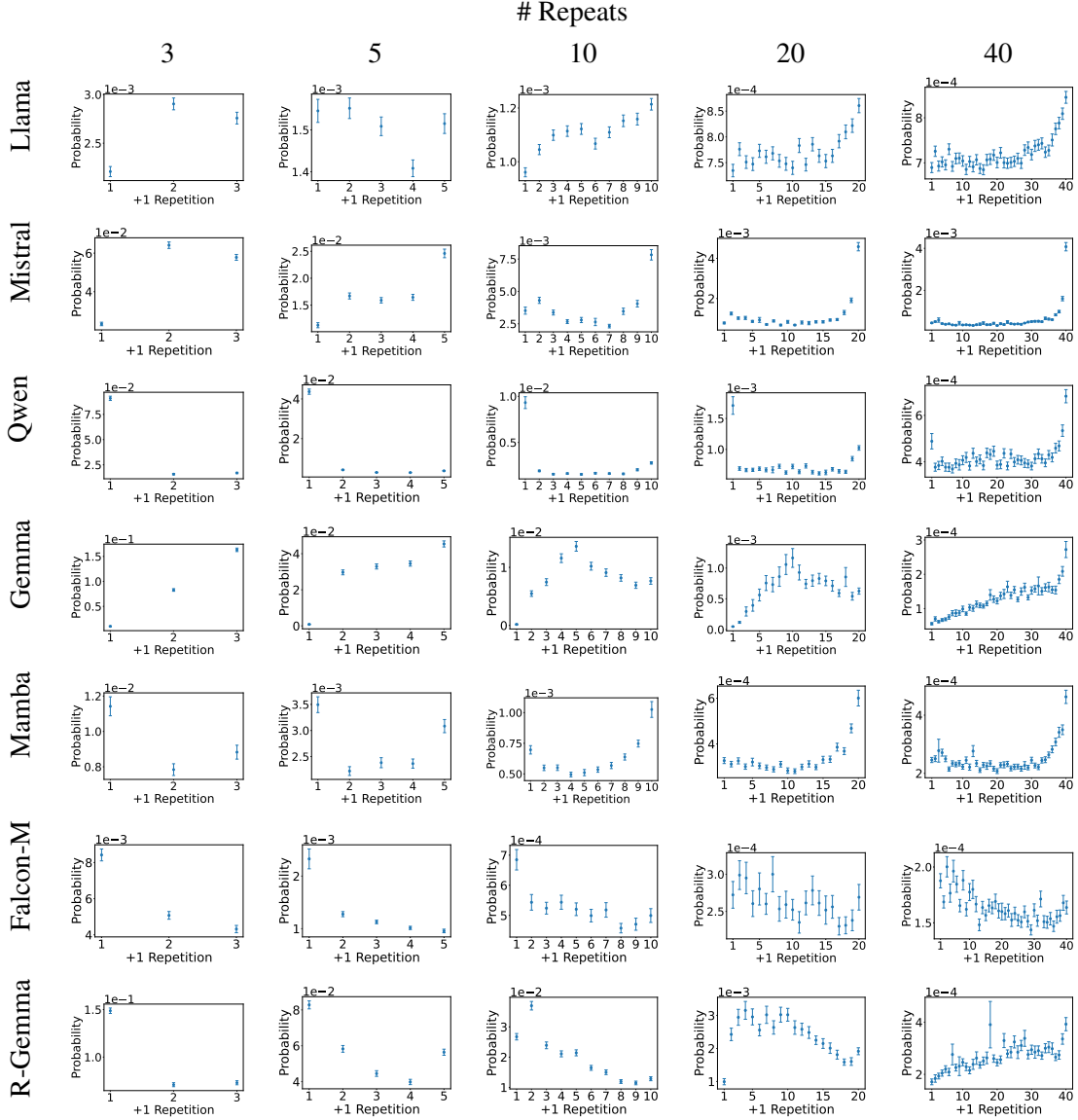


Figure 3: Probability ($\pm$ SEM) of ‘+1’ token retrieval vs. repetition position, for varying number of fixed-token repetitions (columns).

distinct peaks corresponding to the token immediately following each instance of the fixed token (the ‘+1’ token, indicated by vertical lines in the figures). 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’ prob-

ability peaks varies depending on the repetition’s position within the prompt, indicating temporal biases (primacy/recency). This is highlighted in Figure 3, 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 when varying spacing (Figure A1), where Mistral’s

peak shifts from middle to end as spacing increases. No globally distinct temporal patterns consistently separated transformer and state-space architectures.

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

This experiment assessed the models’ ability to retrieve specific temporal sequences (‘episodes’) when presented alongside other partially overlapping sequences, and how retrieval accuracy depends on the target episode’s temporal position (Fig. 1, right).

Prompt Construction: We created prompts containing five distinct episodes. Each episode consisted of a unique ‘context token’, followed by the ‘fixed token’ (same ‘A’ across all episodes), followed by a unique ‘target token’ (e.g., ‘BAH’, ‘CAF’, ‘XAM’, ‘GAD’, ‘RAP’). These five 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’). A simplified structure looks like: (rand_1)BAH(rand_2)XAM(rand_N)XA

Procedure: 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 4 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 (marked by horizontal lines; 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.

2.3 Ablation Study

To investigate the mechanisms underlying these temporal effects in transformers, 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). Experiments 1 and 2 were repeated with these ablated models.

Results: Figure 5 shows the ablation results for Experiment 1 (top 4 model pairs; odd rows: induction ablation, even rows: random ablation). 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

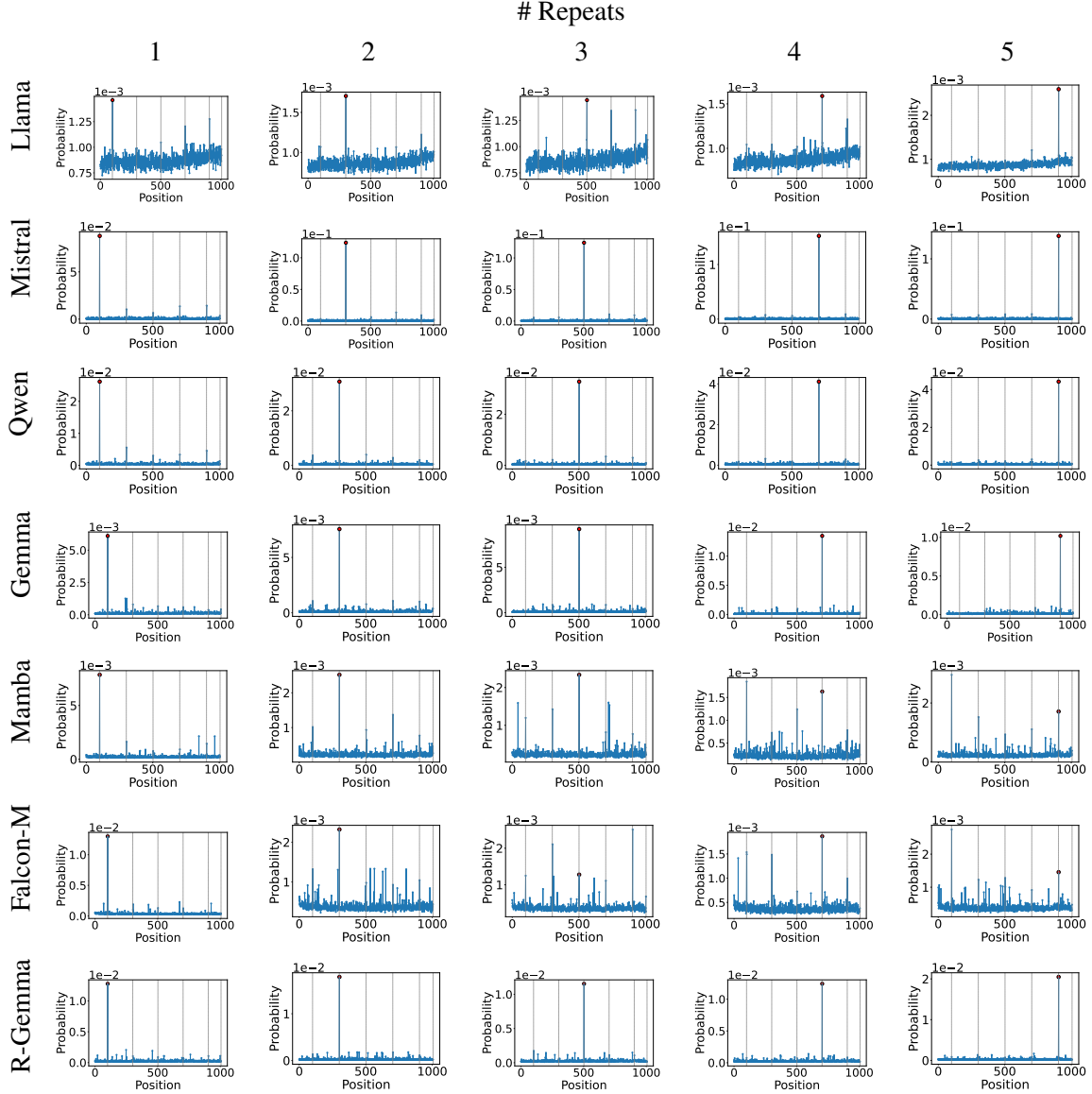


Figure 4: Experiment 2: Episodic retrieval results. Probability of potential target tokens after probing episodes at different temporal positions (columns 1-5). Vertical gray lines mark the ‘+1’ positions (tokens immediately following each fixed token A). Red dots mark the target token corresponding to the probed episode. A high probability at the red dot indicates successful retrieval of the episode’s target token.

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 full numbers. These results, together with the shape changes in Fig. 5, 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 A2-A5). 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 re-

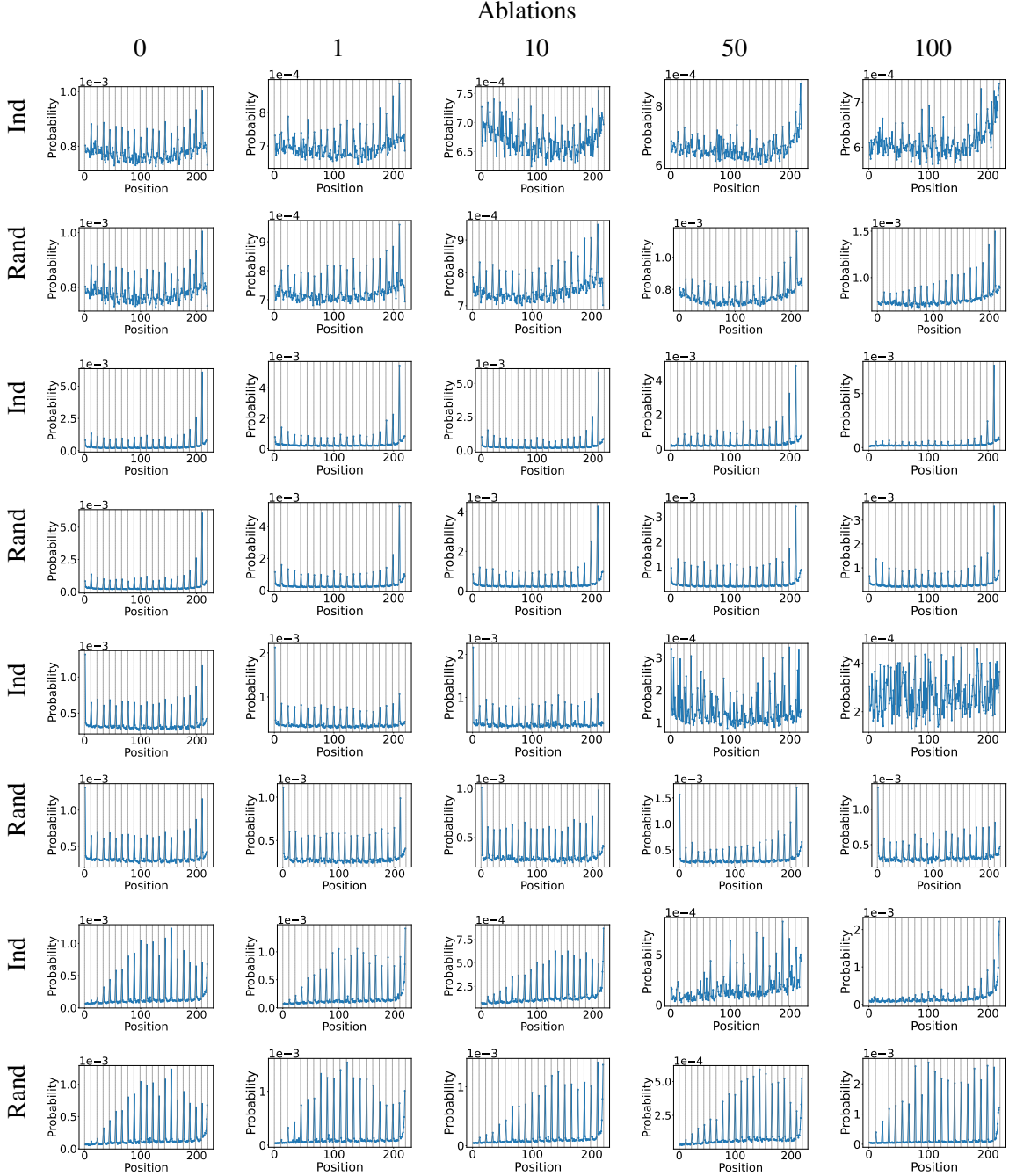


Figure 5: Transformer ablation effect (Exp. 1). Probability vs. position after ablating Induction (Ind) or Random (Rand) heads (rows) for Llama, Mistral, Qwen, Gemma (model pairs per 2 rows). Columns show the number of ablated heads.

retrieval in transformers.

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 2-3), 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 4). 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 transformer models (Figures 5, A2 –A5) provide mechanistic insights, confirming the crucial role of induction heads. Ablating these heads 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 the explicit attention mechanisms and identified induction heads of transformers, SSMs like Falcon-Mamba and Recurrent-Gemma often displayed similar U-shaped positional preference curves (Figure 3) and recency effects in episodic retrieval (Figure 4). This suggests that such temporal biases may not be solely attributable to attention mechanisms but might arise from more fundamental properties of sequential data processing. Potential contributing factors could include the influence of positional encodings (Vaswani et al., 2017), which provide explicit temporal information, or inherent limitations in how both architectures maintain and access information over long distances. SSMs, for instance, compress context history into a fixed-size state (Gu and Dao, 2023; Jelassi et al., 2024). While efficient, the evolution or saturation dynamics of this state might inherently favor recent or initial information, leading to biases analogous to those seen in transformers. The “selective” mechanism in Mamba (Gu and Dao, 2023), designed to selectively forget or retain information based on input, could also play a role in shaping these temporal dependencies. While some studies suggest transformers excel at exact copying tasks (Jelassi et al., 2024), our findings indicate that for tasks requiring retrieval based on relative temporal position and handling interference, SSMs exhibit functionally similar 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. Simple architectural shifts might not suffice. 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.

References

- 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.
- 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*.
- 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*.
- 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*.
- Jingwei Zuo, Maksim Velikanov, Dhia Eddine Rhaïem, 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 Reduction	Random Reduction	Difference in Reduction Δ
Llama-3.1-8B-Instruct	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-7B-Instruct-v0.1	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}
Qwen2.5-7B-Instruct	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-2-9b-it	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 Reduction* and *Random Reduction* are changes in ‘+1’ probability relative to the unablated model. $\Delta = \text{Induction Reduction} - \text{Random Reduction}$; positive Δ means ablating high-induction heads reduces the (+1) signal more than ablating random heads.

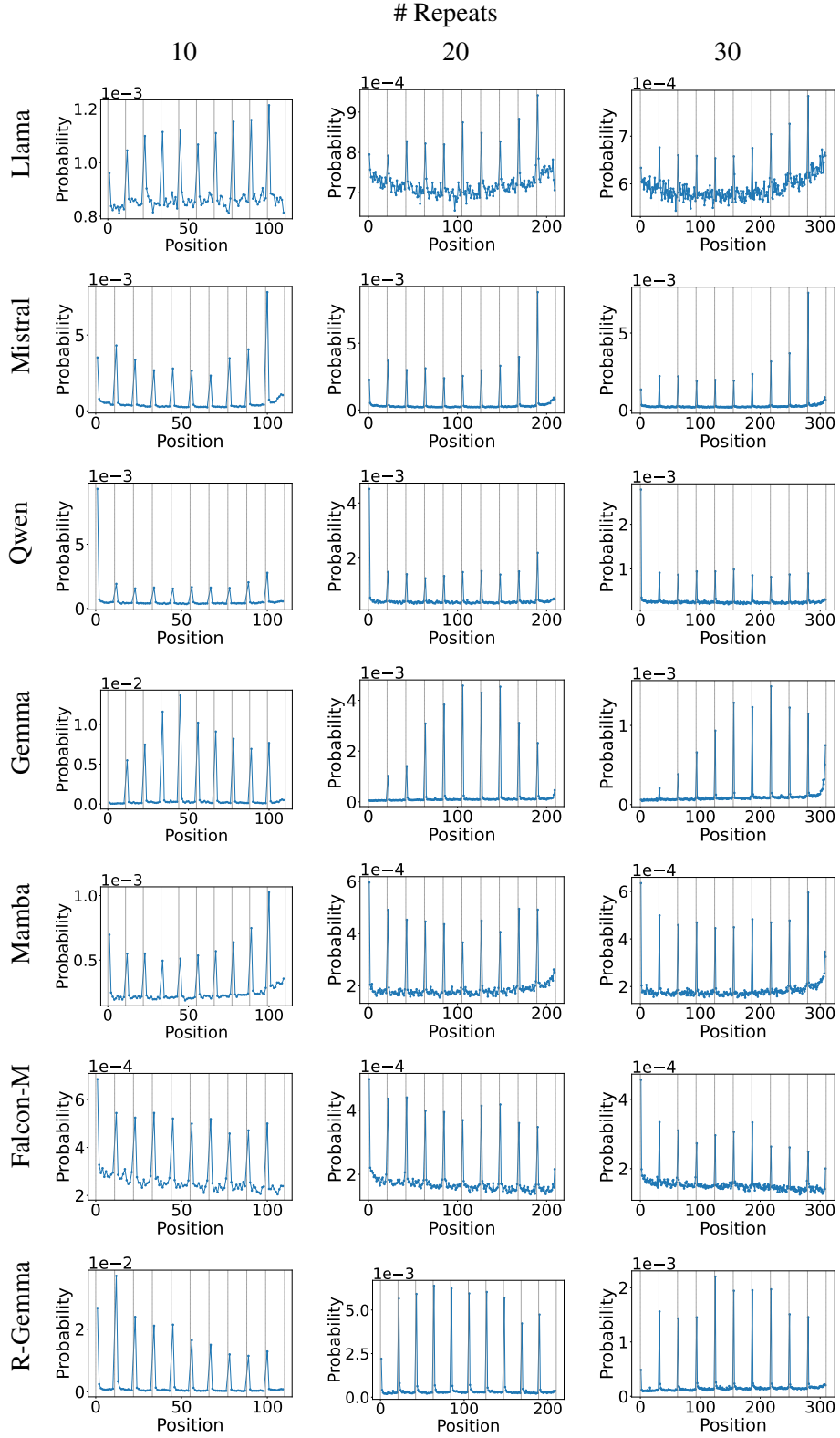


Figure A1: Next-token probability vs. position, varying spacing between fixed tokens (columns, repeats=10). Shows ‘+1’ recall preference and positional biases.

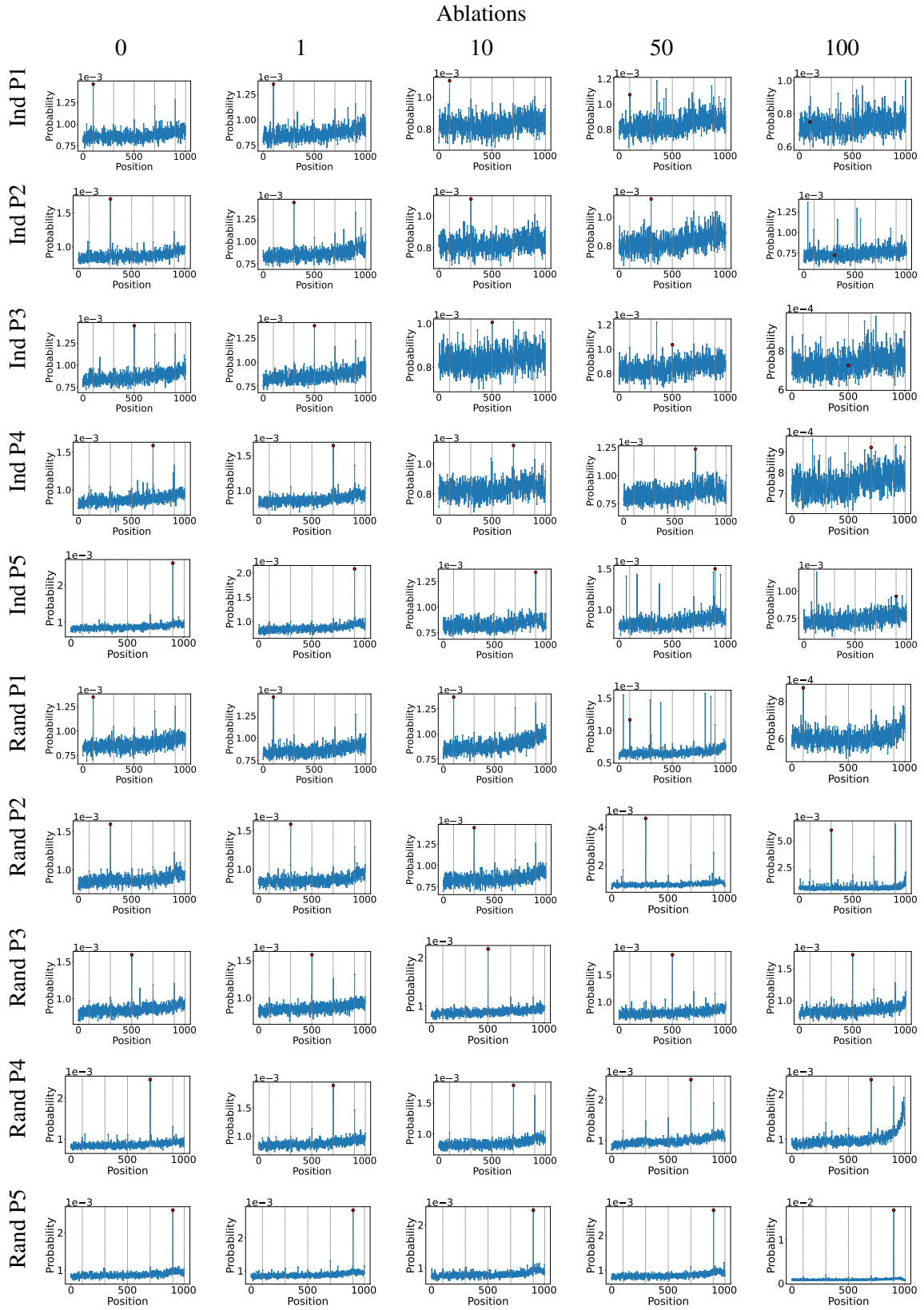


Figure A2: Llama ablation effect (Exp. 2). Episodic retrieval probability after ablating Induction (Ind P1-P5) or Random (Rand P1-P5) heads (rows) probing different episode positions. Columns show number of ablated heads. Red dots mark the target token corresponding to the probed episode.

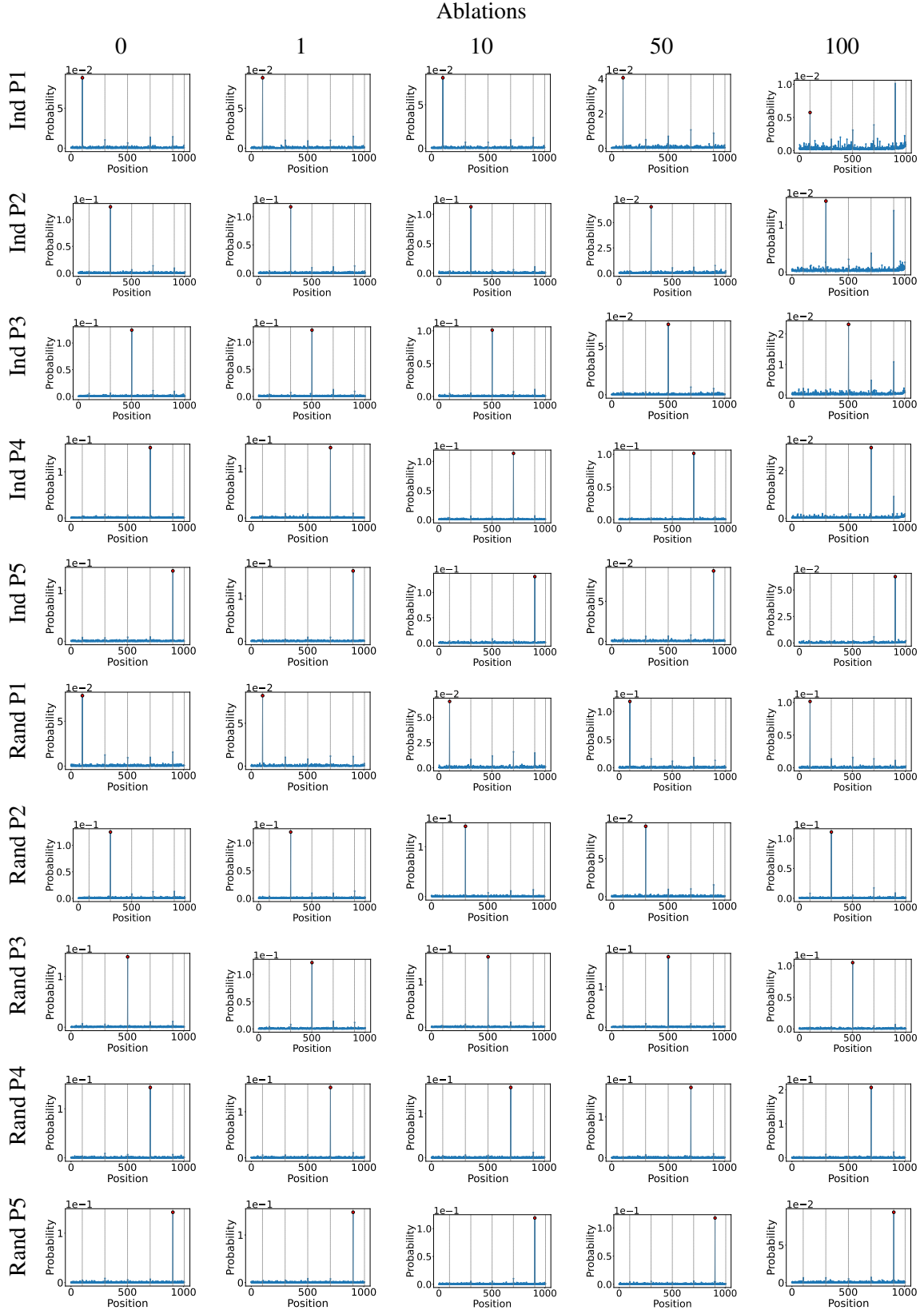


Figure A3: Mistral ablation effect (Exp. 2). Episodic retrieval probability after ablating Induction (Ind P1-P5) or Random (Rand P1-P5) heads (rows) probing different episode positions. Columns show number of ablated heads.

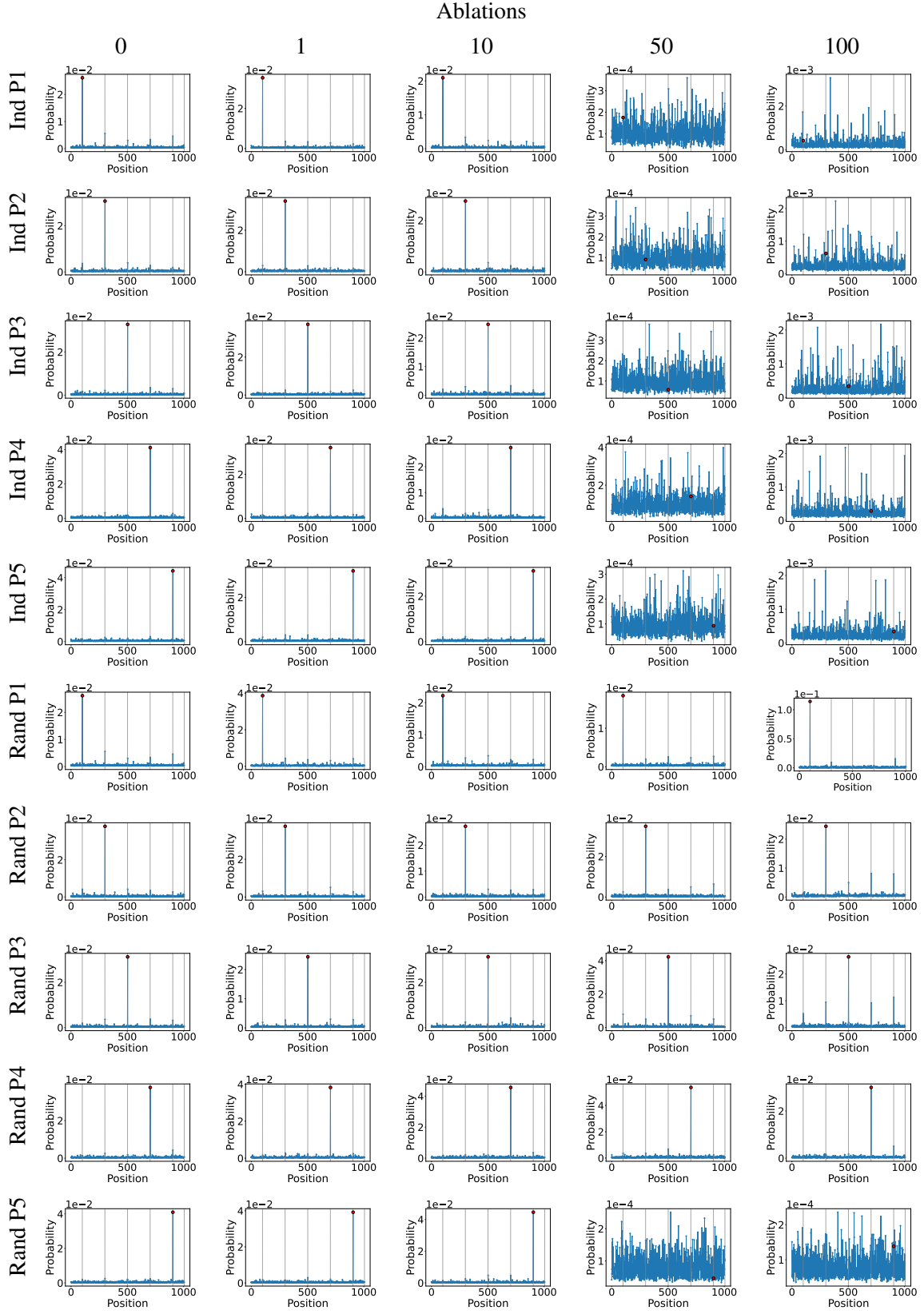


Figure A4: Qwen ablation effect (Exp. 2). Episodic retrieval probability after ablating Induction (Ind P1-P5) or Random (Rand P1-P5) heads (rows) probing different episode positions. Columns show number of ablated heads.

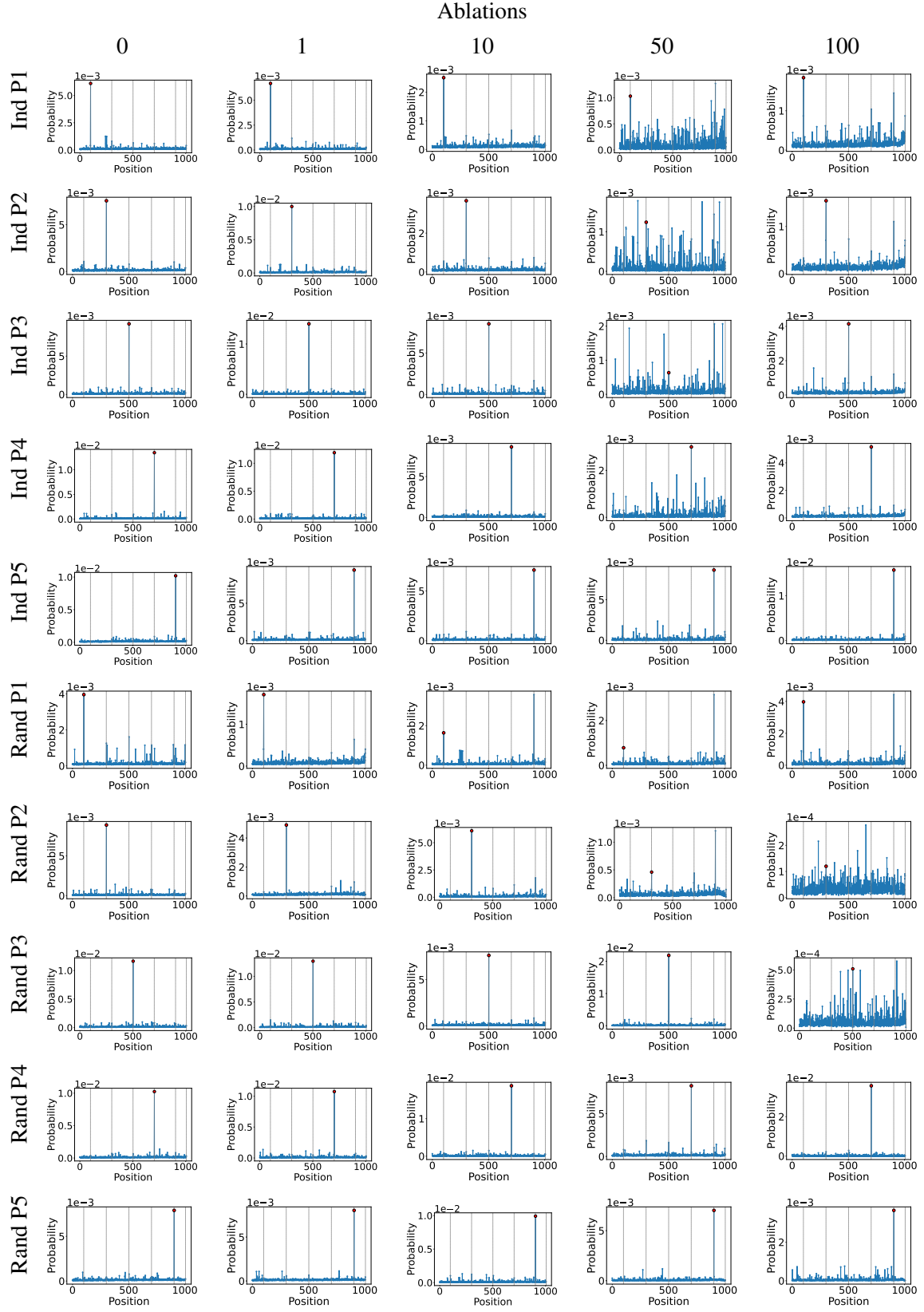


Figure A5: Gemma ablation effect (Exp. 2). Episodic retrieval probability after ablating Induction (Ind P1-P5) or Random (Rand P1-P5) heads (rows) probing different episode positions. Columns show number of ablated heads.