QUANTIFYING THE LINK FROM LOCALIZATION TO CONTEXT INTERVENTIONS IN LONG-CONTEXT QA

Anonymous authors
Paper under double-blind review

ABSTRACT

Search-augmented question answering, knowledge-base question answering, and multi-document analysis require language models to identify the right evidence among many documents and use it to produce an answer. Whether models can do this reliably determines whether longer context windows translate into stronger problem-solving capabilities. Existing methods use localization or relevance signals to prioritize context and improve its utilization through attention calibration, document reordering, compression, or pruning. Although prior work reports both localization quality and downstream QA performance, it does not explicitly quantify how localization successes and failures jointly determine intervention outcomes. We present a systematic analysis linking localization accuracy to downstream QA outcomes across multiple matched context interventions. In the actual locate-then-intervene workflow, we group records by whether the localizer selects the answer document, measuring the average gain after a hit and the average loss after a miss. A separate gold-document control applies the same intervention directly to the answer document to measure the intervention's improvement potential. We further show that document-localization signals concentrate in a small number of attention heads, and select these heads using questions disjoint from evaluation to connect localization quality to intervention outcomes. Experiments across multiple models and datasets show that localization and intervention based on a small number of attention heads improve answer-hit accuracy by 4.46 percentage points on average and yield positive effects in the large majority of configurations. Gold-document controls further show substantial remaining improvement potential across the tested interventions. On unseen models, using only a small labeled set, our method predicts whether an intervention will help or hurt with 78.1% accuracy, compared with 58.8% for the better constant policy, without using held-out full-QA outcomes at prediction time. These results turn localization accuracy from an isolated intermediate metric into a basis for explaining and predicting the effectiveness of context interventions, providing a unified view of how such methods can be analyzed and transferred.

1 INTRODUCTION

Long-context question answering requires models to identify relevant evidence among many documents and use it to produce an answer. Larger context windows provide access to more information but do not guarantee effective use. Evidence position can substantially alter predictions (Liu et al., 2024), while long-context benchmarks show that nominal context-window length does not ensure reliable use of all available context (Bai et al., 2024; Hsieh et al., 2024a). Reliable localization and utilization are therefore central to practical long-context reasoning.

Existing methods use question-to-document attention or specialized retrieval heads to identify or prioritize relevant context (Hsieh et al., 2024b; Peysakhovich & Lerer, 2023; Chen et al., 2025; Zhang et al., 2025), and improve context utilization through reordering, attention calibration, compression, or pruning (Hsieh et al., 2024b; Jiang et al., 2023; 2024; Li et al., 2023). These studies demonstrate end-to-end gains and often report both localization quality and downstream QA performance. However, they do not quantify how localization successes and failures jointly determine intervention outcomes or how accurate localization must be before a fixed intervention helps.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

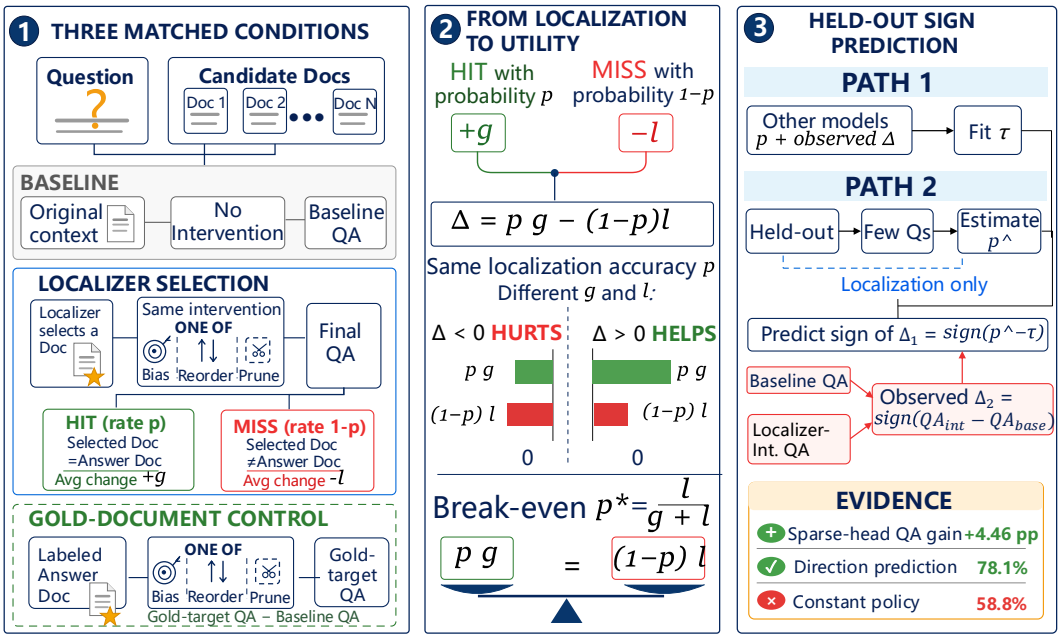


Figure 1: **How localization accuracy determines intervention effects. Left:** We compare three settings: no intervention, intervention on the localizer-selected document, and intervention on the answer document. **Middle:** Correct localization yields gains, whereas incorrect localization causes losses; together, they determine the intervention's break-even point. **Right:** We fit a threshold using other models and use a small sample to predict the effect direction on a held-out model. The method improves QA performance by 4.46 percentage points on average and achieves 78.1% directional accuracy, outperforming the better constant policy at 58.8%.

As illustrated in Figure 1, the net effect depends jointly on how often localization succeeds, how much intervention helps after a hit, and how much it hurts after a miss; equal localization accuracy can therefore yield different outcomes. Our experiments also show substantial intervention headroom when the answer document is supplied and localization signals concentrated in a few attention heads. Localization and intervention capacity should thus be measured separately but analyzed jointly.

In the actual workflow, we partition records by whether the localizer selects the answer document; a separate gold-document control measures intervention headroom. Effective attention heads are selected using questions disjoint from those used for evaluation.

Context interventions are not uniformly effective. Among 163 sparse-head and ensemble configurations, 138 improve over their matched baselines, while 25 do not, despite an average gain of 4.46 pp. Our gain–loss decomposition explains this variation through the benefits of localization hits and the costs of misses, yielding a configuration-specific break-even accuracy at which the expected QA effect changes sign. Because this quantity requires completed QA outcomes, we use it to diagnose observed results; for unseen models, we instead fit a threshold on other models for localization-only direction prediction.

Our contributions are threefold:

1. We establish a **quantitative link between localization accuracy and intervention utility** by decomposing realized effects into gains after localization hits and losses after misses, while separating both from intervention headroom measured by the gold-document control.

2. We show that **document-localization signals concentrate in a few attention heads** and use them to drive three context-intervention families, improving answer-hit accuracy by 4.46 pp on average.

3. We introduce a **localization-only predictor for held-out models** that requires no full-QA outcomes at prediction time and reaches 78.1% directional accuracy versus 58.8% for the better constant policy.

2 Related Work

2.1 Long-Context Retrieval and Position Bias

Long-context QA varies with evidence position, often following a U-shaped pattern (Liu et al., 2024). FITM estimates and subtracts positional attention using uninformative documents (Hsieh et al., 2024b); related work studies attention basins and head-dependent positional encodings (Yi et al., 2026; Zhang et al., 2024). LongBench and RULER evaluate long-context understanding and effective context length (Bai et al., 2024; Hsieh et al., 2024a). We instead study how localization quality affects subsequent intervention utility in natural multi-document QA.

Attention-based localization includes attention sorting (Peysakhovich & Lerer, 2023), calibrated reranking (Chen et al., 2025), sparse retrieval-head identification (Wu et al., 2025), and few-shot selection of query-focused heads (Zhang et al., 2025). Prior mechanistic work has shown that individual attention heads can support specialized in-context behaviors such as induction (Olsson et al., 2022). We similarly select sparse localization heads on independent questions, but ask how gains after hits and losses after misses jointly determine intervention utility.

2.2 Context Intervention and Downstream Utility

Methods for improving context utilization include document reordering (Peysakhovich & Lerer, 2023; Jiang et al., 2024), attention calibration (Hsieh et al., 2024b), and context reduction through document retention or token-level compression (Li et al., 2023; Jiang et al., 2023; 2024). We evaluate matched document-level conditions for these operation families rather than treating them as interchangeable.

Prior work commonly reports retrieval quality and downstream QA performance, while Salemi & Zamani (2024) evaluate retrieval using downstream RAG utility. These measurements do not decompose how localization successes and errors contribute to matched target-specific interventions. We address this relationship by grouping records from the actual localizer-selected intervention into hits and misses, while using a separate gold-document control to measure the intervention's remaining improvement potential.

Small labeled sets have been used to configure positions or attention heads (Yi et al., 2026; Zhang et al., 2025). We instead test whether localization performance alone can predict the direction of a configured intervention's QA effect on a held-out model. The prediction threshold is selected using other-model results and is distinct from the configuration-level break-even point, which requires completed QA experiments. This setting is also different from selective QA: selective QA decides whether to answer an individual question and studies accuracy–coverage tradeoffs (Kamath et al., 2020), whereas our prediction concerns the aggregate effect of an intervention configuration on a new model.

3 From Sparse Localization to Intervention Utility

3.1 Locate–then–Intervene

Each record contains a question q_i, a candidate set D_i, and one labeled answer document d_i^*. Our workflow first selects a target document and then applies a fixed context intervention to that target. The central question is not how to train a new QA model, but how the benefits of correct target selection and the costs of incorrect selection jointly determine the final QA effect. Alternative localizers are introduced as comparison methods in Section 4.

3.2 Sparse Attention-Head Localization

For attention head h, let $\mathcal{T}_i(d)$ be the token span of document d. We score the document by the mean attention from the final question token:

$$a_{ih}(d) = \frac{1}{|\mathcal{T}_i(d)|} \sum_{t \in \mathcal{T}_i(d)} A_h(q_i^{\text{last}}, t). \tag{1}$$

We standardize this score across the candidate documents for each question and head,

$$\widetilde{a}_{ih}(d) = \frac{a_{ih}(d) - \mu_{ih}}{\sigma_{ih} + \epsilon}, \tag{2}$$

so heads with different attention scales can be combined.

Questions are split into two folds by question-ID parity. To evaluate question i, we use only middle-position records from the other fold to select the $k = 4$ heads with highest localization accuracy. If $\mathcal{H}_{-f(i)}$ denotes this selected set, the localizer predicts

$$\widehat{d}_i = \arg\max_{d \in D_i} \frac{1}{k} \sum_{h \in \mathcal{H}_{-f(i)}} \widetilde{a}_{ih}(d). \tag{3}$$

Thus, no question contributes to both head selection and evaluation, and no model parameter is updated. The complete layer–head search avoids assuming that useful localization must occur late. The fixed $k = 4$ rule provides one shared default across models; nested selection of k is examined separately in Appendix B.

3.3 Matched Experimental Conditions

The selected document is passed to one of three operations. **Attention biasing** adds a generation-time attention-logit bias to tokens in the selected document. **Reordering** moves the selected document to the first position while preserving the relative order of the others. **Pruning** retains the selected document and a specified number of highly ranked documents, removing the rest without changing the retained order.

For each model, question set, intervention, and operation strength, we compare three conditions:

1. **Baseline:** the original context is left unchanged.
2. **Localizer selection:** the intervention is applied to the document selected by the localizer.
3. **Gold-document control:** the same intervention is applied directly to the labeled answer document.

The latter two conditions use the same operation code path and differ only in the target index. Whenever the localizer selects the answer document, their record-level generations are identical. The gold-document control is compared with baseline to measure the intervention's improvement potential; it is not used to define the hit gain below. Detailed localizer definitions, operation strengths, pruning scans, and implementation checks appear in Appendices B and G.

3.4 From Localization to Intervention Utility

Let $c = (M, \mathcal{S}, L, A, \lambda)$ denote a configuration consisting of a model, evaluation setting, localizer, intervention, and intervention strength. For record i, the labeled answer document is d_i^* and the localizer selects $\hat{d}_{ic} = L(q_i, D_i)$. We define

$$\begin{aligned} H_{ic} &= \mathbb{1}[\hat{d}_{ic} = d_i^*], & \delta_{ic} &= Y_{ic}^L - Y_i^0, \\ Y_i^0 &= \text{QA score under baseline}, & Y_{ic}^L &= \text{QA score after applying } (A, \lambda) \text{ to } \hat{d}_{ic}. \end{aligned} \tag{4}$$

All expectations below are empirical averages over the evaluation records of configuration c. The localization accuracy, hit-conditional gain, miss-conditional loss, and overall QA effect are

$$\begin{aligned} p_c &= \mathbb{E}[H_{ic}], & g_c &= \mathbb{E}[\delta_{ic} \mid H_{ic} = 1], \\ l_c &= -\mathbb{E}[\delta_{ic} \mid H_{ic} = 0], & \Delta_c &= \mathbb{E}[\delta_{ic}]. \end{aligned} \tag{5}$$

Conditioning on H_{ic} yields the exact decomposition

$$\Delta_c = p_c g_c - (1 - p_c) l_c, \qquad p_c^* = \frac{l_c}{g_c + l_c}. \tag{6}$$

When $g_c, l_c > 0$, p_c^* is the break-even accuracy obtained while holding the observed conditional effects fixed. It diagnoses whether a poor result is associated with low localization accuracy, limited gain after a hit, or high loss after a miss; it is not a universal threshold for changing the localizer.

The gold-document score Y_{ic}^G applies the same (A, λ) directly to d_i^*. Its full-sample headroom is

$$h_c = \mathbb{E}[Y_{ic}^G - Y_i^0]. \tag{7}$$

Crucially, h_c is not g_c: the former averages over all records under a supplied gold target, whereas the latter conditions on the subset that the localizer actually hits.

3.5 Held-Out Prediction Protocol

The prediction protocol evaluates a fully specified configuration rather than choosing its localizer, operation, or strength. For held-out model m, let $\mathcal{C}_{-m}$ contain configurations from all other models. We select a single threshold that maximizes the QA gain retained on those configurations:

$$\tau_{-m} = \arg\max_{\tau \in \mathcal{T}} \sum_{c \in \mathcal{C}_{-m}} \Delta_c \mathbb{1}[p_c \geq \tau], \tag{8}$$

where $\mathcal{T}$ contains the observed localization accuracies from the other models and two endpoint decisions. This is a cross-model calibration rule, not model-parameter training, and is distinct from each configuration's post-hoc p_c^*.

On the held-out model, an external annotation budget of b questions is labeled with answer documents. Moving each answer document to five positions yields the audit set $\mathcal{A}_c^{(b)}$ and

$$\widehat{p}_c^{(b)} = \frac{1}{|\mathcal{A}_c^{(b)}|} \sum_{i \in \mathcal{A}_c^{(b)}} H_{ic}, \qquad \widehat{z}_c = \mathbb{1}[\widehat{p}_c^{(b)} \geq \tau_{-m}], \tag{9}$$

where $\widehat{z}_c = 1$ predicts a positive intervention effect. To match the compact signed notation in Figure 1, we additionally denote the predicted and observed directions by

$$\Delta_{1,c} = \operatorname{sign}\left(\widehat{p}_c^{(b)} - \tau_{-m}\right), \qquad \Delta_{2,c} = \operatorname{sign}(\Delta_c), \tag{10}$$

where Δ_c is the aggregate difference between localizer-intervention QA and baseline QA. Thus, $\Delta_{1,c}$ and $\Delta_{2,c}$ are direction labels rather than effect magnitudes. The reported computation retains the indicator coding in Equations 9 and 11, which makes the tie conventions explicit: $\widehat{p}_c^{(b)} = \tau_{-m}$ predicts a positive effect, whereas $\Delta_c = 0$ is treated as non-positive.

This stage runs the localizer but performs no downstream QA intervention or decoding. Directional accuracy is evaluated only afterward as

$$\mathrm{Acc}_{\mathrm{dir}} = \frac{1}{|\mathcal{C}_m|} \sum_{c \in \mathcal{C}_m} \mathbb{1}[\widehat{z}_c = \mathbb{1}[\Delta_c > 0]]\,. \tag{11}$$

Head selection and held-out prediction use non-overlapping questions. We report $b = 20$ as a representative low-budget operating point and evaluate sensitivity to b separately. Near-threshold estimates may be deferred for additional labels. Appendix E reports the break-even summary, older record-sampling protocol, deferral curve, and semi-synthetic budget analysis.

4 Experimental Design

4.1 Models, Data, and Configurations

We evaluate eight models: Qwen3-4B/8B/14B/32B, Llama-3-8B, Llama-3.1-8B/70B, and Vicuna-7B-16k. The sparse-head layer and ensemble-size ablations cover the seven Qwen and Llama models; Vicuna is used in the FITM transfer analysis. The datasets are an open-domain multi-document

adaptation of Natural Questions (Kwiatkowski et al., 2019), PopQA and a hard-distractor variant (Mallen et al., 2023), and MuSiQue (Trivedi et al., 2022). Together they cover open-domain single-hop QA, a different knowledge distribution, distractors that assert incorrect answers, and multi-hop reasoning. Main experiments use contexts of 20 or 30 candidate documents. Experiments with 10, 60, and 120 documents test changes in context length and are reported as boundary analyses.

Standard experiments place the answer document of each of 100 questions at five positions, producing 500 records. A configuration specifies the model, dataset, document count, localizer, intervention, and strength; after filtering, the main analysis contains 388 configurations from 65 result files.

4.2 Comparison Methods and Intervention Settings

In addition to our sparse attention-head localizer, we compare a sparse-head/model-self-report ensemble, direct model self-report, position-calibrated FITM, an average over all heads in the late layers, BM25, and a cross-encoder reranker. These methods represent three sources of localization evidence: internal attention, explicit model judgments, and external retrieval.

Each localizer is connected separately to attention biasing, reordering to the beginning, and context pruning; the main pruning condition retains only the target. Appendix B reports alternative retention counts, strength sweeps, and model-specific settings.

4.3 Metrics and Data Separation

Answer quality uses normalized answer-hit accuracy. Model outputs and acceptable answers are lowercased; punctuation and English articles are removed; and whitespace is normalized. A prediction is correct when any normalized acceptable answer is a substring of the normalized output. The original Lost-in-the-Middle evaluator calls this metric "Exact Match," but it is not strict full-string equality. Localization quality is the top-1 accuracy of selecting the answer document.

Records that share a question but differ in answer position are dependent. Statistical comparisons involving the gold-document control and localizers therefore use question-clustered permutation tests and bootstrap confidence intervals. All mean QA changes are computed relative to the corresponding baseline. For analyses without completed question-clustered inference, we report directions and point estimates without significance claims.

Data separation has two levels. First, attention-head selection and localization evaluation use different question folds. Second, in held-out prediction, all configurations of one model form the test fold and the threshold is selected using only other-model results. The questions used to select attention heads are also disjoint from the audit questions used to estimate localization accuracy on the new model. Directional accuracy measures whether a rule enables positive configurations and disables negative ones; the better constant policy is the more accurate of always predicting a positive effect and always predicting a non-positive effect. Available gain captured compares the rule's realized gain with the gain obtainable when every configuration's true direction is known in advance. No model parameters are updated. Appendix E reports the older record-sampling protocol and fixed-threshold sensitivity analyses.

5 Results

5.1 Localization Signals Concentrate in a Few Attention Heads

We first test whether document localization benefits from aggregating many attention heads. Figure 2 varies both the layer and the number of aggregated heads. On all seven models, selecting a small number of heads on independent questions outperforms averaging every head in the final 40% of layers. On five models, the late all-head average even falls below the fixed-position prior. The model-specific optimum always lies between one and eight heads; increasing the ensemble to 16 or 32 heads reduces accuracy relative to each model's peak. Effective localization signals are therefore concentrated in a few heads and become diluted by broad averaging.

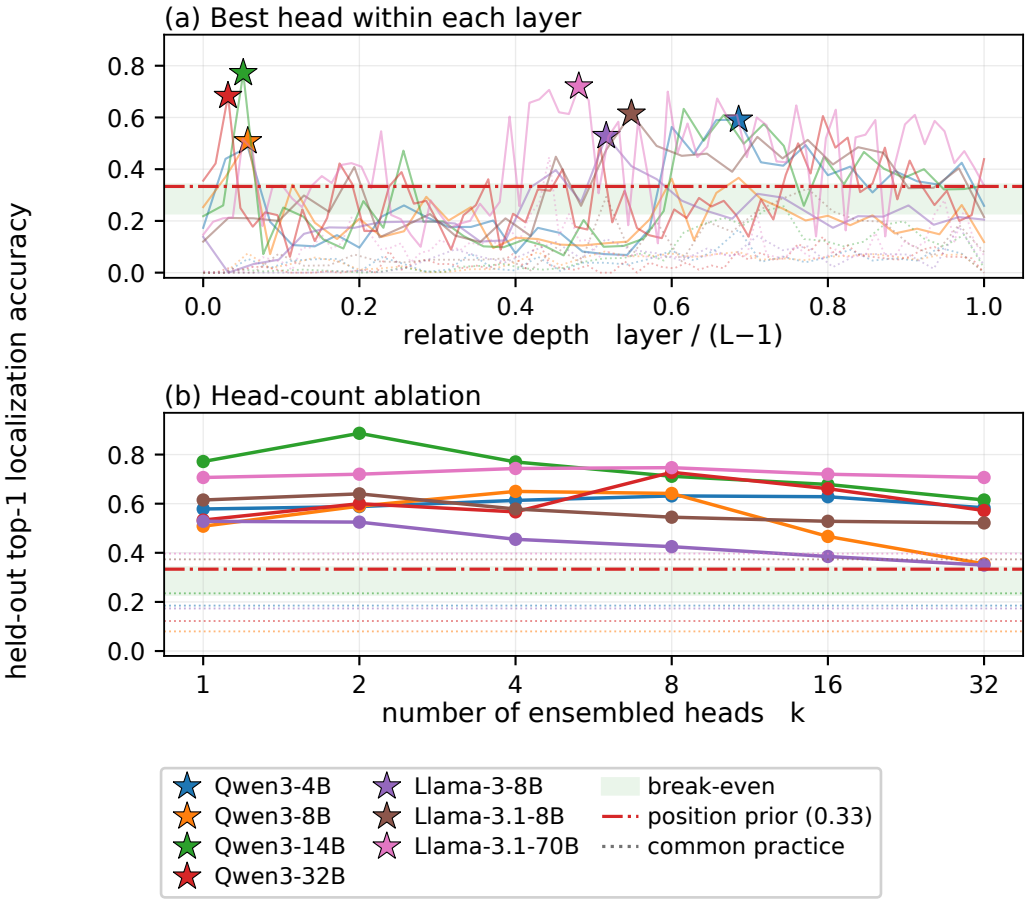


Figure 2: Localization ablations over attention layers and ensemble size. A few heads selected on independent questions outperform late-layer all-head averaging on all seven models; the best ensemble sizes range from one to eight, and adding more heads dilutes the localization signal.

Fixed $k = 4$ gives a mean localization accuracy of 0.625 and provides one common setting across models. Allowing k to be selected in an additional separated-data step raises the mean to 0.669, but the selected ensemble size varies substantially across models and question folds. We therefore use four heads in the main experiments and retain adaptive selection as a supplementary analysis.

Useful layers are model dependent: several Qwen models peak early, whereas Llama models peak more often in middle layers. We therefore search the complete layer–head set on independent questions; this supports signal sparsity but not transfer of head identities. Per-model results appear in Appendix B.

5.2 Sparse-Head Interventions Improve QA

Improved localization must ultimately translate into better answers. We aggregate 163 configurations that use sparse attention heads or their ensemble with model self-report and compare the actual localizer-selection condition with its corresponding baseline. Configuration-weighted answer-hit accuracy rises from 56.7% to 61.2%, an average gain of 4.46 pp, and 138 of the 163 configurations have positive effects. These results come from the actual workflow in which the localizer selects a document before attention biasing, reordering, or pruning; they do not use the gold-document control.

Thus, localization signals from a few attention heads improve not only document selection but also final QA in the large majority of tested configurations. The remaining variation cannot, however, be

attributed to target selection or intervention capacity from the average alone. We next isolate these factors with the gold-document control and the gain–loss decomposition.

5.3 Gold-Document Controls Reveal Intervention Headroom

Sparse attention heads improve localization, but localization quality alone does not establish whether the subsequent intervention is useful. To separate these factors, we apply the same intervention directly to the answer document in 11 model–setting groups while holding the model, question, candidate documents, operation, and strength fixed. This control measures the remaining improvement potential when the target is always correct; it is distinct from the hit-conditional gain of the localizer.

All 11 controls have positive point estimates of 7.8–20.5 pp, and 10 reach nominal question-clustered $p < 0.05$; the full results appear in Appendix G. In the matched FITM transfer example, the end-to-end change is near zero under FITM localization, rises to 16.3 pp with sparse-head selection, and reaches 20.5 pp with the answer document supplied. Because the latter two conditions use the same operation and coincide whenever localization succeeds, their gap isolates target selection.

Headroom depends on document count and operation strength. The gold-document control therefore establishes remaining intervention capacity, not equal effectiveness in every setting.

5.4 Conditional Gains and Losses Explain Configuration Differences

We next compute p_c, g_c, and l_c from the actual localizer-selection records of each configuration. Among configurations with positive hit gains and miss losses, the median break-even accuracy p_c^* is 0.419, with an interquartile range of 0.277–0.512. This spread shows that interventions differ substantially in their tolerance to target errors: a large hit gain or small miss loss permits a positive effect at lower localization accuracy, whereas high localization accuracy need not suffice when this balance is unfavorable.

The median distance between observed localization accuracy and the configuration-level break-even point is 0.188, while approximately 28% of configurations satisfy $|p_c - p_c^*| < 0.10$. These near-boundary configurations are particularly sensitive to localization changes. Because p_c^* requires complete QA outcomes, it is used only to explain completed experiments. Prediction on a new model instead uses an empirical threshold selected on other-model results, not p_c^*.

Selection effects. The localizer is more accurate on questions already answered correctly under the baseline, but controlled comparisons at matched overall localization accuracy show opposing effects on initially correct and incorrect questions that largely cancel in aggregate. Adding this gap also does not improve held-out decisions. We therefore retain overall localization accuracy in the main analysis and report the controlled construction and alternative diagnostics in Appendices C and D.

5.5 Localization-Only Prediction on Held-Out Models

Table 1: Leave-one-model-out deployment decisions. Available gain is the gain captured relative to perfect knowledge of each configuration's direction.

Decision rule	Directional accuracy	Available gain captured
Better constant policy	0.588	8.8%
10 questions, all positions	0.766	62.1%
20 questions, all positions	**0.781**	**64.2%**
50 questions, all positions	0.790	59.8%
All localization data (reference)	0.799	66.9%

Across 388 configurations, the representative $b = 20$ operating point reaches a pooled, configuration-weighted retrospective directional-accuracy estimate of 0.781, compared with 0.588 for the better constant policy, and captures 64.2% of available gain (Table 1). A 5,000-replicate

file-cluster bootstrap that reselects the leave-one-model threshold within each replicate and propagates audit-question sampling gives a 19.4-point advantage over the better constant policy (95% interval, 3.4–29.4 points). Each question contributes five constructed positions but requires only one answer-document annotation, and the audit performs no downstream QA intervention or decoding.

The audit size is an external annotation budget rather than a tuned method parameter. Under the same leave-one-model protocol, 10, 20, and 50 questions yield directional accuracies of 0.766, 0.781, and 0.790. Under a separate fixed-$\tau = 0.454$ resampling analysis, their mean decision-flip rates are 0.093, 0.063, and 0.035. The smooth trade-off has no special elbow at 20; it shows that the conclusion is not tied to a uniquely selected audit size, while additional labels mainly stabilize configurations near the threshold.

The older 20-record protocol reaches 0.765 directional accuracy; the question/all-position protocol's improvement comes mainly from reusing all five positions. Full comparisons remain in Appendix E.

Separate annotation costs. Head selection and held-out auditing consume separate, non-overlapping labels. Their sample sizes therefore remain explicit costs rather than jointly tuned hyperparameters. Appendix E reports the full audit-size, deferral, and semi-synthetic allocation analyses; none is used to claim an optimal universal budget.

5.6 Source of Gate Effectiveness

The aggregate benefit comes primarily from rejecting unreliable localizer families. Attention-head configurations already gain 4.46 pp under unconditional use and 4.62 pp after gating, with an interval for the difference that includes zero, whereas gating brings the clear average losses of FITM and raw-attention configurations to approximately zero. The gate is therefore mainly a cross-family filter, not a consistent improvement within every reliable family.

Removing FITM and retaining only NQ reduces the directional-accuracy advantage over the constant policy to 0.8 pp, with interval $[-21.2, +6.1]$. The full family breakdown and direct-audit analysis appear in Appendices F and D.

5.7 Boundary Conditions Change the Gate Inputs

The gate transfers only while the localization target and distributions of p, g, and l remain comparable. Stress tests vary wrong-target loss, multi-document evidence, model scale, context length, and model family; notably, a threshold calibrated at 20–30 documents rejects three positive 120-document configurations. Appendix F provides full results.

The 120-document failure is informative at the position level. Head localization accuracy is 0.895 and 0.495 at the two edge positions but only 0.065, 0.115, and 0.145 at the three middle positions. Full-position gains are positive, while middle-position gains are zero or negative at the tested doses. Pooling these regimes produces a localization estimate below the frozen threshold even though the overall effect is positive, exposing a threshold-calibration failure rather than a reversal of the position-level relation.

6 Conclusion

Context interventions do not uniformly improve long-context QA: their effects depend on target-localization accuracy and on gains after hits versus losses after misses. Our gain–loss decomposition makes this dependence explicit and yields a configuration-specific break-even accuracy where the expected effect changes sign. Matched baseline, localizer-selected, and gold-document conditions separate localization reliability from intervention capacity.

Empirically, sparse-head localization and intervention improve QA by 4.46 pp on average, but 25 of 163 configurations do not improve. For unseen models, a localization-only audit predicts intervention direction with 78.1% accuracy versus 58.8% for the better constant policy. Thus, the central question is not whether context intervention helps in general, but whether localization is reliable enough for a particular intervention.

AI USE STATEMENT

Generative AI tools were used to draft sections of this manuscript from author-provided notes and to aid or polish the writing, including reorganization of the presentation and edits for readability. They also assisted with literature discovery. Every citation was checked against the original source. No experimental result was generated by the writing assistant, and all numerical claims were taken from author-provided materials.

The language models evaluated in this study are Qwen3-4B, Qwen3-8B, Qwen3-14B, Qwen3-32B, Llama-3-8B, Llama-3.1-8B, Llama-3.1-70B, and Vicuna-7B-16k.

All AI-assisted text was reviewed and revised by the authors. Generative AI tools were not treated as authors and did not make decisions about the study design, experimental interpretation, or final claims. The authors take full responsibility for the content of the manuscript.

ETHICS STATEMENT

This work follows standard ethical research practices and is intended to support scientific understanding of long-context language models. The experiments use existing research datasets, language models, and software resources in accordance with their applicable terms of use. No new human-subject study or collection of personal information was conducted as part of this work.

The study was designed and reported with attention to research integrity, appropriate use of experimental resources, and clear communication of its scope. The authors take responsibility for the design, analysis, interpretation, and presentation of the study.

REPRODUCIBILITY STATEMENT

The paper and appendices document the methodology, experimental setup, evaluation procedures, and implementation details used in this study. We describe the models, datasets, comparison methods, metrics, data separation, and analysis protocol used to obtain the reported results. Additional definitions, experimental settings, and supporting analyses are provided in the appendices.

The experiments follow consistent evaluation procedures across the reported configurations, and the numerical results are linked to the corresponding settings described in the manuscript. These materials are intended to support independent understanding, verification, and reproduction of the reported findings.

REFERENCES

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137, 2024. doi: 10.18653/v1/2024.acl-long.172. URL https://aclanthology.org/2024.acl-long.172/.

Shijie Chen, Bernal Jiménez Gutiérrez, and Yu Su. Attention in large language models yields efficient zero-shot re-rankers. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/b5b1890a7c1a79fe9b32b0f442347089-Abstract-Conference.html.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*, 2024a. URL https://openreview.net/forum?id=kIoBbc76Sy.

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. Found in the middle: Calibrating positional attention bias improves long context utilization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14982–14995, 2024b. doi: 10.18653/v1/2024.findings-acl.890. URL https://aclanthology.org/2024.findings-acl.890/.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LLMLingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13358–13376, 2023. doi: 10.18653/v1/2023.emnlp-main.825. URL https://aclanthology.org/2023.emnlp-main.825/.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1658–1677, 2024. doi: 10.18653/v1/2024.acl-long.91. URL https://aclanthology.org/2024.acl-long.91/.

Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5684–5696, 2020. doi: 10.18653/v1/2020.acl-main.503. URL https://aclanthology.org/2020.acl-main.503/.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. Compressing context to enhance inference efficiency of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6342–6353, 2023. doi: 10.18653/v1/2023.emnlp-main.391. URL https://aclanthology.org/2023.emnlp-main.391/.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric

memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, 2023. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022. doi: 10.48550/arXiv.2209.11895. URL https://arxiv.org/abs/2209.11895.

Alexander Peysakhovich and Adam Lerer. Attention sorting combats recency bias in long context language models. *arXiv preprint arXiv:2310.01427*, 2023. doi: 10.48550/arXiv.2310.01427. URL https://arxiv.org/abs/2310.01427.

Alireza Salemi and Hamed Zamani. Evaluating retrieval quality in retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2395–2400, 2024. doi: 10.1145/3626772.3657957. URL https://doi.org/10.1145/3626772.3657957.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/9b77f07301b1ef1fe810aae96c12cb7b-Abstract-Conference.html.

Zihao Yi, Zhenqing Ling, Delong Zeng, Haohao Luo, Zhe Xu, Wei Liu, Jian Luan, Wanxia Cao, and Ying Shen. Attention basin: Why contextual position matters in large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 26086–26106, 2026. doi: 10.18653/v1/2026.acl-long.1198. URL https://aclanthology.org/2026.acl-long.1198/.

Wuwei Zhang, Fangcong Yin, Howard Yen, Danqi Chen, and Xi Ye. Query-focused retrieval heads improve long-context reasoning and re-ranking. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 23791–23805, 2025. doi: 10.18653/v1/2025.emnlp-main.1214. URL https://aclanthology.org/2025.emnlp-main.1214/.

Zhenyu Zhang, Runjin Chen, Shiwei Liu, Zhewei Yao, Olatunji Ruwase, Beidi Chen, Xiaoxia Wu, and Zhangyang Wang. Found in the middle: How language models use long contexts better via plug-and-play positional encoding. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6ffdbbe354893979367f93e2121e37dd-Abstract.html.

A Limitations

As with any empirical study, the findings should be understood within the scope of the evaluated models, datasets, and experimental settings. Future work may examine the framework in additional settings and explore broader applications.

B Localization and Implementation Details

Localizers. The sparse attention localizer uses length-normalized, standardized layer–head document scores. The ensemble adds a fixed contribution when a model self-report parses to a valid document index. Raw attention averages document attention over all heads in the final 40% of layers. Self-report prompts ask for the relied-upon document index; invalid parses count as failures.

FITM uses uninformative documents to estimate position-dependent attention, with original and simplified implementations reported separately. BM25 and an MS MARCO cross-encoder score question–document pairs independently of generator internals.

Nested ensemble-size selection. Across models, fixed $k = 4$ averages 0.625 and nested selection averages 0.669. The nested choices vary across models and inner folds, including values from 1 to 16, and the gain is not stable enough to replace the shared default. The complete $k \in \{1, 2, 4, 8, 16, 32\}$ curves are shown in Figure 2.

Matched implementation audit. The localizer-selection and gold-document-control conditions call the same operation with different target indices. Whenever the localizer selects the answer document, their record-level generations are identical. This check rules out separate generation paths as the source of the gap between these conditions.

C Selection Effects and Synthetic Localizers

External retrieval signals show the same initially-correct localization advantage, suggesting that question difficulty affects both answering and localization rather than only one internal representation.

For the controlled test, two synthetic localizers are matched in total accuracy: one distributes errors uniformly, and one concentrates errors on originally incorrect questions. The comparison spans three models, three operations, and five accuracy levels. Operation-level aggregation finds a $+1.65$ pp residual for biasing at dose 3, with model- and question-clustered interval $[+0.43, +2.88]$ and $p = 0.014$, but this result does not survive correction across four operation-level tests. Detectable-improvement bounds are 1.2 pp for reordering and 1.8 pp for pruning. Thus the branch placement has visible effects, but does not provide a stable replacement for the overall-accuracy decision rule.

D Alternative Diagnostics and Direct Intervention Audit

Richer localization-only diagnostics do not improve the leave-one-model-out decision. Adding the selection gap changes gain-regression mean absolute error from 0.0469 to 0.0479. Adding baseline QA accuracy yields 0.760 directional accuracy and 58.7% of available gain, compared with 0.799 and 66.9% using all-data localization accuracy alone. Operation-specific thresholds also yield 0.799 while reducing captured gain to 65.5%. Restricting localization accuracy to originally incorrect questions reduces directional accuracy to 0.693.

When interventions may be run on the audit sample, direct gain estimation is a different budget regime. Table 2 reports the complete older 20-record comparison. These results show that intervention runs can improve average gain, while localization-only auditing minimizes audit execution cost.

Table 2: Deployment rules under the older 20-record audit. The worst-model value is an empirical leave-one-model-out result, not a risk guarantee.

Rule	Pooled gain (pp)	Worst model (pp)	Directional acc.	Audit runs
Always use	$+0.25$	-7.48	0.585	Full
Localization gate	$+1.81$	-0.05	0.764	0
Direct estimate, $\hat{\Delta} > 0$	$+1.77$	-0.76	0.649	20
Direct estimate, $\hat{\Delta} \geq 0$	$+1.94$	-1.46	0.690	20
Both, $\hat{\Delta} > 0$	$+1.80$	—	0.654	20
Both, $\hat{\Delta} \geq 0$	$+2.12$	-0.03	0.725	20
Perfect direction	$+2.94$	—	1.000	—

E GATE PROTOCOL, STABILITY, AND BUDGET COMPARISONS

Protocol. At the reported $b = 20$ operating point, sample 20 questions, annotate the answer document once per question, run five constructed localizer evaluations at reordered positions, estimate $\hat{p}$, and compare it with the leave-one-model-out threshold. This audit performs no downstream QA-generation intervention or decoding.

Table 3: Observed conditional break-even summary. Configuration-level p^* requires realized localizer-selection outcomes and is used only as a post-hoc diagnostic.

Quantity	Value
Median p^* (interquartile range)	0.419 (0.277–0.512)
Median $\lvert p - p^* \rvert$	0.188
Configurations with $\lvert p - p^* \rvert < 0.10$	approximately 28%

Question-level sensitivity. Under the fixed threshold $\tau = 0.454$, 10, 20, and 50 annotated questions yield mean flip rates 0.093, 0.063, and 0.035; near-threshold flip rates 0.305, 0.243, and 0.141; and directional accuracies 0.777, 0.796, and 0.812. These fixed-threshold sensitivity results are separate from the 0.781 leave-one-model-out refitted-threshold estimate.

Older record protocol. Sampling 20 records gives 0.765 directional accuracy, 1.6 pp below the question/all-position protocol. Sampling 20 question IDs but one position per question gives approximately 0.767, only about 0.2 pp above record sampling; using all five positions supplies the larger stability gain. In this older protocol alone, empirical worst-model gain changes from -7.48 pp under unconditional use to -0.05 pp after gating. Separately fitted model thresholds range from 0.410 to 0.515.

Deferral. Under the older protocol, deferring configurations with $|\hat{p} - \tau| < 0.10$ abstains on 21.9% and increases accuracy on decided configurations from 0.773 to 0.819. This is a descriptive conditional-coverage result from the observed curve, not a full-coverage or safety guarantee.

Semi-synthetic budget allocation. With 40 total labels, selection/audit splits 5/35, 10/30, 20/20, 30/10, and 35/5 produce net effects $+0.0502$, $+0.0590$, $+0.0608$, $+0.0592$, and $+0.0550$ and sign accuracies 0.852, 0.892, 0.897, 0.880, and 0.829. The 20/20 split has the highest tested point estimates but lies within a broad 10–30 plateau. Wrong-target outcomes are partly imputed, and simulator sign fidelity is 0.807, so these results compare relative budget allocation rather than provide absolute end-to-end validation.

F LOCALIZATION-FAMILY BREAKDOWN AND BOUNDARY CONDITIONS

Removing FITM and retaining only NQ reduces the gate's directional-accuracy advantage over the constant policy to 0.8 pp, with interval $[-21.2, +6.1]$, and its captured-gain advantage to 2.5 pp, with interval $[-18.3, +7.1]$. This confirms that aggregate effectiveness depends on family composition.

Table 4: Mean QA change (pp) by localizer family before and after gating. The five named families cover 373 of 388 configurations; 15 heterogeneous configurations are grouped as other and omitted.

Family	Configurations	Always use	After gate
Attention heads and ensembles	163	$+4.46$	$+4.62$
Model self-report	103	-0.34	$+0.09$
FITM	64	-5.89	$+0.00$
Raw attention	36	≈ -2.7	≈ 0
Off-the-shelf retrievers	7	-3.56	-0.24

Hard distractors. When distractors explicitly assert wrong answers, gold-document gain is 35.5 pp but incorrect targeting costs approximately 30.0 pp. Break-even accuracy rises to 0.458–0.550, the best label-free localizer reaches 0.394, and net change is -4.2 pp and not significant.

Multi-hop targets. MuSiQue contains connected multi-hop evidence (Trivedi et al., 2022). A retriever selects a relevant bridge document without the final answer in 81.5% of cases; single-document intervention then changes normalized answer-substring hit accuracy by -10.0 pp, whereas gold-document targeting gives $+11.4$ pp. Multi-hop use requires an evidence set or chain rather than one answer-document target.

Model scale. On Llama-3.1-70B, localizer-selected pruning is approximately $+0.2$ pp and not significant, while matched gold-document pruning gains 19.8 pp and gold-document biasing gains 8.6 pp. Similar localization accuracy across scales need not imply similar realizable gain.

Context length. At 120 documents, the frozen threshold rejects three configurations whose measured gains are $+2.4$, $+2.9$, and $+3.2$ pp and remain positive after clustered inference and multiple-comparison correction. Threshold transfer across context lengths can therefore forgo gains.

FITM transfer. In the original Vicuna setting, the middle-position point estimate is $+3.7$ pp with clustered $p = 0.083$. After transfer to Qwen, none of eight Qwen3-14B temperatures is positive. Three localization accuracies are 0.152, 0.403, and 0.250, below corresponding break-even points 0.832, 0.745, and 0.615.

G Operation Strength, Pruning, and Metric Audits

Table 5: Gold-document intervention headroom relative to baseline. Asterisks denote nominal question-clustered $p < 0.05$ before correction across the 11 groups. Llama-3-8B had no preregistered dose, so all scanned doses are shown.

Model and setting	Documents	Answer-document gain (pp)
Qwen3-4B	20	$+14.4^*$
Qwen3-8B	20	$+10.2^*$
Qwen3-8B (ensemble file)	20	$+11.4^*$
Qwen3-14B	20	$+20.5^*$
Qwen3-32B	20	$+20.0^*$
Qwen3-8B	10	$+12.7^*$
Qwen3-14B	10	$+18.7^*$
Qwen3-8B	30	$+8.0, p = 0.095$
Qwen3-14B	30	$+20.0^*$
Llama-3.1-8B	20	$+8.6^*$
Llama-3-8B	20	$+7.8/+10.0/+13.6^*$

Operation strength. Main-table attention-bias doses are primarily 2 or 3 for Qwen and smaller for Llama. Llama-3-8B has no preregistered dose, so all three scanned doses are retained. On Llama-3.1-8B, a high dose reduces normalized answer-substring hit accuracy by 13.6 pp; at dose 3, gold-document gain shrinks to $+1.4$ pp and is not significant.

Pruning scan. Pruning to one retained document is the strongest localizer-selected setting tested. Gains are $+4.8$, $+8.4$, $+16.2$, $+14.0$, and $+9.0$ pp for Qwen3-4B, Qwen3-8B, Qwen3-14B, Qwen3-32B, and Llama-3.1-8B, respectively. After question clustering and Benjamini–Hochberg correction, the 14B, 32B, and Llama-3.1-8B results remain significant; the 4B and 8B results do not. Additional scans retain 2, 3, 5, or 10 documents. They are treated as operation-strength checks rather than independently selected main results because the answer can be removed whenever targeting is wrong.

Output length. For Qwen3-8B, localizer-selected pruning to one document shortens output by 31.1 characters and improves normalized answer-substring hit accuracy by 10.9 pp; gold-document pruning shortens it by 32.3 characters and improves normalized answer-substring hit accuracy by 27.1 pp. For Llama-3.1-8B, the corresponding changes are -5.0 characters/$+5.3$ pp and -12.1 characters/$+31.9$ pp. Six operations with positive gain and significantly longer outputs were stratified by whether each output became longer. In every case, gain is larger on the not-longer branch; for example, Llama gold-document pruning to two documents gives $+5.2$ pp on the longer branch and $+19.1$ pp on the not-longer branch, while localizer-selected pruning to two gives -5.1 pp and

+7.1 pp. This audit makes a pure output-length explanation less plausible. Stricter full-output exact match is not evaluated; all answer-quality claims use normalized answer-substring hit accuracy.