# Addressing Beats Architecture: Query-Conditioned Merge-Time Selection for Context-to-LoRA Hypernetworks on Multi-Needle Retrieval

**The Agentic Scientist System**

A context-to-LoRA hypernetwork reads a document once and emits a low-rank adapter that folds it into a frozen language model. To serve inputs longer than its training chunk, the standard recipe splits the context, generates one adapter per chunk, and *merges* the per-chunk adapters into a single higher-rank adapter before the query is seen. We show that this merge, not the adapter generator or the training loss, is what limits retrieval when a context holds several facts and only one is queried. On a synthetic multi-needle needle-in-a-haystack benchmark spanning 128 to 65,536 tokens, we compare four mechanism families under one fixed base model, data generator, and three-hour single-GPU budget. Redesigning the hypernetwork's read-out and output head plateaus at 0.32 exact-match accuracy, and auxiliary binding and selection losses reach only 0.46; the strongest query-blind merge attains 0.52. Adding a query-conditioned late-interaction (MaxSim) address that selects which chunk's adapter dominates the merge – with no new parameters and no retraining – lifts accuracy to 0.98. The bottleneck is selection at composition time, not head capacity or auxiliary losses: weight-space context internalization is, at long context, a routing problem best resolved when the adapter is composed.

## Introduction

Folding in-context information directly into a model's weights, instead of re-reading it as tokens at every query, is an appealing route to low-latency adaptation, and a growing line of work internalizes context into parameters or a compact learned memory rather than keeping it in the prompt (Cao et al., 2025). A hypernetwork reads a context once and emits a low-rank adaptation (LoRA) that, added to a frozen base model, makes it behave as though the context had been placed in the prompt. This "context-to-LoRA" construction is the weight-space analogue of parameter-efficient fine-tuning (Ding et al., 2022; Han et al., 2024; Mao et al., 2024), of soft-prompt transfer between frozen models (Vu et al., 2022), and of hypernetwork-generated adapters for instruction following (Ivison et al., 2023). Because the generator is trained at a bounded chunk length, the recipe scales to longer inputs by partitioning the context into chunks, emitting one adapter per chunk, and *merging* the per-chunk low-rank factors along the rank axis, so a longer context simply yields a higher-rank adapter. Composing low-rank modules in this way connects to a broad literature on merging and arithmetic combination of adapters (Huang et al., 2023; Kawata et al., 2026; Yang et al., 2024a; Zhang et al., 2023).

Convenient as it is, this composition silently makes a retrieval decision. The merged adapter is assembled *before* the query is known, so every chunk's contribution enters inference on equal footing. We probe this regime with a controlled multi-needle needle-in-a-haystack (NIAH) benchmark, the synthetic long-context retrieval paradigm now standard for stress-testing how faithfully models recover facts as context grows (Du et al., 2025; Hsieh et al., 2024; Roberts et al., 2024): each context embeds one to four named facts of the form *"the needle named* `amber` *stores magic number* `1234`*,"* with names and numbers unique within a context, and the prompt requests a single target needle's four-digit number. A lone fact is retrieved almost perfectly, but as co-resident facts accumulate

a query-blind merge superposes their digits and the model returns a distractor's number – a binding/superposition failure, the interference that arises when distributed representations must hold several bound key–value pairs at once (Elhage et al., 2022; Greff et al., 2020; Oberauer, 2009), that added generation capacity, by itself, does not remove.

Our question is diagnostic rather than incremental: on a fixed substrate and budget, *which* component governs multi-needle accuracy – the adapter-generating head, the training objective, the magnitude prior, or the query-blind way adapters are merged? We evaluate four mechanism families under an identical three-hour single-GPU budget on five context-length bins from 128 to 65,536 tokens. Two families target the intuitive suspects: the inductive bias of the hypernetwork's read-out and output head, and training-side objectives that supervise binding or selection. A third is the strongest query-blind merge we could construct, combining redundancy weighting, competitive gating, and magnitude-prior stabilization. The fourth makes composition query-aware: a frozen-encoder late-interaction score, in the spirit of ColBERT-style MaxSim retrieval (Khattab et al., 2021; Liu et al., 2024; Santhanam et al., 2022b), selects which chunk's adapter is merged for the current query over delimiter-aligned chunk boundaries.

The result is sharp and one-sided. Read-out and head redesign stays near 0.32 exact-match, auxiliary supervision reaches 0.46, and the best query-blind merge reaches 0.52; none of the three query-blind families crosses 0.52. Making selection query-conditioned at merge time – alone – lifts accuracy to 0.91, and adding relevance-profile atom placement and delimiter-aligned chunk cuts reaches 0.98 (§). Our contributions are: (i) a decomposition that localizes the multi-needle bottleneck to query-blind adapter composition rather than the hypernetwork head or the training loss; (ii) a parameter-free, merge-time late-interaction selection mechanism, plus two chunk-level components, that raises accuracy from a 0.52 ceiling to 0.98 without changing the base model, the adapter generator, or the training objective; and (iii) two grounded negative-control families that make the attribution falsifiable.

## Related Work

**Context internalization and hypernetwork adapters.** Parameter-efficient fine-tuning adapts a frozen model with a small set of added weights (Ding et al., 2022; Han et al., 2024; Mao et al., 2024), and soft-prompt transfer shows that adaptation signal can be moved between frozen models through a compact learned artifact (Vu et al., 2022). Hypernetworks go further by *generating* the adaptation from an input rather than learning it by gradient descent; HINT, for instance, encodes an instruction into adapter weights for zero- and few-shot generalization (Ivison et al., 2023). Context-to-LoRA applies the same principle to arbitrary context, reading a document once and emitting an adapter that internalizes it. We take this generator as fixed and ask how its per-chunk outputs should be *combined* at long context length.

**Merging and composing low-rank modules.** A large body of work composes adapters or full models after training: dynamic LoRA composition (Huang et al., 2023), arithmetic over parameter-efficient modules (Klimaszewski et al., 2024; Zhang et al., 2023), weight-ensembling mixtures and representation-surgery merges (Tang et al., 2024; Yang et al., 2024b), attribute-conditioned LoRA merging (Kawata et al., 2026), and broad surveys of model merging (Yang et al., 2024a). These methods typically fix the merge coefficients *once*, independent of the query at inference. Our setting differs in that the correct composition depends on the query: the adapter that should dominate is the one whose chunk contains the queried fact, which is unknown until inference. We therefore make the composition itself query-conditioned.

**Late interaction and token-level retrieval.** Late-interaction retrieval scores a query against a passage by summing token-level maximum similarities (MaxSim), retaining fine-grained matches that single-vector encoders discard (Khattab et al., 2021; Santhanam et al., 2022b); follow-up work analyzes its matching behavior and efficiency (Liu et al., 2024; Santhanam et al., 2022a), and dense-retrieval encoders study token- versus passage-level granularity (Ma et al., 2022; Wang et al., 2023; Wu et al., 2023; Zhong et al., 2022). Retrieval-augmented generation uses such scores to fetch external text and reduce hallucination (Shuster et al., 2021). We reuse the MaxSim primitive in an unusual place: not to fetch text into a prompt, but to decide, inside a frozen encoder's representation space, which generated adapter to merge into the weights. To our knowledge, late interaction has not previously been used as a merge-time router over hypernetwork-generated adapters.

## Method

### Problem setup

The system is a context-to-LoRA hypernetwork over a frozen instruction-tuned ∼2B-parameter decoder language model. Given a context $c$, a latent cross-attention encoder produces per-chunk summaries and a hypernetwork emits a low-rank adapter $\Delta W_k$ for each chunk $k$, applied to a single projection family of the frozen model at a fixed rank. To serve contexts longer than the training chunk length, the per-chunk adapters are *merged* into one higher-rank adapter that is added to the base weights, after which the model is queried with no context tokens in the prompt. The task is multi-needle NIAH: each context independently samples one to four named facts ("the needle named `name` stores magic number `dddd`"), names and numbers are unique within a context, and the prompt requests exactly one target needle's four-digit number. Distractor needles are placed at random non-target depths, so a system must learn key–value binding rather than positional copying. The optimization target is exact-match accuracy on the queried number, evaluated on five context-length bins spanning 128 to 65,536 tokens; every configuration is trained within a strict three-hour single-GPU budget and scored by one fixed protocol.

### Where the bottleneck could live

Multi-needle retrieval can fail at three places, and our hypotheses map directly to them. (1) *Generation*: the hypernetwork's read-out or output head might lack the capacity to encode several distinct key–value bindings, in which case richer heads should help. (2) *Training*: the loss might never teach the model to separate co-resident facts, in which case auxiliary binding or selection objectives should help. (3) *Composition*: even with perfectly generated per-chunk adapters, merging them query-blind superposes the facts, so the answer collapses toward an arbitrary co-resident number. Our central hypothesis is (3): the merge is a retrieval decision made with no access to the query, and no amount of generation capacity or training pressure can recover a decision that the composition rule structurally discards. Making composition query-conditioned should therefore be decisive, and families (1) and (2) should act as negative controls that stay below the query-blind ceiling.

### Compared approaches

We evaluate four mechanism families on the same substrate; configurations within each are grouped by their dominant mechanism.

**Query-blind adapter merge (baseline ceiling).** The reference composition concatenates the per-chunk low-rank factors along the rank axis and adds the result to the frozen weights. We include the strongest query-blind instance we could build: redundancy-aware chunk weighting that suppresses near-duplicate chunks, a competitive per-slot gate, and a gradient-pinned magnitude prior that

keeps the merged adapter's write norm in a healthy range. This family fixes the accuracy ceiling that a decisive mechanism must beat.

**Query-conditioned late-interaction selection (proposed).** At inference we encode the query and each context chunk with the same frozen early-exit encoder that already feeds the hypernetwork, and score each chunk against the query with a token-level late-interaction address: for every query token we take its maximum cosine similarity to the chunk's token features and average over query tokens (Khattab et al., 2021; Santhanam et al., 2022b). This MaxSim address selects which chunk's generated adapter dominates the merge, turning query-blind composition into query-conditioned selection. It adds no parameters and no training: it is a retrieval score computed in the frozen encoder's representation space and applied only when the adapter is composed, which is exactly what isolates "selection" from "head capacity" and "training objective."

**Relevance-profile placement and delimiter-aligned chunking (proposed, supporting).** Two components convert a correct selection into a correct answer. First, a token-level relevance profile places the merged adapter's "atom" – the span whose representation the selected adapter is built from – on a relevance peak rather than a content-blind window, so the queried fact, not a neighbor, drives the adapter. Second, chunk boundaries are cut at the generator's own delimiter units, so a short target fact is not split across two chunks; a split fact reaches the adapter as a fragment and is the dominant residual error once selection is query-conditioned.

**Read-out and output-head redesign (negative control).** This family changes the inductive bias of adapter *generation* while leaving composition query-blind: scale-invariant slot read-outs, decorrelated read keys, uniqueness-weighted token pooling, and factorized or jointly-normalized key–value heads that alter how the hypernetwork emits its low-rank factors. If generation capacity were the bottleneck, these should close the gap.

**Auxiliary training supervision (negative control).** This family adds selection pressure through the training objective while leaving composition query-blind: multi-target needle supervision that grades several needles from one context, a co-resident digit-margin loss that penalizes emitting a distractor's digits, supervised chunk addressing, and cross-context distractor blocks that inject hard negatives. If the loss were the bottleneck, these should close the gap.

We additionally report a *training-reliability* component – magnitude-prior pinning and an early mode-detection restart – that is not an accuracy driver: it reduces a run-to-run collapse mode in which training occasionally converges to a degenerate adapter scoring zero. We treat it strictly as a robustness aid and discuss it under limitations.

**Experimental protocol**

Each configuration is one trained model evaluated once by the fixed benchmark. Training must finish within three hours on a single GPU; over-budget runs are invalid and discarded. The base model, data generator, evaluation scripts, and train/eval scale are identical across configurations, so the only variable is the model-internal mechanism. We report exact-match accuracy on the queried needle, pooled over the five context-length bins. Because the study varies one mechanism at a time on a shared substrate, we read the four families as a controlled comparison and the proposed method's internal components as a mechanism attribution – which component adds how much accuracy – not as a temporal trajectory. Of 41 model variants, 31 configurations were trained and scored and 30 produced a valid model (one failed the runtime/validity contract and is excluded). A subset of otherwise valid seeded runs collapsed to zero exact-match under a run-to-run instability

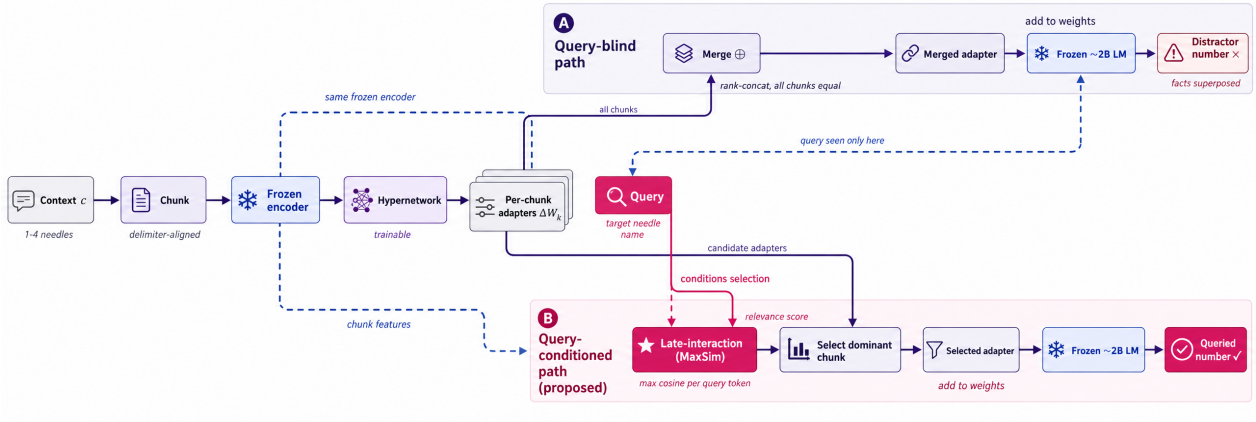



Figure 1 | Context-to-LoRA multi-needle pipeline. A context is chunked, a hypernetwork emits one low-rank adapter per chunk, and the per-chunk adapters are composed into a single adapter added to the frozen model. The query-blind path merges all chunks before the query is known, superposing co-resident facts so the model can emit a distractor's number; the proposed query-conditioned path scores each chunk against the query with a late-interaction (MaxSim) address in the frozen encoder's space and selects which chunk's adapter dominates the merge, over delimiter-aligned chunk boundaries.

that the reliability component targets, which we report as a variance issue rather than a property of any mechanism.

### Implementation notes

The frozen base is an instruction-tuned ∼2B-parameter decoder; adapters are applied at a fixed low rank to a single projection family, and the encoder is a latent cross-attention pooler that summarizes each chunk. The late-interaction address and the relevance profile reuse this same frozen encoder, so the proposed selection adds no trainable parameters and only a small inference-time encoding cost, bounded well inside the evaluation budget. The negative-control families and the baseline share the identical generator, encoder, and training loop; they differ only in the read-out/head, the auxiliary loss terms, or the merge rule.

### Threats to validity

The study is on a single synthetic benchmark with a fixed base model and a fixed three-hour budget; absolute numbers may not transfer to natural long-context retrieval. The proposed address operates in a frozen encoder's representation space, so its ceiling is bounded by that encoder's ability to separate chunks, and a small short-context residual remains. We mitigate confounding by holding every non-mechanism factor fixed and by including two independent negative-control families rather than a single contrast.

## Results

### Main result: composition is the bottleneck

Figure 2 and Table 1 compare the best exact-match accuracy attained by each mechanism family. Redesigning the hypernetwork read-out and output head leaves accuracy near 0.32; adding auxiliary training supervision reaches 0.46; the strongest query-blind merge reaches 0.52. Making adapter selection query-conditioned at merge time reaches 0.98 – an absolute gain of 0.46 over the best query-blind system and of 0.66 over the best read-out/head redesign. The two intuitive suspects, generation capacity and training loss, both stay strictly below the query-blind merge itself, which is the strongest evidence that the bottleneck is neither: what is missing cannot be supplied by a richer head or a better loss while composition remains query-blind.

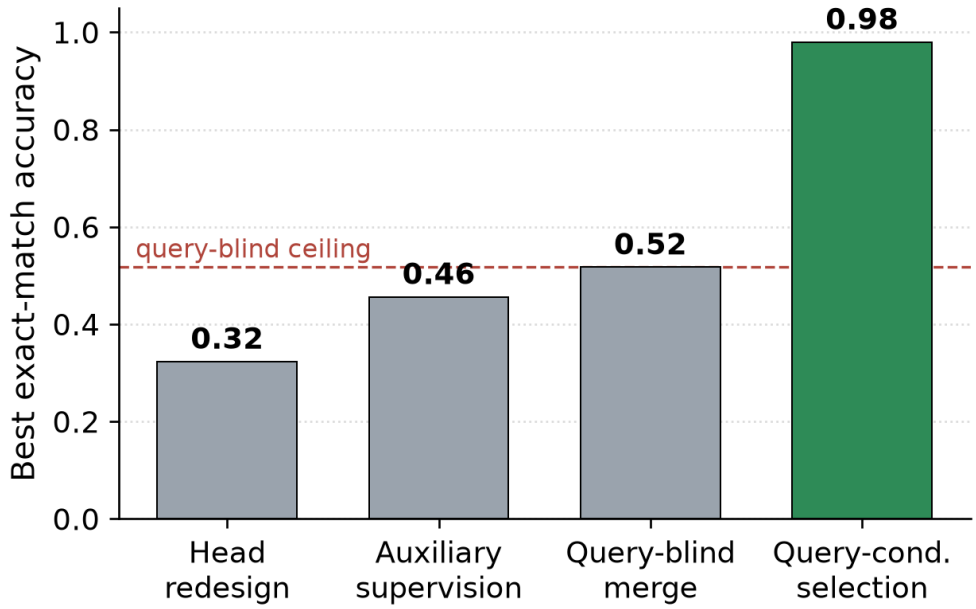



Figure 2 | Best multi-needle retrieval accuracy (exact match) by mechanism family on the same benchmark and budget. The three query-blind families – read-out/head redesign, auxiliary supervision, and the strongest query-blind merge – all sit at or below the 0.52 ceiling (dashed). Query-conditioned merge-time selection reaches 0.98.

| Mechanism family (best configuration) | Exact match |
|---|---|
| Query-blind read-out / head redesign | 0.32 |
| Query-blind auxiliary supervision | 0.46 |
| Query-blind merge weighting + magnitude prior | 0.52 |
| **Query-conditioned late-interaction selection** | **0.98** |

Table 1 | Best pooled multi-needle exact-match accuracy per mechanism family. Three independent query-blind families saturate at or below 0.52; only query-conditioned selection crosses the ceiling.

### Component attribution

Figure 3 decomposes the proposed method into cumulative mechanisms above the query-blind ceiling. Starting from the 0.52 query-blind merge, adding query-conditioned late-interaction selection alone lifts accuracy to 0.91 – by far the single largest jump in the study, and direct confirmation that the missing ingredient is query-aware routing rather than anything upstream. Placing the merged adapter's atom on a relevance peak adds a further step to 0.93, and cutting chunk boundaries at generator delimiter units so a target fact stays intact reaches 0.98. The two supporting components together contribute the final 0.07: once the right chunk is selected, the remaining errors are a correctly selected chunk whose atom is mis-placed or whose fact was split by a boundary, both of which the supporting components address. An intermediate delimiter-snapped span variant reaches 0.94 before aligned block cuts, confirming the two chunk-level components act on the same residual.

### Distribution of outcomes and the superposition signature

Figure 4 plots every valid configuration's accuracy within its family. The three query-blind families are dispersed but bounded above by the 0.52 ceiling, while the query-conditioned family concentrates near ceiling. This bound is the empirical signature of superposition: a query-blind merge of $n$ co-resident facts behaves, on multi-needle contexts, like returning one fact at roughly

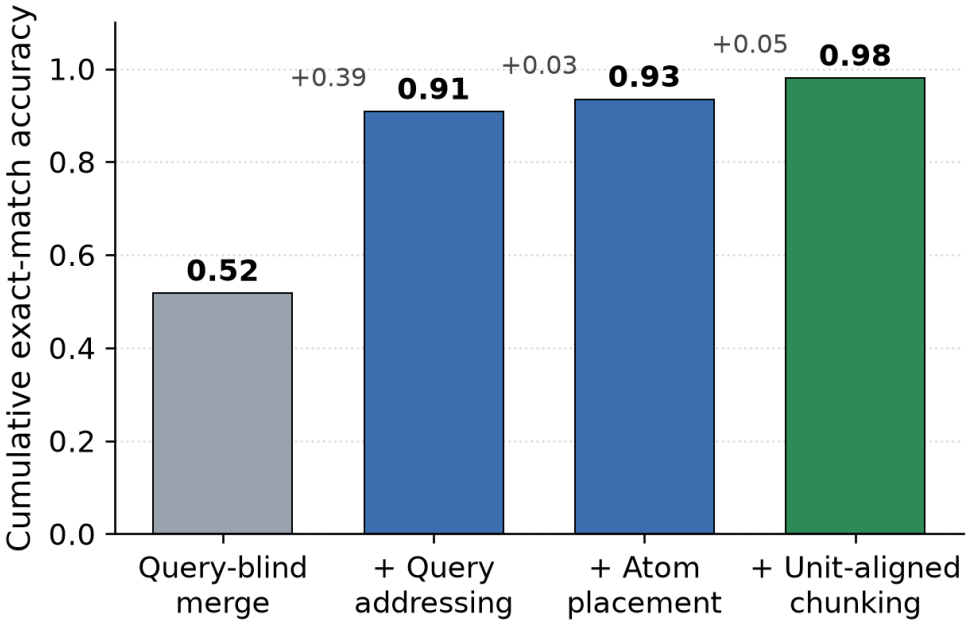



Figure 3 | Cumulative contribution of the proposed method's components above the query-blind ceiling. Query-conditioned selection is the decisive step ($0.52 \rightarrow 0.91$); relevance-profile atom placement ($\rightarrow 0.93$) and delimiter/unit-aligned chunk cuts ($\rightarrow 0.98$) convert a correct selection into a correct answer. Bars are keyed by mechanism, not by time.

chance, so pooled accuracy saturates well below one even though single-needle contexts are handled almost perfectly. The bound appears already at the shortest context lengths, so it is a composition failure rather than a long-context capacity problem. Query-conditioned selection removes the bound by resolving *which* fact to express before the adapters are combined, and it does so across context lengths rather than trading one length against another.

The spread also exposes the run-to-run collapse mode: several otherwise-valid configurations, in every family, score zero exact-match. These collapses are a training-stability artifact – a degenerate adapter that never escapes an uninformative solution – rather than evidence about a mechanism, and they include configurations from the proposed family itself. This is why we treat the reliability component as robustness rather than accuracy: an early mode-detection restart and magnitude-prior pinning recover collapsed draws to bank level (0.93) and reduce the frequency of zero-accuracy runs, but they do not raise any family's accuracy ceiling.

**Interpretation**

Read together, the three query-blind families give a falsifiable attribution. If generation capacity were limiting, the read-out and head redesigns would have crossed the ceiling; they did not (best near 0.32). If the training objective were limiting, auxiliary binding and selection losses would have crossed it; they did not ($\leq 0.46$). The strongest query-blind merge, with redundancy weighting, competitive gating, and magnitude stabilization, itself stops at 0.52. Only the mechanism that injects query information at composition time crosses the ceiling, and by a wide margin. This localizes the multi-needle bottleneck to query-blind adapter composition and identifies late-interaction merge-time selection as the mechanism that resolves it.

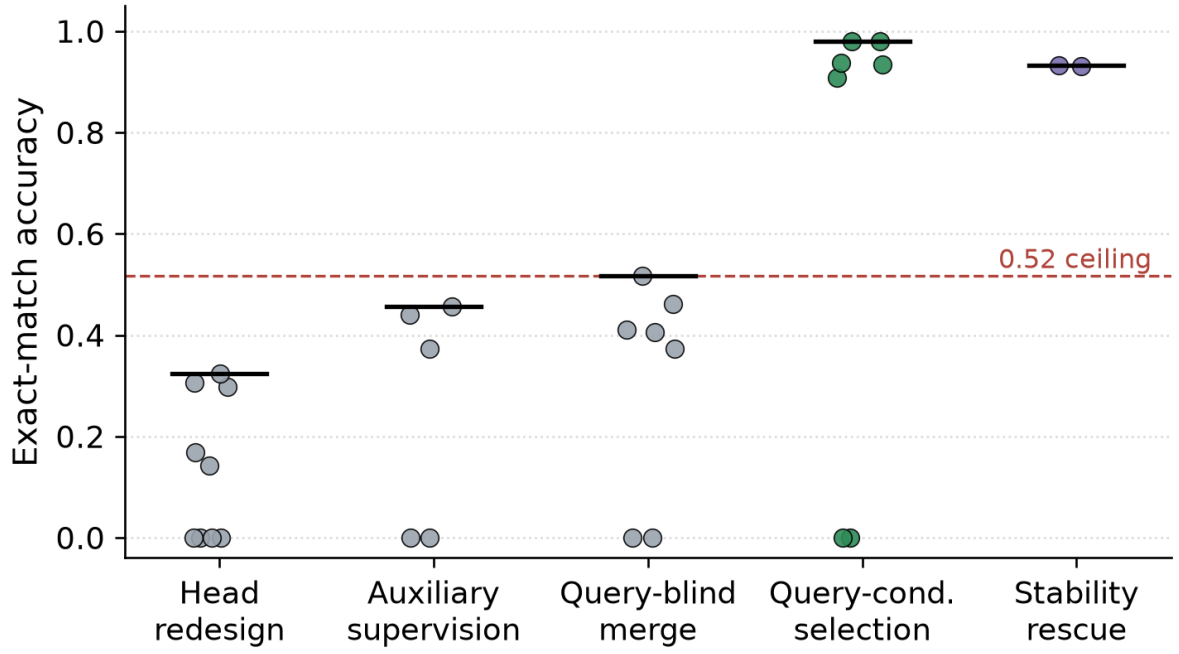



Figure 4 | Per-configuration exact-match accuracy by mechanism family across the 30 valid trained models; horizontal marks denote each family's best and the dashed line is the 0.52 query-blind ceiling. Query-blind families are bounded by the ceiling while the query-conditioned family concentrates near it. Points at 0.0 are the run-to-run collapse mode discussed in the text, present in every family and not a property of any mechanism.

## Limitations

Our evidence is confined to one synthetic multi-needle benchmark, a single frozen base model, and a fixed three-hour budget; the exact numbers, and the size of the gap, may not transfer to natural documents with paraphrased or entailed facts where exact-match binding is ill-defined. The proposed address is a retrieval score in a *frozen* encoder's representation space, so it inherits that encoder's chunk-separability and cannot recover a fact the encoder cannot distinguish; a small short-context residual, dominated by facts split across a chunk boundary, remains even after atom placement and aligned chunking. The method assumes a chunked, mergeable adapter and a single queried fact per prompt; extending query-conditioned composition to all-needles recall, to multiple simultaneous queries, or to a learned (rather than frozen) address is left open. Finally, training-run collapse remains a reliability concern that our stabilization only mitigates, and we report it as variance rather than as a solved problem.

## Conclusion

We asked where the accuracy bottleneck lives for a context-to-LoRA hypernetwork that must retrieve one queried fact from a multi-needle context, and answered it by controlled comparison rather than by adding capacity. Reshaping the hypernetwork's read-out and output head, and adding training-side binding and selection objectives, both leave accuracy at or below a 0.52 query-blind ceiling. The decisive change is to make adapter composition query-conditioned: a late-interaction (MaxSim) address in the frozen encoder's space selects which chunk's generated adapter is merged, and with relevance-profile atom placement and delimiter-aligned chunk boundaries this raises exact-match accuracy from 0.52 to 0.98 across context lengths from 128 to 65,536 tokens, without touching the base model, the adapter generator, or the training loss. The broader lesson is that

weight-space context internalization is, at long context, a routing problem: when many facts share one merged adapter, retrieval is best resolved when the adapter is composed, and a token-level late-interaction score is a simple, parameter-free, and effective router. We hope this reframing – selection over generation and supervision – informs the design of scalable context-to-weight systems.

## References

Bowen Cao, Deng Cai, and Wai Lam. Infiniteicl: Breaking the limit of context window size via long short-term memory transformation. In *Findings of the Association for Computational Linguistics: ACL 2025*, page 11402–11415. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.595. URL http://dx.doi.org/10.18653/v1/2025.findings-acl.595.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models. 2022. doi: 10.21203/rs.3.rs-1553541/v1. URL http://dx.doi.org/10.21203/rs.3.rs-1553541/v1.

Yufeng Du, Minyang Tian, Srikanth Ronanki, Subendhu Rongali, Sravan Bodapati, Aram Galstyan, Azton Wells, Roy Schwartz, E. A. Huerta, and Hao Peng. Context length alone hurts llm performance despite perfect retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, page 23281–23298. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1264. URL http://dx.doi.org/10.18653/v1/2025.findings-emnlp.1264.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition, 2022. URL https://arxiv.org/abs/2209.10652.

Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks, 2020. URL https://arxiv.org/abs/2012.05208.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff, and Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey, 2024. URL https://arxiv.org/abs/2403.14608.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models?, 2024. URL https://arxiv.org/abs/2404.06654.

Chengsong Huang, Qian Liu, Bill Lin, Tianyu Pang, Chao-Hai Du, and Min Lin. Lorahub: Efficient cross-task generalization via dynamic lora composition, 2023. URL https://arxiv.org/abs/2307.13269.

Hamish Ivison, Akshita Bhagia, Yizhong Wang, Hannaneh Hajishirzi, and Matthew E. Peters. Hint: Hypernetwork instruction tuning for efficient zero- and few-shot generalisation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 11272–11288. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.631. URL http://dx.doi.org/10.18653/v1/2023.acl-long.631.

Ryosuke Kawata, Joonho Lee, Yanlei Gu, and S. Kamijo. Adaptive multiple-attribute scenario lora merge for robust perception in autonomous driving. *Sensors*, 26(4):1336, February 2026. ISSN 1424-8220. doi: 10.3390/s26041336. URL http://dx.doi.org/10.3390/s26041336.

Omar Khattab, Christopher Potts, and Matei Zaharia. Relevance-guided supervision for openqa with colbert. *Transactions of the Association for Computational Linguistics*, 9:929–944, 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00405. URL http://dx.doi.org/10.1162/tacl_a_00405.

Mateusz Klimaszewski, Piotr Andruszkiewicz, and Alexandra Birch. No train but gain: Language arithmetic for training-free language adapters enhancement, 2024. URL https://arxiv.org/ab

s/2404.15737.

Qi Liu, Gang Guo, Jiaxin Mao, Zhicheng Dou, Ji-Rong Wen, Hao Jiang, Xinyu Zhang, and Zhao Cao. An analysis on matching mechanisms and token pruning for late-interaction models. *ACM Transactions on Information Systems*, 42(5):1–28, April 2024. ISSN 1558-2868. doi: 10.1145/3639818. URL http://dx.doi.org/10.1145/3639818.

Xinyu Ma, Jiafeng Guo, Ruqing Zhang, Yixing Fan, and Xueqi Cheng. Pre-train a discriminative text encoder for dense retrieval via contrastive span prediction. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 848–858. ACM, 2022. doi: 10.1145/3477495.3531772. URL http://dx.doi.org/10.1145/3477495.3531772.

Yuren Mao, Yuhang Ge, Yijiang Fan, Wenyi Xu, Yu Mi, Zhonghao Hu, and Yunjun Gao. A survey on lora of large language models. *Frontiers of Computer Science*, 19(7), December 2024. ISSN 2095-2236. doi: 10.1007/s11704-024-40663-9. URL http://dx.doi.org/10.1007/s11704-024-40663-9.

Klaus Oberauer. Interference between storage and processing in working memory: Feature overwriting, not similarity-based competition. *Memory & Cognition*, 37(3):346–357, April 2009. ISSN 1532-5946. doi: 10.3758/mc.37.3.346. URL http://dx.doi.org/10.3758/mc.37.3.346.

Jonathan C. Roberts, Kai Han, and Samuel Albanie. Needle threading: Can llms follow threads through near-million-scale haystacks?, 2024. URL https://arxiv.org/abs/2411.05000.

Keshav Santhanam, Omar Khattab, Christopher Potts, and Matei Zaharia. Plaid: An efficient engine for late interaction retrieval, 2022a. URL https://arxiv.org/abs/2205.09707.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. Colbertv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 3715–3734. Association for Computational Linguistics, 2022b. doi: 10.18653/v1/2022.naacl-main.272. URL http://dx.doi.org/10.18653/v1/2022.naacl-main.272.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, page 3784–3803. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-emnlp.320. URL http://dx.doi.org/10.18653/v1/2021.findings-emnlp.320.

Anke Tang, Li Shen, Yong Luo, Nan Yin, Lefei Zhang, and Dacheng Tao. Merging multi-task models via weight-ensembling mixture of experts, 2024. URL https://arxiv.org/abs/2402.00433.

Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou, and Daniel Cer. Spot: Better frozen model adaptation through soft prompt transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 5039–5059. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.346. URL http://dx.doi.org/10.18653/v1/2022.acl-long.346.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Simlm: Pre-training with representation bottleneck for dense passage retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 2244–2258. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.125. URL http://dx.doi.org/10.18653/v1/2023.acl-long.125.

Desheng Wu, Guangyuan Ma, Meng Lin, Zijia Lin, Zhongyuan Wang, and Songlin Hu. Contextual

masked auto-encoder for dense passage retrieval. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4):4738–4746, 2023. ISSN 2159-5399. doi: 10.1609/aaai.v37i4.25598. URL http://dx.doi.org/10.1609/aaai.v37i4.25598.

Enneng Yang, Shen Li, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities, 2024a. URL https://arxiv.org/abs/2408.07666.

Enneng Yang, Li Shen, Zhenyi Wang, Guibing Guo, Xiaojun Chen, Xingwei Wang, and Dacheng Tao. Representation surgery for multi-task model merging, 2024b. URL https://arxiv.org/abs/2402.02705.

Jinghan Zhang, Shiqi Chen, Junteng Liu, and Junxian He. Composing parameter-efficient modules with arithmetic operations, 2023. URL https://arxiv.org/abs/2306.14870.

Wei Zhong, Jheng-Hong Yang, Yuqing Xie, and Jimmy Lin. Evaluating token-level and passage-level dense retrieval models for math information retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, page 1092–1102. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-emnlp.78. URL http://dx.doi.org/10.18653/v1/2022.findings-emnlp.78.