# CALIBER: CONTEXT-AWARE LLM INTELLIGENT BOUNDARY ESTIMATION AND ROUTING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Large language models (LLMs) support increasingly long context windows, yet the amount of context they can reliably use often falls well below their nominal limits and varies substantially across models. Existing long-context evaluations primarily measure synthetic retrieval capacity or aggregate task performance, providing limited guidance for selecting an operational context budget. We propose Context-Aware LLM Inference via Boundary Estimation and Routing (**CALIBER**), a framework that empirically estimates a model-specific effective context threshold ($\tau_{\text{eff}}$) and uses it to organize long-context inference. CALIBER measures answer correctness, faithfulness, and consistency risk across context lengths and identifies a conservative boundary based on sustained quality degradation and consistency criteria. It then hierarchically indexes documents into threshold-bounded subcontexts, routes queries to relevant indices, processes selected subcontexts in parallel, and aggregates their partial responses. Across 14 LLMs, CALIBER reveals substantial nominal-to-effective context gaps and, on three frontier models, significantly improves average EQ compared to baselines.

## 1 INTRODUCTION

Retrieval-augmented generation (RAG) enables large language models (LLMs) to access external knowledge without encoding an entire corpus in their parameters (Lewis et al., 2020). Its effectiveness, however, depends critically on how documents are indexed and retrieved: relevant evidence omitted by the retriever cannot be recovered by the downstream generator. The expanding context windows of modern LLMs offer a complementary alternative by allowing more documents to be processed directly and reducing the information loss induced by restrictive retrieval. Recent comparisons reveal a quality–cost trade-off between retrieval and long-context inference (Li et al., 2024). Nevertheless, the ability to accept a long input does not imply the ability to use it reliably. LLMs exhibit positional sensitivity, reasoning degradation as context grows, and performance loss even when all relevant evidence is perfectly retrieved (Liu et al., 2024; Levy et al., 2024; Du et al., 2025). This raises a practical question for deployment: how much context can a target model reliably use before additional input becomes detrimental?

Existing long-context evaluations only partially address this question. RULER (Hsieh et al., 2024) evaluates context utilization primarily through controlled synthetic tasks, whereas HELMET (Yen et al., 2025) measures performance across diverse application-oriented workloads. Related work on the maximum effective context window further demonstrates that usable context can be substantially shorter than the advertised window and vary across tasks (Paulsen, 2025). These approaches provide valuable diagnostic evaluations, but do not translate multiple dimensions of response quality into a model- and workload-specific operating boundary or incorporate such a boundary into downstream inference.

To bridge this gap, we propose Context-Aware LLM Inference via Boundary Estimation and Routing (**CALIBER**), a framework that estimates an effective context boundary $\tau_{\text{eff}}$ for a target model and calibration workload. During offline calibration, CALIBER evaluates answer correctness, faithfulness, and estimated consistency risk across increasing context lengths and summarizes them using an Effective Quality (EQ) score. From the resulting quality trajectory, CALIBER detects both sustained degradation in overall response quality and excessive consistency risk, selecting the more conservative



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.

boundary as $\tau_{\text{eff}}$. This procedure translates observed long-context behavior into a practical inference budget without relying on the model's advertised context window.

After constructing the boundary $\tau_{\text{eff}}$, CALIBER uses $\tau_{\text{eff}}$ to organize documents into threshold-bounded subcontexts. And then at inference time, CALIBER routes queries to relevant indices, processes the selected subcontexts in parallel, and aggregates their partial responses into a final answer. This design retains efficient embedding-based filtering while complementing it with LLM-based routing, thereby reducing dependence on a single context-selection decision. Across 14 API-based and open-source LLMs, CALIBER identifies substantial gaps between nominal and effective context lengths. On three frontier models and five long-context benchmark categories, it improves average EQ by 19.2% over dense RAG and 8.6% over Self-RAG, with the largest gains on multi-document reasoning tasks.

## 2 PRELIMINARIES

### 2.1 PROBLEM SETUP

Let $M$ denote a target LLM with tokenizer $T_M$ and nominal context window $W_M$. For a query $q$ and an input context $C$, we denote the context length by $\ell_M(C) = |T_M(C)|$ and the model response by $r \sim p_M(\cdot \mid q, C)$. We consider a calibration workload $\mathcal{D}_{\text{cal}} = \{(q_i, C_i, y_i)\}_{i=1}^{N}$, where $y_i$ denotes the reference answer when available. For a context length $L$, let $Q_M(L; \mathcal{D}_{\text{cal}})$ denote the aggregate response quality of $M$ under the calibration workload at that length.

While $W_M$ specifies the maximum context that the model can technically accept, it does not necessarily characterize how much context the model can reliably utilize (Liu et al., 2024; Levy et al., 2024; Du et al., 2025). We therefore seek an *effective context boundary* $\tau_{\text{eff}} \leq W_M$ beyond which response quality exhibits persistent degradation. Importantly, this boundary may depend on both the target model and the workload; formally, $\tau_{\text{eff}} = \tau_{\text{eff}}(M, \mathcal{D}_{\text{cal}})$, which we abbreviate as $\tau_{\text{eff}}$ when clear from context. CALIBER instantiates $Q_M$ and estimates this boundary in Section 3.

### 2.2 CONTEXT BUDGETING IN RETRIEVAL-AUGMENTED INFERENCE

Given an external corpus $\mathcal{K} = \{d_j\}_{j=1}^{J}$, a standard RAG system assigns each document a relevance score $s(q, d_j)$ and selects relevant documents for generation (Lewis et al., 2020). We can generally view this process as constructing a query-specific context under a token budget $B$,

$$C_q(B) = \bigoplus_{d \in \mathcal{S}_q(B)} d, \qquad \ell_M(C_q(B)) \leq B,$$

where $\mathcal{S}_q(B) \subseteq \mathcal{K}$ is the selected document set and $\oplus$ denotes context concatenation. Conventional top-$k$ retrieval implicitly induces such a budget, while recent adaptive approaches vary whether, when, or how much information is retrieved according to passage relevance, query complexity, or inference cost (Asai et al., 2024; Jeong et al., 2024; Li et al., 2024; Taguchi et al., 2025).

These approaches improve how the available context budget is allocated, but the budget itself is typically determined from retrieval-side signals or computational considerations rather than the target LLM's observed long-context reliability. Consequently, a context can be highly relevant and remain within the nominal window $W_M$, yet still exceed the range that $M$ can reliably utilize. This mismatch motivates a complementary model-side calibration: estimating the usable context budget from the target model's response quality and incorporating it directly into downstream context construction. CALIBER provides such a mechanism by estimating $\tau_{\text{eff}}$ and using it as the operating budget for subsequent retrieval and inference.

## 3 METHOD

### 3.1 OVERVIEW

As discussed in Section 2, existing retrieval strategies can allocate contexts that remain within the nominal window $W_M$ but exceed the range that the target LLM can reliably utilize. In order to

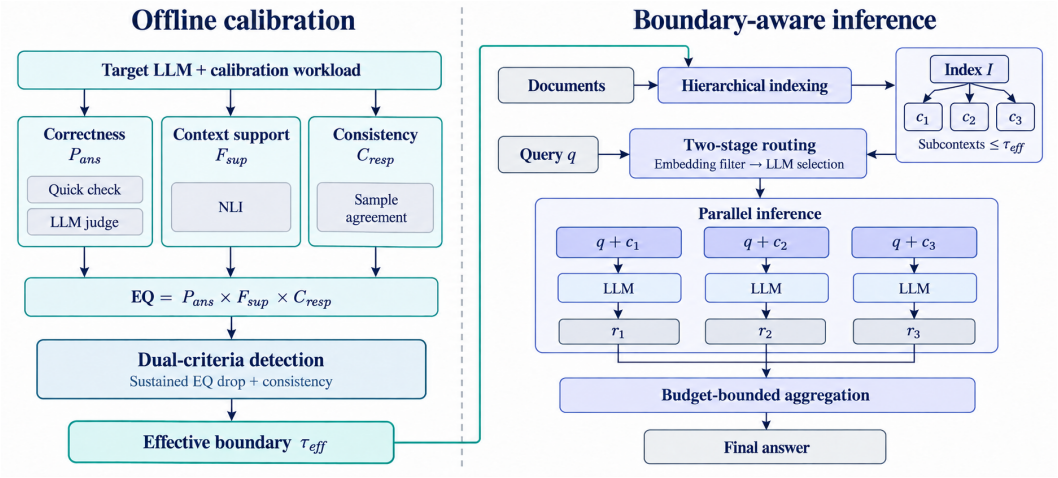



Figure 1: Overview of the CALIBER framework. **Left (Offline):** Given a target LLM and calibration workload, CALIBER measures answer correctness ($P_{\text{ans}}$), context support ($F_{\text{sup}}$), and response consistency ($C_{\text{resp}}$), and estimates $\tau_{\text{eff}}$ via dual-criteria detection. **Right (Online):** The calibrated boundary guides hierarchical context organization, two-stage routing, parallel inference, and budget-bounded aggregation.

address this issue, we introduce **CALIBER**, a two-phase framework that calibrates and exploits a model-specific effective context boundary $\tau_{\text{eff}}$. First, CALIBER measures response quality across context lengths and identifies $\tau_{\text{eff}}$ from persistent quality degradation. Second, it uses the calibrated boundary to organize and route threshold-bounded contexts, process them in parallel, and aggregate their responses. Together, CALIBER translates observed long-context reliability into a practical inference budget for retrieval-augmented generation.

### 3.2 Multidimensional Quality Calibration

To characterize how reliably a target LLM utilizes increasingly long contexts, we evaluate its responses along three complementary dimensions: answer correctness, contextual support, and response consistency. Each captures a distinct failure mode that may emerge as the input context grows.

**Answer correctness.** For a calibration example $(q, C, y)$ with model response $r$, we define $P_{\text{ans}} \in [0, 1]$ as the degree to which $r$ correctly answers $q$ with respect to the reference $y$. We first apply a task-specific deterministic quick check $\mathrm{QC}$ for directly verifiable cases, such as exact-match, binary-answer, and answer-containment tasks; otherwise, an LLM-based $\mathrm{Judge}$ evaluates semantic correctness following prior model-based evaluation protocols (Liu et al., 2023). Detailed evaluation rules and judge prompts are provided in Appendix B.

$$P_{\text{ans}} = \begin{cases} 1, & \text{if } \mathrm{QC}(y, r) = 1, \\ \mathrm{Judge}(q, y, r), & \text{otherwise.} \end{cases} \tag{1}$$

Here, by explicitly measuring answer correctness, CALIBER ensures that the calibrated boundary reflects degradation in actual task utility rather than changes in context-processing behavior alone.

**Contextual support.** Correctness alone does not indicate whether the generated answer is supported by the provided context. Following NLI-based document evaluation (Laban et al., 2022), we split input context $C$ into segments $\mathcal{G}(C)$ that fit within the input limit of the NLI evaluator and decompose the response $r$ into sentences $\mathcal{S}(r)$. For each pair $(g, s)$, the NLI model treats the context segment $g$ as a premise and the response sentence $s$ as a hypothesis, producing an entailment probability $p_{\text{ent}}(s \mid g)$. We then define

$$F_{\text{sup}} = \frac{1}{|\mathcal{S}(r)|} \sum_{s \in \mathcal{S}(r)} \max_{g \in \mathcal{G}(C)} p_{\text{ent}}(s \mid g). \tag{2}$$

Max-pooling allows supporting evidence to occur at arbitrary positions within the context. We interpret $F_{\text{sup}}$ as an NLI-based context-support score rather than a complete measure of factuality; segmentation and evaluator configurations are detailed in Appendix B. With contextual support, CALIBER distinguishes genuinely grounded responses from answers that may appear correct but are unsupported by the supplied evidence.

**Response consistency.** We further measure whether the model produces stable responses when conditioned on the same query and context. Given $m$ independently sampled responses $r^{(1)}, \ldots, r^{(m)}$, we define

$$C_{\text{resp}} = \frac{1}{\binom{m}{2}} \sum_{a<b} \text{BERTScore}\left(r^{(a)}, r^{(b)}\right), \tag{3}$$

following sampling-based self-consistency evaluation (Manakul et al., 2023; Zhang et al., 2019). $C_{\text{resp}}$ measures semantic agreement across generations rather than factual correctness, which is separately captured by $P_{\text{ans}}$. Sampling configurations are provided in Appendix B. By incorporating this complementary signal, CALIBER can capture increasing generation instability even when average correctness or contextual support has not yet substantially deteriorated.

**Effective Quality.** The three metrics capture complementary aspects of reliable long-context inference. We combine them into an Effective Quality score,

$$\text{EQ} = P_{\text{ans}} \cdot F_{\text{sup}} \cdot C_{\text{resp}}. \tag{4}$$

We adopt multiplicative aggregation because CALIBER treats reliability as a conjunctive requirement, which means that a response with a severe failure in correctness, contextual support, or consistency should not be considered reliable solely because the other dimensions remain strong. In contrast, additive aggregation can substantially compensate for a near-zero score with high values on the remaining dimensions, whereas the product increasingly penalizes such imbalance. Equivalently, for positive components, $\log \text{EQ}$ decomposes into the sum of their log-scores, making the aggregate particularly sensitive when any individual dimension approaches zero:

$$\log \text{EQ} = \log P_{\text{ans}} + \log F_{\text{sup}} + \log C_{\text{resp}}. \tag{5}$$

Consequently, EQ provides a single quality signal that responds to degradation in any of the three reliability dimensions as context length increases.

### 3.3 Effective Context Boundary Estimation

Given the quality measurements in Section 3.2, we next estimate the maximum context length at which the target model exhibits empirically reliable behavior on the calibration workload. A key challenge is that long-context quality profiles need not decrease monotonically: performance may temporarily deteriorate at an intermediate context length and recover at a later bucket. Consequently, defining the effective boundary as the first observed quality drop can be overly sensitive to local fluctuations, while selecting the globally best-performing context length does not provide a meaningful operating limit.

CALIBER therefore seeks a persistent transition in the quality regime rather than an isolated degradation. It combines sustained quality degradation with an absolute consistency criterion, allowing transient drops and subsequent recovery while detecting persistent deterioration. The resulting $\tau_{\text{eff}}$ is thus interpreted as an operational boundary beyond which reliability is no longer consistently maintained, rather than as the first context length at which performance decreases.

Let $L_1 < \cdots < L_K$ denote the ordered context-length buckets, and let $Q_k$ and $C_k$ denote the average EQ and response consistency within bucket $k$, respectively.

**Sustained quality degradation.** We first identify the context length at which overall response quality exhibits a persistent decline relative to the model's preceding quality level. Let $\overline{Q}_{a:b}$ denote the mean EQ from buckets $a$ through $b$. We directly define the degradation boundary as

$$\tau_{\text{drop}} = \min\left\{ L_k \,\middle|\, \Delta_k \geq \delta_{\text{drop}}, \max_{k<j\leq k+w} Q_j < \rho_{\text{rec}}\, \overline{Q}_{1:k} \right\}, \quad \Delta_k = \frac{\overline{Q}_{1:k} - \overline{Q}_{k+1:k+w}}{\overline{Q}_{1:k}},$$

Here, $\delta_{\text{drop}}$ specifies the minimum relative quality decline, $w$ is the number of subsequent buckets over which the decline must persist, and $\rho_{\text{rec}}$ specifies the fraction of the preceding quality level required to consider performance recovered. Thus, a local quality drop does not determine the boundary if performance recovers within the persistence window. This allows CALIBER to distinguish persistent context-induced degradation from transient fluctuations while retaining longer contexts at which reliable performance re-emerges.

**Consistency criterion.** While $\tau_{\text{drop}}$ captures sustained degradation in overall response quality, relative EQ changes do not directly specify an absolute level of response consistency. We therefore introduce a complementary consistency criterion. Let $C_k$ denote the average response consistency within context-length bucket $k$. We define

$$\tau_{\text{cons}} = \max\{L_k \mid C_k \geq \theta_{\text{cons}}\}, \tag{6}$$

where $\theta_{\text{cons}}$ specifies the consistency threshold used for calibration. Selecting the longest qualifying bucket accommodates recovery after a transient decline without assuming monotonic degradation. This criterion complements $\tau_{\text{drop}}$ by incorporating the observed consistency level alongside relative changes in EQ.

**Effective context boundary.** CALIBER combines the two estimates by selecting the smaller context budget:

$$\tau_{\text{eff}} = \min\{\tau_{\text{drop}}, \tau_{\text{cons}}\}. \tag{7}$$

This rule allows either sustained EQ degradation or the longest consistency-qualified length to limit the operating budget. Because consistency may be non-monotonic, the resulting budget does not guarantee that every bucket at or below $\tau_{\text{eff}}$ satisfies the consistency threshold. If no sustained EQ degradation is detected within the evaluated range, $\tau_{\text{drop}}$ is set to the largest evaluated context length, indicating that no qualifying decline was observed. The detector parameters $w$, $\delta_{\text{drop}}$, $\rho_{\text{rec}}$, and $\theta_{\text{cons}}$ are calibration hyperparameters, with their values and sensitivity analyses reported in the experimental setup and appendix.

We interpret $\tau_{\text{eff}}$ as an empirical operating boundary conditioned on the target model, calibration workload, and evaluation procedure. We evaluate its stability across calibration splits and evaluator configurations in Section 4.

### 3.4 Boundary-Aware Context Organization

Once $\tau_{\text{eff}}$ is calibrated, CALIBER uses it to constrain the source context supplied to each model invocation. Rather than modifying the underlying document corpus, CALIBER constructs an auxiliary hierarchical index over the original documents. Each parent index groups semantically related content for efficient routing, while its source content is maintained in one or more independently processable subcontexts bounded by $\tau_{\text{eff}}$.

**Semantic grouping.** For each incoming document $d$, CALIBER generates lightweight routing metadata consisting of a semantic description, its embedding $e_d$, and a keyword set $K_d$. Given an existing parent index $I_i$ with corresponding metadata $(e_i, K_i)$, we measure their compatibility as

$$s(d, I_i) = \lambda \cos(e_d, e_i) + (1 - \lambda) J(K_d, K_i), \tag{8}$$

where $J$ denotes Jaccard similarity (Jaccard, 1912) and $\lambda$ balances semantic and keyword similarity. CALIBER associates each document with all parent indices whose similarity exceeds a routing threshold, optionally retaining only the top-scoring parents to control index density. These associations are logical references to the original source rather than duplicated document copies. Otherwise, a new parent index is created. The routing metadata is updated as new content is assigned. Importantly, descriptions and keywords are used only for organization and routing, while downstream answer generation operates on the original source content.

**Boundary-aware subcontexts.** Semantic grouping is separated from the context budget imposed on individual model calls. We reserve a small budget $b_{\text{res}}$ for the query and inference instructions and constrain every source subcontext $c$ to satisfy $\ell_M(c) \leq \tau_{\text{eff}} - b_{\text{res}}$. When a document assigned to a parent can be accommodated within an existing subcontext, its source content is appended

without rewriting or summarization; otherwise, a new subcontext is created under the same parent. Documents that individually exceed this budget are split at natural structural boundaries with a small overlap to preserve local continuity. Thus, the parent index preserves semantic relationships across potentially large collections of documents, while each subcontext respects the empirically calibrated operating boundary. Exact grouping, overlap, and reserved-budget configurations are provided in Appendix B.

### 3.5 Routing and Parallel Inference

Given a query $q$, CALIBER first retrieves candidate parent indices using similarity between the query embedding and index descriptions, and an LLM router further selects those relevant to the query. Subcontexts referenced by the selected parents are then collected and deduplicated. Each selected subcontext $c$ is independently processed with the original query,

$$r_c \sim p_M(\cdot \mid q, c), \qquad c \in \mathcal{C}_{\text{sel}}, \tag{9}$$

while respecting the calibrated context budget, and the resulting responses are processed in parallel.

Routing and parallel inference serve complementary roles: routing suppresses irrelevant evidence, while parallelization allows evidence distributed across multiple contexts to be processed without concatenating them into a single overlong prompt. The resulting partial responses are combined by an LLM aggregator. To avoid reintroducing the long-context problem at this stage, CALIBER applies the same budget constraint to aggregation; when the partial responses exceed the available budget, they are recursively aggregated in budget-bounded groups until a single final response is obtained. Thus, no individual inference or aggregation call exceeds the calibrated effective boundary. Figure 1 summarizes the overall procedure.

## 4 Experiments

We evaluate CALIBER along three questions: (i) whether it identifies meaningful effective context boundaries across different LLMs, (ii) whether the calibrated boundaries improve retrieval-augmented inference, and (iii) which components contribute to the resulting performance gains.

### 4.1 Experimental Setup

**Benchmarks.** We construct the calibration workload from 1,833 examples drawn from NarrativeQA (Kočiský et al., 2018), LongBench (Bai et al., 2023), Multi-LexSum (Shen et al., 2022), FanOutQA (Zhu et al., 2024), and InfiniteBench (Zhang et al., 2024), covering long-document question answering, multi-document reasoning, summarization, and retrieval-oriented tasks. Inputs are partitioned into 15 context-length buckets ranging from 2K to 128K+ tokens. We estimate $\tau_{\text{eff}}$ using five-fold calibration, where the boundary is determined on four folds and evaluated on the held-out fold. For end-to-end evaluation, we consider FanOutQA, NarrativeQA, LongBench, InfiniteBench, and L-Eval (An et al., 2024). Dataset composition, length distributions, and split construction are provided in Appendix D.

**Models.** We evaluate 14 LLMs spanning both proprietary and open-weight model families. The six API models are GPT-5.2 (OpenAI, 2025), GPT-4o (OpenAI, 2024), Claude Sonnet 4.5 (Anthropic, 2025b), Claude Haiku 4.5 (Anthropic, 2025a), Gemini 2.5 Pro (Google DeepMind, 2025b), and Gemini 2.5 Flash (Google DeepMind, 2025a). The eight open-weight models are Llama-3.1-70B (Grattafiori et al., 2024), Mistral-NeMo-12B (Mistral AI Team, 2024), Jamba-1.5-Mini (Lenz et al., 2025), Phi-3-Mini-128K (Abdin et al., 2024), Qwen2.5-7B-Instruct and Qwen2.5-14B-Instruct (Qwen et al., 2024), Qwen3-14B (Yang et al., 2025), and Yi-34B-200K (Young et al., 2024). Hardware configurations and official model resources are provided in Appendix D.1; calibration and inference settings are described in Appendix D.

**Baselines and evaluation.** We compare CALIBER's estimated context boundaries with RULER (Hsieh et al., 2024) and evaluate end-to-end inference against dense-retrieval RAG (Lewis et al., 2020) and Self-RAG (Asai et al., 2024). All methods are evaluated using the multidimensional

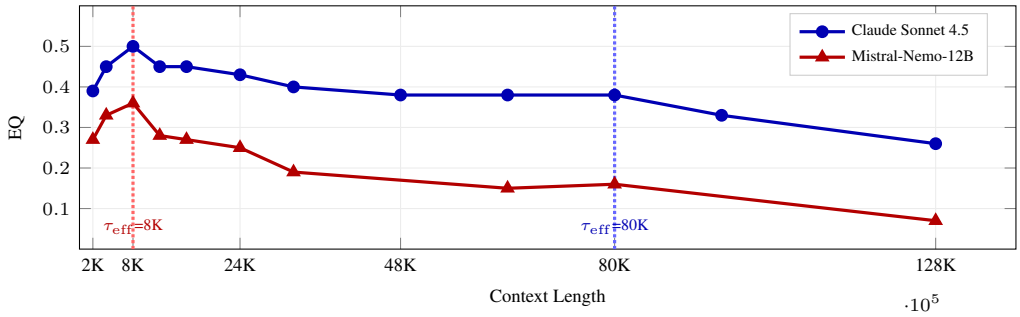



Figure 2: EQ across context lengths for two representative models. The empirically calibrated boundary can be substantially shorter than the advertised context window.

EQ metric defined in Section 3.2. Unless otherwise specified, the same evaluation protocol is applied across methods. Calibration hyperparameters, evaluator configurations, retrieval settings, and decoding parameters are provided in Appendix D.

### 4.2 Effective Context Calibration

Table 1 compares each model's advertised context window with its calibrated $\tau_{\text{eff}}$. The gap between nominal and effective context is neither uniform nor explained by advertised window size alone: $\tau_{\text{eff}}$ ranges from approximately 2% to 63% of the nominal window across the evaluated models. Notably, Gemini 2.5 Pro and Gemini 2.5 Flash both advertise 1M-token windows yet yield substantially different boundaries of 112K and 24K, respectively, while Jamba-Mini-1.5 retains a 128K effective boundary. This variation supports empirically calibrating the usable context of the target model rather than inferring it directly from nominal capacity.

Table 1: Advertised context windows and calibrated effective context boundaries.

| Type | Model | Window | $\tau_{\text{eff}}$ |
|---|---|---|---|
| API | GPT-5.2 | 256K | 112K |
| | Gemini 2.5 Pro | 1M | 112K |
| | GPT-4o | 128K | 80K |
| | Claude Sonnet 4.5 | 200K | 80K |
| | Claude Haiku 4.5 | 200K | 56K |
| | Gemini 2.5 Flash | 1M | 24K |
| Open | Jamba-Mini-1.5 | 256K | 128K |
| | Yi-34B-200K | 200K | 80K |
| | Qwen2.5-7B-Inst | 128K | 16K |
| | Qwen2.5-14B-Inst | 128K | 24K |
| | Llama-3.1-70B | 128K | 24K |
| | Phi-3-Mini-128K | 128K | 16K |
| | Qwen3-14B | 131K | 12K |
| | Mistral-Nemo-12B | 128K | 8K |

Figure 2 further illustrates the quality profiles of Claude Sonnet 4.5 and Mistral-Nemo-12B. Claude maintains relatively stable EQ over a broad intermediate range before degrading at longer contexts, whereas Mistral-Nemo exhibits a substantially earlier decline. The trajectories also show that quality need not decrease strictly monotonically with context length, supporting CALIBER's use of sustained degradation and recovery rather than terminating at the first local quality drop. Detailed decompositions of $P_{\text{ans}}$, $F_{\text{sup}}$, and $C_{\text{resp}}$ across context lengths are provided in Appendix E.2.

The resulting boundaries are stable across several calibration perturbations. Across five folds, the detected boundary varies by at most one bucket for 12 of 14 models. Replacing the default answer judge with Claude Sonnet 4.5 yields identical boundaries for two of three tested models and a one-bucket difference for the remaining model. In blinded human validation on five representative models, CALIBER exactly matches the human-estimated boundary for three models and remains within one adjacent bucket for all five. Full analyses of detector hyperparameters, judge choice, response sampling, and tokenizer sensitivity are reported in Appendix E.3.

### 4.3 End-to-End Performance

We next evaluate whether the calibrated boundary translates into improved downstream inference. Table 2 compares CALIBER with dense RAG and Self-RAG across three frontier models and five long-context benchmarks. CALIBER improves EQ over both baselines in all 15 model–benchmark combinations, yielding average gains of 19.2% over dense RAG and 8.6% over Self-RAG.

The largest relative improvements over dense RAG occur on InfiniteBench and FanOutQA, at approximately 25%. This pattern is consistent with CALIBER's intended advantage, which means

Table 2: EQ comparison across long-context benchmarks. R, S, and C denote RAG-Dense, Self-RAG, and CALIBER, respectively.

| Benchmark | Gemini 2.5 Pro | | | GPT-4o | | | Claude Sonnet 4.5 | | |
|---|---|---|---|---|---|---|---|---|---|
| | R | S | C | R | S | C | R | S | C |
| FanOutQA | .25 | .28 | **.31** | .22 | .25 | **.27** | .23 | .26 | **.29** |
| NarrativeQA | .34 | .37 | **.40** | .31 | .34 | **.36** | .33 | .36 | **.39** |
| LongBench | .38 | .41 | **.44** | .35 | .38 | **.40** | .37 | .40 | **.43** |
| InfiniteBench | .28 | .31 | **.35** | .25 | .28 | **.31** | .26 | .29 | **.33** |
| L-Eval | .36 | .39 | **.42** | .33 | .36 | **.38** | .35 | .38 | **.41** |
| **Average** | .32 | .35 | **.38** | .29 | .32 | **.34** | .31 | .34 | **.37** |

EQ is computed as $P_{\text{ans}} F_{\text{sup}} C_{\text{resp}}$; higher is better.

that when useful evidence is distributed across long or multiple contexts, boundary-constrained parallel processing can preserve broader evidence coverage without forcing a single model invocation beyond its empirically reliable range. The consistent improvement over Self-RAG further suggests that adaptive retrieval alone does not fully address model-side context-utilization limits. Complete baseline configurations and task-specific evaluation protocols are provided in Appendix D. For concrete examples of outputs generated by CALIBER and baselines, see the **Baseline implementations and adaptations** and Figure 3 in Appendix E.8.

### 4.4 Ablation and Robustness

Table 3 disentangles the contribution of context calibration from the remaining CALIBER pipeline. Replacing $\tau_{\text{eff}}$ with the larger RULER-derived budget reduces Loose Accuracy by 1.7–2.0% on Claude Haiku 4.5 and GPT-4o, while the direct baseline performs approximately 12% worse than full CALIBER. Thus, quality-based calibration provides a consistent benefit over a retrieval-oriented context estimate, while the larger Direct–CALIBER gap indicates that effective context management requires more than boundary estimation alone.

Table 3: Ablation on FanOutQA using Loose Accuracy (LA).

| Model | Configuration | $\tau$ | LA | Δ |
|---|---|---|---|---|
| Claude Haiku 4.5 | Direct | Full | .665 | -11.6% |
| | CALIBER ($\tau_R$) | 128K | .739 | -1.7% |
| | CALIBER ($\tau_{\text{eff}}$) | 56K | **.752** | – |
| GPT-4o | Direct | Full | .636 | -12.0% |
| | CALIBER ($\tau_R$) | 128K | .709 | -1.9% |
| | CALIBER ($\tau_{\text{eff}}$) | 80K | **.723** | – |
| *Component ablation on GPT-4o* | | | | |
| | w/o Orchestration | 80K | .674 | -6.8% |
| | w/o Description Routing | 80K | .694 | -4.0% |

The component ablation further clarifies this distinction. Removing parallel orchestration decreases performance by 6.8%, whereas removing description-based routing decreases it by 4.0%. These results suggest that $\tau_{\text{eff}}$ provides an effective operating constraint, while routing and orchestration preserve evidence coverage under that constraint. The relatively small gap between RULER- and CALIBER-based budgets additionally indicates that CALIBER remains robust to moderate boundary variation once inference is conducted within an appropriate context regime. Full threshold comparisons and additional calibration analyses are provided in Appendix E.5, with the detector sensitivity analysis below detailed further in Appendix E.3.

**Detector hyperparameter sensitivity.** We examine how the operating boundary responds to the four detector hyperparameters using the pooled calibration scores described in Appendix E.6. We jointly vary $\delta_{\text{drop}} \in \{0.10, 0.15, 0.20\}$, $w \in \{2, 3, 4\}$, $\rho_{\text{rec}} \in \{0.85, 0.90, 0.95\}$, and $\theta_{\text{cons}} \in \{0.35, 0.40, 0.45\}$, yielding 81 configurations per model. Comparing the 80 nondefault configurations with the default gives 1,040 comparisons across the 13 models with a defined default boundary. Boundary shifts are measured on the ordered 16-bucket grid.

Table 4 summarizes the results. Among 959 comparisons with defined boundaries, 689 (71.85%) remain within one bucket of the default. The mean absolute shift across all defined comparisons is 1.56 buckets; the other 81 comparisons contain an undefined boundary and are excluded from these statistics. Jamba-1.5-Mini and Mistral-NeMo retain a boundary within one bucket in 92.50% and 88.75% of comparable settings, respectively, while GPT-5.2 has a maximum shift of two buckets. Larger shifts for other models show that parameter sensitivity varies with the quality trajectory.

Table 4: Detector sensitivity across 80 nondefault configurations for each of the 13 models with a defined default boundary using pooled calibration scores. Within-one-bucket percentages and maximum shifts use defined comparisons.

| Model | Defined / 80 | Within 1 | Within 1 (%) | Max shift |
|---|---|---|---|---|
| Claude Haiku 4.5 | 80 | 47 | 58.75 | 6 |
| Claude Sonnet 4.5 | 80 | 59 | 73.75 | 7 |
| Gemini 2.5 Flash | 80 | 53 | 66.25 | 11 |
| Gemini 2.5 Pro | 80 | 44 | 55.00 | 3 |
| GPT-4o | 80 | 59 | 73.75 | 7 |
| GPT-5.2 | 80 | 62 | 77.50 | 2 |
| Jamba-1.5-Mini | 80 | 74 | 92.50 | 10 |
| Llama-3.1-70B | 80 | 59 | 73.75 | 9 |
| Mistral-NeMo-12B | 80 | 71 | 88.75 | 2 |
| Qwen2.5-7B-Instruct | 53 | 53 | 100.00 | 1 |
| Qwen3-14B | 53 | 53 | 100.00 | 0 |
| Phi-3-Mini-128K | 53 | 26 | 49.06 | 3 |
| Yi-34B-200K | 80 | 29 | 36.25 | 7 |
| **Total** | **959 / 1,040** | **689** | **71.85** | **11** |

## 5 Discussion

**Interpreting the effective context boundary.** Our results suggest that the advertised context window should not be interpreted as a reliable inference budget. CALIBER instead estimates $\tau_{\text{eff}}$ from observed response quality, providing an empirical operating boundary for downstream inference. Importantly, $\tau_{\text{eff}}$ is not an intrinsic property of a model: it depends on the target model, calibration workload, and evaluation protocol. This interpretation is consistent with recent evidence that increasing context length can degrade reasoning quality even when relevant information is successfully retrieved (Du et al., 2025). Accordingly, $\tau_{\text{eff}}$ is best viewed as a calibrated deployment-specific inference budget rather than a universal measure of model context capability.

**Multidimensional calibration.** The three components capture complementary failure modes of long-context inference: a response may be correct with respect to the reference yet unsupported by the provided context, or remain well supported while becoming unstable across repeated generations. Their joint use therefore provides a broader signal for detecting context-induced degradation than any individual metric alone. Importantly, $F_{\text{sup}}$ measures NLI-based contextual support and $C_{\text{resp}}$ measures semantic agreement across sampled responses; neither is intended as a standalone measure of factual reliability. Accordingly, EQ is best interpreted as an operational calibration signal for estimating $\tau_{\text{eff}}$, rather than a general-purpose factuality metric.

**Deployment trade-offs and limitations.** CALIBER introduces an offline calibration cost and may require multiple parallel model calls during inference. These costs are most readily amortized in repeated-query settings, while conventional RAG may remain preferable for low-volume applications. Moreover, because $\tau_{\text{eff}}$ is workload-dependent, substantial domain or task shifts may require recalibration. Its estimation can also depend on evaluator choice and calibration hyperparameters. We examine these sensitivities in Appendix E.3, while broader domain-specific calibration remains an important direction for future work.

## 6 Conclusion

We presented CALIBER, a framework that estimates a model- and workload-dependent effective context boundary from multidimensional response quality and uses it to guide context organization, routing, and parallel inference. Across 14 proprietary and open-weight LLMs, CALIBER reveals substantial gaps between advertised and empirically effective context lengths. On three frontier models across five long-context benchmarks, CALIBER improves average EQ by 19.2% over dense RAG and 8.6% over Self-RAG. These results demonstrate that explicitly calibrating how much context a model can reliably use provides a practical basis for more robust long-context inference.

## Ethics Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

## Disclosure of Generative AI Use

In this work, we used generative AI tools to assist with research execution, including implementing experimental pipelines, running experiments, organizing experimental results, and supporting qualitative analysis and interpretation of the results. We also used generative AI tools to help structure and draft parts of the experimental sections, including deciding how to organize and present experiments and which results to highlight in the main paper. In addition, generative AI tools were used to search for relevant information, create or modify scientific figures and other artifacts, suggest improvements to manuscript organization, and edit the manuscript to improve clarity and readability. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the final content of this work.

## Reproducibility statement

The calibration and inference procedures are described in Section 3. Hardware configurations and official model resources are provided in Appendix D.1. Dataset composition, evaluation metrics, calibration hyperparameters, and inference settings are described in Appendix D, with prompt templates in Appendix C.

## References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. URL https://arxiv.org/abs/2404.14219.

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. L-Eval: Instituting standardized evaluation for long context language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14388–14411, 2024. doi: 10.18653/v1/2024.acl-long.776. URL https://aclanthology.org/2024.acl-long.776/.

Anthropic. Claude haiku 4.5 system card. https://www-cdn.anthropic.com/7aad69bf12627d42234e01ee7c36305dc2f6a970.pdf, October 2025a.

Anthropic. Claude sonnet 4.5 system card. https://www-cdn.anthropic.com/963373e433e489a87a10c823c52a0a013e9172dd.pdf, September 2025b.

A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *Proceedings of the 12th International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Y. Bai, X. Lv, J. Zhang, H. Lyu, J. Tang, Z. Huang, Z. Du, X. Liu, A. Zeng, L. Hou, et al. LongBench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*, 2023. URL https://arxiv.org/abs/2308.14508.

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

Google DeepMind. Gemini 2.5 flash model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Flash-Model-Card.pdf, 2025a. Model card updated December 2025; accessed September 26, 2026.

Google DeepMind. Gemini 2.5 pro model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Pro-Model-Card.pdf, 2025b. Accessed September 26, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

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? *arXiv preprint arXiv:2404.06654*, 2024. URL https://arxiv.org/abs/2404.06654.

Paul Jaccard. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50, 1912. doi: 10.1111/j.1469-8137.1912.tb05611.x. URL https://doi.org/10.1111/j.1469-8137.1912.tb05611.x.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7036–7050, 2024. doi: 10.18653/v1/2024.naacl-long.389. URL https://aclanthology.org/2024.naacl-long.389/.

Ziyan Jiang, Xueguang Ma, and Wenhu Chen. LongRAG: Enhancing retrieval-augmented generation with long-context LLMs. *arXiv preprint arXiv:2406.15319*, 2024. URL https://arxiv.org/abs/2406.15319.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018. doi: 10.1162/tacl_a_00023. URL https://aclanthology.org/Q18-1023/.

P. Laban, T. Schnabel, P. N. Bennett, and M. A. Hearst. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177, 2022. doi: 10.1162/tacl_a_00453. URL https://aclanthology.org/2022.tacl-1.10/.

Alex Laitenberger, Christopher D. Manning, and Nelson F. Liu. Stronger baselines for retrieval-augmented generation with long-context language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 32559–32569, 2025. doi: 10.18653/v1/2025.emnlp-main.1656. URL https://aclanthology.org/2025.emnlp-main.1656/.

Barak Lenz, Opher Lieber, Alan Arazi, Amir Bergman, Avshalom Manevich, Barak Peleg, Ben Aviram, Chen Almagor, Clara Fridman, Dan Padnos, et al. Jamba: Hybrid transformer-Mamba language models. In *International Conference on Learning Representations*, volume 2025, pp. 67959–67984, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a9ed43fa31dc8b4a7d7a673d713dcb5f-Abstract-Conference.html.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15339–15353, 2024. doi: 10.18653/v1/2024.acl-long.818. URL https://aclanthology.org/2024.acl-long.818/.

P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. Retrieval augmented generation or long-context LLMs? a comprehensive study and hybrid approach. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 881–893, 2024. doi: 10.18653/v1/2024.emnlp-industry.66. URL https://aclanthology.org/2024.emnlp-industry.66/.

N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. 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/.

Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

P. Manakul, A. Liusie, and M. J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, 2023. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Mistral AI Team. Mistral NeMo. https://mistral.ai/news/mistral-nemo/, July 2024.

OpenAI. GPT-4o system card. https://openai.com/index/gpt-4o-system-card/, August 2024.

OpenAI. Update to GPT-5 system card: GPT-5.2. https://openai.com/index/gpt-5-system-card-update-gpt-5-2/, December 2025.

Norman Paulsen. Context is what you need: The maximum effective context window for real world limits of LLMs. *arXiv preprint arXiv:2509.21361*, 2025. URL https://arxiv.org/abs/2509.21361.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. Multi-LexSum: Real-world summaries of civil rights lawsuits at multiple granularities. In *Advances in Neural Information Processing Systems*, volume 35, pp. 13158–13173, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/552ef803bef9368c29e53c167de34b55-Abstract-Datasets_and_Benchmarks.html.

Chihiro Taguchi, Seiji Maekawa, and Nikita Bhutani. Efficient context selection for long-context QA: No tuning, no iteration, just adaptive-$k$. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20105–20130, 2025. doi: 10.18653/v1/2025.emnlp-main.1017. URL https://aclanthology.org/2025.emnlp-main.1017/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. HELMET: How to evaluate long-context language models effectively and thoroughly. In *Proceedings of the 13th International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.02694.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. Yi: Open foundation models by 01.AI. *arXiv preprint arXiv:2403.04652*, 2024. URL https://arxiv.org/abs/2403.04652.

T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*, 2019. URL https://arxiv.org/abs/1904.09675.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. $\infty$Bench: Extending long context evaluation beyond 100K tokens. *arXiv preprint arXiv:2402.13718*, 2024. URL https://arxiv.org/abs/2402.13718.

L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Andrew Zhu, Alyssa Hwang, Liam Dugan, and Chris Callison-Burch. FanOutQA: A multi-hop, multi-document question answering benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 18–37, 2024. doi: 10.18653/v1/2024.acl-short.2. URL https://aclanthology.org/2024.acl-short.2/.

# A Related Work

## A.1 Context Length Degradation

Recent studies provide empirical evidence that context length itself degrades reasoning performance. Du et al. (2025) show that LLMs exhibit performance degradation of 13.9–85% even with perfect retrieval, demonstrating that context length—independent of retrieval quality—limits reasoning capability. The FLenQA study (Levy et al., 2024) reports accuracy decline from 0.92 to 0.68 as context expands from 500 to 3,000 tokens. *Liu et al.* (Liu et al., 2024) reveal the "lost in the middle" phenomenon, where models struggle when relevant content appears in the middle of long contexts.

These findings suggest that models should not be assumed to effectively utilize their nominal context window, and the threshold at which quality degradation occurs must be empirically determined.

## A.2 Long-Context Evaluation

RULER evaluates long-context behavior through controlled synthetic tasks, including retrieval, multi-hop tracing, and aggregation (Hsieh et al., 2024). HELMET complements these evaluations with seven application-oriented categories and controllable context lengths (Yen et al., 2025). Both characterize performance across context lengths. CALIBER focuses on translating multidimensional quality measurements into a workload-specific operating budget and using that budget in downstream context organization and inference.

In this work, we evaluate answer correctness, contextual support, and response consistency. Our contextual-support score follows the NLI-based evaluation approach of SummaC (Laban et al., 2022). Our response-consistency score draws on sampling-based self-consistency evaluation in SelfCheck-GPT (Manakul et al., 2023) and measures pairwise semantic agreement using BERTScore (Zhang et al., 2019); it is not a direct estimate of hallucination rate. For answer correctness, we use task-specific LLM-based evaluation following prior model-based evaluation protocols (Zheng et al., 2023; Liu et al., 2023).

# B Metric Computation Details

**Answer Correctness ($P_{\text{ans}}$; LLM-as-Judge)**: Two-stage evaluation to handle verbose but correct answers:

*Stage 1 - Quick Check*: Deterministic evaluation for clear cases:

- **Retrieval**: Exact substring match (UUID, passkey)
- **Binary**: First word match (Yes/No, True/False)
- **List GT**: Any acceptable answer contained in response
- **Short GT**: Normalized containment check

*Stage 2 - LLM Judge*: For ambiguous cases, G-Eval style (Liu et al., 2023) chain-of-thought evaluation with task-specific guidelines. Uses GPT-4o as judge with 5-point rubric (1=wrong, 5=semantically equivalent), normalized to [0,1].

**Contextual Support ($F_{\text{sup}}$)**: SummaC-ZS approach using DeBERTa-small-long-nli. Context chunked into 1,400-token segments with 200-token overlap. For each response sentence, max-pool entailment scores across chunks, then mean across sentences.

**Response Consistency ($C_{\text{resp}}$)**: $m = 5$ responses at $T = 0.7$, evaluated using average pairwise BERTScore as defined in Equation 3.

# C Prompt Templates

## C.1 Description Generation

```
System: Generate a concise description
```

```
summarizing the document's main topics
and key information for retrieval.
Document: {content}
Output: {description}
```

### C.2 Router Prompt (Stage B)

```
System: Select relevant context indices.
Query: {query}
Indices: {descriptions with token counts}
Output: {"selected": [indices],
         "reasoning": "..."}
```

### C.3 LLM-as-Judge Prompt (Query Adequacy)

```
Query: {query}
Context: {context}
Response: {response}

Does the response adequately answer?
Output: {"adequate": bool, "gaps": [...]}
```

### C.4 LLM-as-Judge Prompt (Answer Precision)

```
You are evaluating answer correctness for
a {task_type} task.
Task Guidelines: {task_guidelines}
Question: {question}
Ground Truth: {ground_truth}
Model Response: {response}
Scoring: 5(1.0)=equivalent, 4(0.8)=minor
diff, 3(0.6)=partial, 2(0.4)=mostly wrong,
1(0.2)=wrong.
Output JSON: {"score": 1-5,
 "normalized_score": 0.2-1.0,
 "reasoning": "..."}
```

Task-specific guidelines: QA treats "No" $\equiv$ "No, because..." (semantic equivalence); multi-hop accepts verbose responses containing GT; summarization evaluates key info coverage, not length/style.

### C.5 Query Decomposition

```
Original: {query}
Gaps: {gaps}
Generate sub-queries to fill gaps.
Output: ["sub-query-1", ...]
```

## D Experimental Details

### D.1 Model Details

**Computing infrastructure.** We conducted open-weight model experiments across three GPU configurations: four NVIDIA RTX A6000 GPUs with 48 GB of memory per GPU, four NVIDIA A100 GPUs with 80 GB per GPU, and eight NVIDIA H100 GPUs. These configurations were used across experiments to run the open-weight models. Proprietary models were accessed through remote APIs; their inference hardware was managed by the service providers.

Table 5: System Hyperparameters

| Parameter | Value |
|---|---|
| Embedding threshold (Stage A) | 0.4 |
| Index merge threshold | 0.75 |
| Quality aggregation | Arithmetic mean |
| Max routing candidates $k$ | 10 |
| Chunk overlap ratio | 10% |
| Token buffer ($\tau_{buffer}$) | 1,000 |
| Calibration dataset size | 1,833 |
| Temp. (generation) | 0.0 |
| Temp. (hallucination) | 0.7 |
| *SummaC-ZS Faithfulness* | |
| NLI model | deberta-small-long-nli |
| NLI chunk size | 1,400 tokens |
| NLI chunk overlap | 200 tokens |
| Aggregation | Max-pool + Mean |

Table 6: Samples by Length Bucket (v2)

| Bucket | N | Task Distribution |
|---|---|---|
| 2K | 60 | QA: 30, Summ: 20, Multi: 10 |
| 4K–16K | 100 ea. | Balanced across 6 types |
| 24K–64K | 100 ea. | Narr: 35, Sum: 30, QA: 20, Multi: 15 |
| 80K–96K | 100 ea. | Narr: 40, Sum: 35, Multi: 15, QA: 10 |
| 112K | 101 | +1 FanOutQA |
| 128K | 157 | +57 FanOutQA |
| 128K+ | 352 | +252 FanOutQA (avg 172K) |
| **Total** | **1,833** | (+310 from FanOutQA) |

**Models and official resources.** Table 7 provides official model cards and release repositories for the six API models and eight open-weight models evaluated in this work. The links identify the corresponding model releases and documentation; they do not specify immutable checkpoint revisions or dated API snapshots for every run. The supplementary qualitative pilot uses the dated `gpt-4o-2024-08-06` snapshot, as specified in Appendix E.8.

**Inference implementation.** The local calibration implementation supports vLLM and Hugging Face Transformers backends using PyTorch. Generation and consistency-sampling settings are summarized in Table 9 and Appendix B.

### D.2 Calibration Workload and Evaluation Protocol

Our calibration workload contains 1,833 examples from five long-context benchmarks. Table 8 summarizes the source composition. The workload covers narrative and single-document QA, summarization, retrieval, and multi-hop reasoning, with contexts ranging from 2K to beyond 128K tokens. We partition examples into 15 length buckets at 2K, 4K, 8K, 12K, 16K, 24K, 32K, 40K, 48K, 56K, 64K, 80K, 96K, 112K, and 128K+.

For boundary estimation, we use 5-fold cross-validation. In each fold, $\tau_{\text{eff}}$ is estimated using four folds and evaluated exclusively on the held-out fold. End-to-end evaluation additionally includes L-Eval and uses held-out examples whenever its benchmark source overlaps with the calibration workload. This separates boundary estimation from downstream evaluation while allowing all calibration examples to contribute across folds.

Table 7: Official resources for the evaluated models. Model-card and repository names are clickable links.

| Model | Official resource |
|---|---|
| *API models* | |
| GPT-5.2 | OpenAI GPT-5.2 system-card update |
| GPT-4o | OpenAI GPT-4o system card |
| Claude Sonnet 4.5 | Anthropic Claude Sonnet 4.5 system card |
| Claude Haiku 4.5 | Anthropic Claude Haiku 4.5 system card |
| Gemini 2.5 Pro | Google DeepMind Gemini 2.5 Pro model card |
| Gemini 2.5 Flash | Google DeepMind Gemini 2.5 Flash model card |
| *Open-weight models* | |
| Llama-3.1-70B | `meta-llama/Llama-3.1-70B-Instruct` |
| Mistral-NeMo-12B | `mistralai/Mistral-Nemo-Instruct-2407` |
| Jamba-1.5-Mini | `ai21labs/AI21-Jamba-Mini-1.5` |
| Phi-3-Mini-128K | `microsoft/Phi-3-mini-128k-instruct` |
| Qwen2.5-7B-Instruct | `Qwen/Qwen2.5-7B-Instruct` |
| Qwen2.5-14B-Instruct | `Qwen/Qwen2.5-14B-Instruct` |
| Qwen3-14B | `Qwen/Qwen3-14B` |
| Yi-34B-200K | `01-ai/Yi-34B-200K` |

Table 8: Composition of the calibration workload.

| Source | Task | Length | N |
|---|---|---|---|
| NarrativeQA (Kočiský et al., 2018) | Narr. QA | 8K–128K+ | 569 |
| LongBench (Bai et al., 2023) | QA/Multi/Sum | 2K–32K | 399 |
| Multi-LexSum (Shen et al., 2022) | Summ. | 8K–128K+ | 358 |
| FanOutQA (Zhu et al., 2024) | Multi-hop | 112K–200K+ | 310 |
| InfiniteBench (Zhang et al., 2024) | QA/Retr. | 64K–128K+ | 197 |
| **Total** | | 2K–200K+ | **1,833** |

### D.3 Evaluation Metrics

For $P_{\text{ans}}$, deterministic quick checks are used for directly verifiable cases, including exact-match retrieval, binary answers, and answer containment. Remaining responses are evaluated by GPT-4o using task-specific semantic-correctness rubrics. We use Claude Sonnet 4.5 as an alternative judge in the evaluator-sensitivity analysis.

For $F_{\text{sup}}$, the input context is split into overlapping segments of at most 1,400 tokens. Each response sentence is treated as an NLI hypothesis and each context segment as a premise, and entailment probabilities are computed using a DeBERTa-based NLI evaluator. $F_{\text{sup}}$ averages the maximum entailment probability obtained for each response sentence.

For $C_{\text{resp}}$, we generate $m = 5$ responses at temperature $T = 0.7$ and compute their average pairwise BERTScore (Zhang et al., 2019). Detailed judge prompts, quick-check rules, NLI checkpoints, BERTScore configurations, and decoding parameters are provided with the reproducibility materials.

### D.4 Calibration and Inference Configuration

Table 9 summarizes the default configuration. The detector values are treated as calibration hyperparameters rather than fixed components of the method definition.

During indexing, documents are logically associated with parent indices satisfying the semantic-grouping criterion; source content itself is stored only once. At inference time, query-description similarity produces at most ten candidate parents, which are further filtered by the LLM router. Source subcontexts referenced by multiple selected parents are deduplicated before parallel generation. If partial responses exceed the aggregation budget, they are recursively aggregated in budget-bounded groups.

Table 9: Default CALIBER configuration.

| Component | Configuration |
|---|---|
| EQ drop $\delta_{\text{drop}}$ | 0.15 |
| Persistence window $w$ | 3 buckets |
| Recovery $\rho_{\text{rec}}$ | 0.90 |
| Consistency $\theta_{\text{cons}}$ | 0.40 |
| NLI segment length | $\leq$1,400 tokens |
| Response samples $m$ | 5 |
| Sampling temperature | 0.7 |
| Similarity weight $\lambda$ | 0.7 |
| Parent similarity threshold | 0.75 |
| Query similarity threshold | 0.40 |
| Maximum parent candidates | 10 |
| Reserved context budget | 1,000 tokens |
| Subcontext overlap | 10% |

Table 10: Distractor control on shared base examples using Qwen2.5-7B. $N$ is the number of available results in each condition; SD denotes the standard deviation of example-level F1.

| Distractors | $N$ | Mean token F1 | SD |
|---|---|---|---|
| 0 | 344 | .1675 | .2092 |
| 1 | 344 | .1088 | .1584 |
| 2 | 344 | .1067 | .1588 |
| 4 | 344 | .0992 | .1520 |
| 8 | 344 | .0979 | .1436 |

# E Additional Experiments

## E.1 Matched-Example Distractor Control

To complement the length-bucket profiles, we evaluate Qwen2.5-7B on the same base query–evidence examples while adding 0, 1, 2, 4, or 8 distractors. This design holds the underlying question and relevant evidence fixed across conditions. Table 10 reports the available-result counts and token-level F1 for each condition.

Mean F1 decreases across the five conditions, from .1675 without distractors to .0979 with eight distractors, a difference of .0696 (41.6% relative to the no-distractor mean). This supports the motivation that additional irrelevant context can impair answer quality even when the base evidence is retained. These are descriptive, condition-wise means with the reported counts, rather than a paired effect estimate over an identical set of completed outputs. The experiment does not separate token-count effects from distractor interference, evidence position, or possible input truncation, and does not directly validate EQ or the numerical value of $\tau_{\text{eff}}$.

## E.2 Detailed Calibration Results

Table 11 reports the three EQ components at an 8K context length. The results illustrate why no individual metric alone is sufficient to characterize response reliability. For example, several open-source models maintain high contextual support while exhibiting substantially lower response consistency.

Table 12 provides the complete quality decomposition for the two models visualized in Figure 2. Claude Sonnet 4.5 maintains relatively high correctness and contextual support across a broad range before degradation, whereas Mistral-Nemo-12B exhibits an earlier decline in EQ.

## E.3 Calibration Robustness

Across five calibration folds, the standard deviation of detected bucket indices is at most one bucket for 12 of 14 models. Replacing GPT-4o with Claude Sonnet 4.5 as the answer judge yields identical $\tau_{\text{eff}}$ for two of three evaluated models and a one-bucket difference for the third, indicating limited sensitivity to the choice of judge.

Table 11: Quality components at 8K context.

| Type | Model | $P_{\text{ans}}$ | $F_{\text{sup}}$ | $C_{\text{resp}}$ | EQ |
|---|---|---|---|---|---|
| API | GPT-5.2 | .93 | .83 | .71 | .54 |
| | Claude Sonnet 4.5 | .92 | .83 | .67 | .51 |
| | Gemini 2.5 Pro | .92 | .84 | .65 | .51 |
| | Gemini 2.5 Flash | .90 | .88 | .63 | .50 |
| | Claude Haiku 4.5 | .91 | .82 | .62 | .46 |
| | GPT-4o | .86 | .83 | .61 | .44 |
| Open | Llama-3.1-70B | .85 | .89 | .52 | .40 |
| | Mistral-Nemo-12B | .81 | .89 | .51 | .37 |
| | Jamba-Mini-1.5 | .75 | .92 | .48 | .33 |
| | Phi-3-Mini-128K | .73 | .91 | .38 | .25 |
| | Qwen2.5-7B-Inst | .81 | .74 | .41 | .24 |
| | Qwen3-14B | .86 | .64 | .43 | .24 |
| | Yi-34B-200K | .73 | .90 | .37 | .24 |
| | Qwen2.5-14B-Inst | .86 | .57 | .32 | .16 |

Table 12: Quality decomposition across context lengths for two representative models.

| | Claude Sonnet 4.5 | | | | Mistral-Nemo-12B | | | |
|---|---|---|---|---|---|---|---|---|
| Length | $P_{\text{ans}}$ | $F_{\text{sup}}$ | $C_{\text{resp}}$ | EQ | $P_{\text{ans}}$ | $F_{\text{sup}}$ | $C_{\text{resp}}$ | EQ |
| 2K | .94 | .73 | .57 | .39 | .82 | .76 | .43 | .27 |
| 4K | .90 | .79 | .64 | .45 | .77 | .86 | .50 | .33 |
| 8K | .91 | .82 | .66 | .50 | .80 | .89 | .50 | .36 |
| 12K | .93 | .82 | .58 | .45 | .76 | .87 | .43 | .28 |
| 16K | .88 | .85 | .59 | .45 | .67 | .90 | .44 | .27 |
| 24K | .87 | .87 | .57 | .43 | .69 | .90 | .40 | .25 |
| 32K | .85 | .86 | .54 | .40 | .60 | .91 | .35 | .19 |
| 48K | .89 | .86 | .49 | .38 | .66 | .85 | .27 | .15 |
| 64K | .88 | .87 | .49 | .38 | .52 | .91 | .32 | .15 |
| 80K | .91 | .87 | .48 | .38 | .60 | .89 | .31 | .16 |
| 96K | .87 | .85 | .44 | .33 | – | – | – | – |
| 128K | .78 | .57 | .58 | .26 | .65 | .21 | .50 | .07 |

For response consistency, we evaluate $m \in 3, 5, 7$ and $T \in 0.5, 0.7, 1.0$. Using three responses increases the variance of the pairwise similarity estimate by approximately 8%, whereas seven responses provides little change relative to five at roughly 40% additional generation cost. Lower temperature reduces response diversity, while $T = 1.0$ introduces greater sampling variability; we therefore use $m = 5$ and $T = 0.7$ by default.

We additionally validate the automatically estimated boundaries with two NLP-experienced annotators on five representative models. Inter-annotator agreement reaches Cohen's $\kappa = 0.78$. Automated and human boundaries match exactly for three of five models and differ by at most one adjacent bucket for all five.

**Detector hyperparameter sensitivity.** Using the supplementary pooled-score protocol described in Appendix E.6, we vary $\delta_{\text{drop}} \in \{0.10, 0.15, 0.20\}$, $w \in \{2, 3, 4\}$, $\rho_{\text{rec}} \in \{0.85, 0.90, 0.95\}$, and $\theta_{\text{cons}} \in \{0.35, 0.40, 0.45\}$. The Cartesian grid contains 81 settings per model. Excluding the default setting leaves 1,120 model–setting comparisons across 14 models. We measure absolute shifts using the ordered 16-bucket grid specified in that protocol; a comparison is defined only when both the default and perturbed boundaries exist.

Of these comparisons, 959 have defined boundaries, and 689/959 (71.85%) remain within one bucket of the default. Table 4 reports the 13 models with a defined default boundary; Qwen2.5-14B-Instruct has an undefined default boundary and contributes the 80 comparisons omitted from the main-text table. The remaining 161/1,120 (14.38%) include an undefined boundary and are excluded from the conditional percentage, rather than counted as zero shifts. Thus, 689/1,120 (61.52%) of all comparisons are both defined and within one bucket. The mean absolute shift among defined comparisons is 1.56 buckets. Sensitivity is model-dependent: most comparable settings yield small shifts, but some models exhibit larger changes or no qualifying consistency boundary. These results characterize parameter sensitivity separately from the five-fold analysis above; agreement with the default is not evidence of downstream optimality.

Table 13: Workload composition diagnostics under the archived proxy-metric protocol. Entries are $\tau_{\text{proxy}}$; – denotes an unreported task subset, and 128K+ is an open-ended bucket label.

| Model | Mixed | QA | Multi-hop | Summ. | Narr. QA |
|---|---|---|---|---|---|
| Jamba-1.5-Mini | 128K | 16K | 2K | – | 8K |
| GLM4-9B-1M | 2K | 24K | 2K | 32K | 32K |
| Llama-3.1-70B | 96K | 16K | 128K+ | 40K | 96K |
| Mistral-NeMo-12B | 8K | 4K | 16K | – | 8K |
| Qwen2.5-14B-Instruct | 4K | 16K | 2K | 64K | – |
| Qwen2.5-7B-Instruct | 4K | 8K | 8K | 24K | 8K |
| Qwen3-14B | 2K | 4K | 12K | 12K | – |

Table 14: Dataset-omission diagnostics for GPT-4o under the proxy-metric protocol. Shifts are measured on the ordered 16-bucket grid relative to the full-workload boundary.

| Omitted dataset | $\tau_{\text{proxy}}$ | Bucket shift |
|---|---|---|
| None (full workload) | 16K | 0 |
| FanOutQA | 16K | 0 |
| InfiniteBench | 16K | 0 |
| L-Eval | 16K | 0 |
| LongBench | 128K+ | 11 |
| Multi-LexSum | 16K | 0 |
| NarrativeQA | 64K | 6 |

### E.4 Calibration Workload Composition

The sensitivity analysis above varies detector parameters while holding the calibration scores fixed. We next examine how changing the calibration workload affects the estimated boundary. These supplementary diagnostics use an archived proxy-metric protocol with token F1 as the correctness input and word-overlap-based consistency. The archived calibration report also identifies proxy contextual-support scores. We denote the resulting boundary by $\tau_{\text{proxy}}$ to distinguish it from both the main five-fold estimates and the pooled-score analysis in Appendix E.6. The comparisons below describe sensitivity within this protocol; their absolute boundaries are not replacements for Table 1.

**Task composition.** Table 13 compares mixed-task calibration with task-specific subsets for the seven models with available detailed results. For Qwen2.5-7B, the boundary changes from 4K under mixed-task calibration to 8K for QA and 24K for summarization. For Mistral-NeMo, the corresponding mixed-task, QA, and multi-hop boundaries are 8K, 4K, and 16K. The direction of change varies across models and tasks, consistent with interpreting the boundary as conditional on the calibration workload.

Task-specific subsets can differ in sample count and observed context-length coverage. Consequently, these shifts do not isolate task difficulty from changes in the length distribution, nor do they establish that task-specific calibration is more accurate. They motivate reporting the workload and its bucket coverage alongside any operating boundary.

**Dataset omission.** Table 14 complements the task-level comparison with a dataset-omission analysis for GPT-4o. Starting from the archived full-workload boundary of 16K, removing LongBench changes the boundary to 128K+, and removing NarrativeQA changes it to 64K. The other reported omissions leave the boundary at 16K. These changes show that the boundary can depend strongly on which datasets contribute calibration observations.

The archived omission analysis includes an L-Eval condition and therefore should not be identified with the five-source calibration split in Appendix D. Its summary does not provide per-dataset retained counts or performance on the omitted dataset. We interpret it as a composition-sensitivity diagnostic, not as a completed test of cross-task generalization. In particular, movement to the open-ended upper bucket does not by itself establish reliable utilization at all longer lengths.

Table 15: Context-boundary and quality comparison. $\tau_R$ denotes the RULER-derived context estimate and $\tau_{\text{eff}}$ the CALIBER boundary.

| | Model | Window | $\tau_R$ | $\text{EQ}_R$ | $\tau_{\text{eff}}$ | $\text{EQ}_{\text{long}}$ | $\text{EQ}_{\text{eff}}$ | $\Delta_R$ | $\Delta_{\text{long}}$ |
|---|---|---|---|---|---|---|---|---|---|
| API | GPT-5.2 | 256K | 128K+ | .26 | 112K | .26 | .44 | +69% | +69% |
| | Gemini 2.5 Pro | 1M | 128K+ | .26 | 112K | .26 | .39 | +50% | +50% |
| | GPT-4o | 128K | 64K | .32 | 80K | .10 | .35 | +9% | +250% |
| | Claude Sonnet 4.5 | 200K | – | – | 80K | .11 | .39 | – | +255% |
| | Claude Haiku 4.5 | 200K | – | – | 56K | .08 | .34 | – | +325% |
| | Gemini 2.5 Flash | 1M | 128K+ | .23 | 24K | .23 | .41 | +78% | +78% |
| Open | Jamba-Mini-1.5 | 256K | 256K | .43 | 128K | .43 | .43 | +0% | +0% |
| | Yi-34B-200K | 200K | 128K | .17 | 80K | .13 | .18 | +6% | +38% |
| | Qwen2.5-7B-Inst | 128K | 128K | .15 | 16K | .15 | .19 | +27% | +27% |
| | Qwen2.5-14B-Inst | 128K | 128K | .19 | 24K | .19 | .19 | +0% | +0% |
| | Llama-3.1-70B | 128K | 128K | .05 | 24K | .06 | .32 | +540% | +433% |
| | Phi-3-Mini-128K | 128K | 128K | .18 | 16K | .18 | .22 | +22% | +22% |
| | Qwen3-14B | 131K | 128K | .24 | 12K | .24 | .24 | +0% | +0% |
| | Mistral-Nemo-12B | 128K | 16K | .28 | 8K | .09 | .37 | +32% | +311% |

$\text{EQ}_{\text{long}}$: EQ at the longest evaluated bucket; $\text{EQ}_{\text{eff}}$: EQ at $\tau_{\text{eff}}$.

### E.5 Threshold Comparison and Additional Ablations

Table 15 compares CALIBER's effective boundary with RULER-derived context estimates and long-context evaluation at the longest available bucket. The comparison highlights that a longer admissible or retrieval-effective context does not necessarily correspond to higher multidimensional response quality.

The RULER comparison shows that retrieval-oriented effective lengths can differ substantially from quality-based operating boundaries. At the same time, the relatively modest downstream gap between RULER- and CALIBER-based budgets on FanOutQA indicates that the exact boundary is not the sole source of CALIBER's improvement; the orchestration mechanism is separately responsible for preserving evidence across bounded calls.

### E.6 Supplementary Calibration Ablations

**Shared analysis protocol.** To examine how the detector criteria and quality signals affect boundary selection, we additionally analyze archived example-level scores pooled within each model's observed length buckets. These supplementary diagnostics are separate from the five-fold boundary estimates in Table 1. For reproducibility, the analysis uses the stored answer-judge score, substitutes token F1 when that score is missing, and combines it with stored contextual support and response consistency. We denote this analysis-specific correctness input by $\tilde{P}$ to distinguish its missing-score rule from Equation 1. All nonempty buckets are retained from the ordered grid 2K, 4K, 8K, 12K, 16K, 24K, 32K, 40K, 48K, 56K, 64K, 80K, 96K, 112K, 128K, and 128K+. The persistence window spans subsequent observed buckets; 128K+ is an open-ended bucket label. We use the default detector settings in Table 9, except in the hyperparameter sweep. When no bucket passes the consistency threshold, the consistency and dual boundaries are undefined. When no sustained drop is detected, the drop boundary returns the largest observed bucket; this endpoint return does not establish degradation beyond that length.

**Detection-criterion ablation.** Table 16 compares the sustained-drop criterion, the consistency criterion, and their minimum using the same $\tilde{P}F_{\text{sup}}C_{\text{resp}}$ trajectory. Among the 13 models with defined dual boundaries, the criteria differ for nine: the drop criterion selects a shorter boundary for seven models, whereas consistency selects a shorter boundary for Qwen2.5-7B and Phi-3-Mini. They agree for four models. Thus, neither criterion uniformly subsumes the other on these scores, supporting their complementary roles in boundary selection. This comparison establishes different detector behavior, rather than a downstream performance advantage of the dual criterion.

Table 16: Detection-criterion ablation on pooled archived scores. All columns use the same observed-bucket inclusion policy.

| Model | Drop only | Consistency only | Dual |
|---|---|---|---|
| Claude Haiku 4.5 | 64K | 128K+ | 64K |
| Claude Sonnet 4.5 | 80K | 128K+ | 80K |
| Gemini 2.5 Flash | 16K | 128K+ | 16K |
| Gemini 2.5 Pro | 128K+$^{\dagger}$ | 128K+ | 128K+$^{\dagger}$ |
| GPT-4o | 80K | 128K+ | 80K |
| GPT-5.2 | 128K+$^{\dagger}$ | 128K+ | 128K+$^{\dagger}$ |
| Jamba-1.5-Mini | 128K$^{\dagger}$ | 128K | 128K$^{\dagger}$ |
| Llama-3.1-70B | 80K | 128K+ | 80K |
| Mistral-NeMo-12B | 8K | 128K+ | 8K |
| Qwen2.5-14B-Instruct | 24K$^{\dagger}$ | – | – |
| Qwen2.5-7B-Instruct | 128K+$^{\dagger}$ | 8K | 8K |
| Qwen3-14B | 12K$^{\dagger}$ | 12K | 12K$^{\dagger}$ |
| Phi-3-Mini-128K | 16K | 4K | 4K |
| Yi-34B-200K | 80K | 128K+ | 80K |

†: largest observed bucket returned without a detected sustained drop; –: no qualifying consistency boundary. Qwen3-14B has observations only through 12K in this analysis.

Table 17: Quality-signal ablation under a shared consistency criterion. Each entry is the resulting dual boundary. $\tilde{P}$ follows the supplementary missing-score rule; $F = F_{\text{sup}}$ and $C = C_{\text{resp}}$.

| Model | $\tilde{P}$ | $\tilde{P}F$ | $\tilde{P}C$ | $\tilde{P}FC$ |
|---|---|---|---|---|
| Claude Haiku 4.5 | 128K+$^{\dagger}$ | 128K+$^{\dagger}$ | 16K | 64K |
| Claude Sonnet 4.5 | 128K+$^{\dagger}$ | 128K+$^{\dagger}$ | 16K | 80K |
| Gemini 2.5 Flash | 128K+$^{\dagger}$ | 128K+$^{\dagger}$ | 16K | 16K |
| Gemini 2.5 Pro | 128K+$^{\dagger}$ | 128K+$^{\dagger}$ | 128K+$^{\dagger}$ | 128K+$^{\dagger}$ |
| GPT-4o | 128K+$^{\dagger}$ | 96K | 128K+$^{\dagger}$ | 80K |
| GPT-5.2 | 128K+$^{\dagger}$ | 128K+$^{\dagger}$ | 128K+$^{\dagger}$ | 128K+$^{\dagger}$ |
| Jamba-1.5-Mini | 128K$^{\dagger}$ | 128K$^{\dagger}$ | 128K$^{\dagger}$ | 128K$^{\dagger}$ |
| Llama-3.1-70B | 96K | 96K | 8K | 80K |
| Mistral-NeMo-12B | 12K | 48K | 8K | 8K |
| Qwen2.5-14B-Instruct | – | – | – | – |
| Qwen2.5-7B-Instruct | 8K | 8K | 8K | 8K |
| Qwen3-14B | 12K$^{\dagger}$ | 12K$^{\dagger}$ | 12K$^{\dagger}$ | 12K$^{\dagger}$ |
| Phi-3-Mini-128K | 4K | 4K | 4K | 4K |
| Yi-34B-200K | 128K+$^{\dagger}$ | 96K | 128K+$^{\dagger}$ | 80K |

†: observed upper endpoint returned without a detected sustained drop; –: undefined because no bucket passes the shared consistency threshold.

**Quality-signal ablation.** We next replace the quality trajectory with $\tilde{P}$, $\tilde{P}F_{\text{sup}}$, or $\tilde{P}C_{\text{resp}}$, while retaining the same absolute consistency criterion in every variant (Table 17). For Claude Haiku 4.5, Claude Sonnet 4.5, Gemini 2.5 Flash, and GPT-4o, the $\tilde{P}$ trajectory returns the observed upper endpoint, whereas the full product yields a shorter boundary. The additional quality signals therefore affect boundary selection beyond the correctness trajectory alone. Their effect is not uniformly toward shorter boundaries: adding support to $\tilde{P}C_{\text{resp}}$ moves the boundary from 16K to 64K for Haiku and from 16K to 80K for Sonnet. This is a quality-signal ablation under a shared consistency criterion; it does not compare aggregation functions or test removal of correctness via an $F_{\text{sup}}C_{\text{resp}}$ variant.

**Dependence on score availability.** The supplementary results are conditional on the stated missing-score and bucket-inclusion rules. Excluding missing-judge examples instead of substituting token F1 changes the full-product boundary from 64K to 24K for Haiku and from 16K to 24K for Flash. Some retained buckets are sparse: Llama-3.1-70B has four examples at 112K, while Mistral-NeMo and Qwen2.5-7B each have a one-example bucket at 112K and 128K+, respectively. Requiring at least five examples per bucket changes Llama's $\tilde{P}$ and $\tilde{P}F$ boundaries from 96K to the observed upper endpoint. These diagnostics support distinct contributions of the criteria and signals, while emphasizing that the resulting numerical boundaries depend on the calibration data and evaluation procedure.

Table 18: Progressive inference pilot with Qwen2.5-7B ($N = 100$). Differences are absolute token-F1 changes relative to direct inference.

| Configuration | Mean token F1 | ΔF1 |
|---|---|---|
| Direct | .1150 | – |
| + Context budgeting (16K) | .1615 | +.0465 |
| + Parallel processing and aggregation | .1506 | +.0356 |

Table 19: Separate direct-versus-budgeted Qwen2.5-7B pilot ($N = 100$). Latencies are reproduced as reported for this run.

| Configuration | Input budget | Mean token F1 | Latency (s) |
|---|---|---|---|
| Direct | 32,768 | .0980 | 85.1 |
| Budgeted (reported as CALIBER) | 16,000 | .1548 | 136.8 |

### E.7 Supplementary Budgeting and Inference Comparisons

To complement the frontier-model results, we report two Qwen2.5-7B pilot experiments run on an A100 system. Each uses 100 examples and a 16,000-token operating budget. We report token F1, since the archived evaluation protocol differs from the multidimensional EQ protocol in Section 3.2. The two runs are summarized separately: their available records do not establish that they use identical example sets or identical inference configurations.

**Progressive inference components.** Table 18 compares direct inference, the addition of context budgeting, and the subsequent addition of parallel processing and aggregation. Budgeting increases mean F1 from .1150 to .1615, an absolute gain of .0465. Adding parallel processing and aggregation yields .1506, which remains .0356 above direct inference but is .0109 below budgeting alone. Thus, budgeting helps in this pilot, while adding an aggregation stage does not automatically improve the final answer. This result complements the main component analysis by illustrating that the benefit of orchestration depends on the model and inference setting. The pilot does not separately isolate hierarchical indexing, description-based routing, or recursive budget-bounded aggregation.

**Direct inference and the budgeted run.** In the second pilot, the run labeled CALIBER in the archived report uses a 16K budget, while direct inference permits inputs up to 32,768 tokens (Table 19). Mean F1 increases from .0980 to .1548, a gain of .0568, or 5.68 percentage points on a 0–100 scale. The reported latency increases from 85.1 to 136.8 seconds, approximately 1.61 times the direct-inference value. This comparison illustrates a quality–latency trade-off in the evaluated setting rather than a speed improvement.

The archived description characterizes the budgeted condition as context truncation and does not fully specify its routing and aggregation configuration. It also notes a short-context bias in the latency comparison. Accordingly, this pilot provides supplementary evidence for context budgeting, but does not establish the efficiency of the full CALIBER pipeline or isolate calibration gains from additional computation. A budget-matched comparison with recorded input/output tokens, inference calls, and concurrency remains necessary for that purpose.

### E.8 Qualitative Comparisons under Bounded Input-Token Ratios

**Case selection and shared setup.** We conduct a supplementary qualitative pilot on three FanOutQA questions requiring attributes from multiple documents: the page counts of five books, the career rushing attempts of five NFL players, and the full birth dates of five notable people from West Reading. The cases were purposively selected from questions with successful archived CALIBER responses; they are not a random evaluation sample. For this pilot, standalone CALIBER was executed on a shared corpus comprising the 19 source articles associated with these questions, and its recorded predictions were reused for the subsequent baseline comparisons. All generation and critic calls use `gpt-4o-2024-08-06` at temperature zero. Final answers allow up to 4,096 output tokens. Retrieval uses `text-embedding-3-small` and the `o200k_base` tokenizer. Baseline retrieval-context budgets were fixed at 8K, 16K, and 40K tokens for the book, NFL, and birth-date

cases, respectively, before collecting their responses. Questions and reference answers were excluded from index-description generation, and reference answers were not supplied to retrieval or answer generation.

**Standalone CALIBER configuration.** Each original source bundle serves as a parent index with a generated description. Description-embedding routing selects a parent, and source subcontexts use a reused 80K operating boundary with a 1K instruction reserve and 10% overlap when splitting is required. The book and NFL cases each use one generation call; the birth-date case uses two source-processing calls followed by aggregation. Each of the latter calls remains below 80K input tokens. No other RAG framework is combined with CALIBER. In these cases, only one parent passes the embedding filter, so the LLM routing-selection stage is skipped. This pilot therefore does not isolate automatic semantic grouping, multi-parent indexing, or the independent effect of boundary calibration.

**Baseline implementations and adaptations.** Dense RAG uses 512-token chunks with 51-token overlap and cosine-ranked retrieval within the budget. Our LongRAG adaptation (Jiang et al., 2024) executes the official long-reader prompt on retrieved source units of at most 4K tokens. It substitutes the shared embedding model for BGE/Tevatron retrieval, does not perform Wikipedia hyperlink grouping, and omits single-entity short-answer extraction for these multi-item questions. Our DOS RAG adaptation (Laitenberger et al., 2025) executes the official sentence-based chunking routine with a 100-token target and the retrieval routine that restores source order. Its original embedding model and tokenizer are replaced with the shared configuration. Dense RAG and DOS RAG use the same answer prompt as CALIBER's source processor.

The Self-RAG-style baseline uses GPT-4o critics derived from the official Self-RAG repository (Asai et al., 2024), rather than its trained reflection-token model. An initial context-level relevance gate rejected the retrieved context for all three questions. After observing this behavior, we separately specified and ran a stronger retrieval-retry variant on all three cases. Starting from the Dense answer, it generates up to six missing-evidence queries, combines the original and generated queries' ranked hits in round-robin order, deduplicates them, and retains the same retrieval-context budget. It then answers, evaluates support and utility, and permits one revision if either score is below four on a five-point scale. Relevance is recorded after retry but does not discard the entire context. Table 20 and Figure 3 report this stronger variant. It does not implement learned reflection-token probabilities or segment-wise beam search; these comparisons are common-backbone adaptations, not complete reproductions of the original methods.

**Attribute scoring and token accounting.** We compare each response against the dataset reference and source evidence, counting complete correct entity–attribute pairs out of five. A birth date requires month, day, and year; a year-only answer is incomplete. The birthplace is checked separately and is not included in the five-date denominator. These descriptive counts are not EQ scores. Input usage sums API-reported prompt tokens across all online generation, critic, revision, and aggregation calls, including cached tokens. The Dense answer reused to initialize the retry baseline is charged to that method as well. Offline indexing and embedding costs are excluded. For CALIBER input usage $I_C$ and baseline usage $I_B$, the pairwise ratio is $\max(I_C, I_B)/\min(I_C, I_B)$. All such ratios are below 3.32, but the methods are not compute-matched.

**Observed answer differences.** The birth-date case provides the clearest contrast. CALIBER returns West Reading and all five complete dates. Dense RAG and LongRAG return the borough but no complete dates. DOS RAG returns Chad Henne's full date and only the birth years for Michael W. Kirst and Andrew H. Knoll. The retry baseline recovers full dates for John Fetterman and Chad Henne, but only years for the other three people, despite using 241,574 input tokens compared with CALIBER's 131,746. For book page counts, CALIBER recovers all five attributes, compared with zero, two, and three for Dense RAG, LongRAG, and DOS RAG; the retry baseline also recovers all five and therefore ties CALIBER on answer completeness. For NFL attempts, Dense RAG adds Frank Gore in place of Eric Shelton. The retry baseline removes Gore but still omits Shelton's eight attempts. LongRAG and DOS RAG match CALIBER on all five players, so this case does not demonstrate an advantage over those two methods.

Table 20: Qualitative FanOutQA comparisons. Correct counts complete entity–attribute pairs out of five; input/output totals include all online calls. Asterisks denote the adaptations described in the text. Ratio is the larger of CALIBER's and the baseline's input totals divided by the smaller.

| Case | Method | Correct | Input | Output | Calls | Ratio |
|---|---|---|---|---|---|---|
| Book pages | Dense RAG | 0/5 | 8,008 | 22 | 1 | 2.67 |
| | LongRAG* | 2/5 | 8,030 | 86 | 1 | 2.66 |
| | DOS RAG* | 3/5 | 8,127 | 105 | 1 | 2.63 |
| | Self-RAG* + retry | **5/5** | 41,728 | 318 | 6 | 1.95 |
| | CALIBER | **5/5** | 21,389 | 99 | 1 | 1.00 |
| NFL attempts | Dense RAG | 4/5 | 15,993 | 99 | 1 | 2.67 |
| | LongRAG* | **5/5** | 16,188 | 60 | 1 | 2.63 |
| | DOS RAG* | **5/5** | 16,006 | 81 | 1 | 2.66 |
| | Self-RAG* + retry | 4/5 | 98,029 | 501 | 7 | 2.30 |
| | CALIBER | **5/5** | 42,653 | 82 | 1 | 1.00 |
| Birth dates | Dense RAG | 0/5 | 39,918 | 28 | 1 | 3.30 |
| | LongRAG* | 0/5 | 39,713 | 26 | 1 | 3.32 |
| | DOS RAG* | 1/5 | 40,493 | 59 | 1 | 3.25 |
| | Self-RAG* + retry | 2/5 | 241,574 | 501 | 7 | 1.83 |
| | CALIBER | **5/5** | 131,746 | 281 | 3 | 1.00 |

**Reproducibility and scope.** The experiment records retain the question identifiers, model versions, upstream implementation revisions, adaptation scripts, prompts, retrieval selections, responses, and API-reported token usage. The adaptations and their configurations are described above; upstream repository revisions alone do not specify the evaluated systems. These cases were purposively selected from questions with successful CALIBER responses and illustrate differences in answer completeness, rather than estimate aggregate benchmark performance.

## Recovering complete attributes across documents

**FanOutQA qualitative pilot.** GPT-4o-2024-08-06; temperature 0; same 19-article corpus.
Baseline retrieval budgets: 8K (left), 40K (right). CALIBER: source-bundle routing and 80K boundary.
**All CALIBER–baseline total-input-token ratios are below 3.32×.**

### (a) Missing book attributes

**Question**

What is the page count for each of the books in the five largest book deals?

**Evidence needed**

Deal ranking → five book articles
6 source documents; 5 page-count attributes.

**Dense RAG** — 8,008 input tokens

Page counts not provided in the context.
**0/5 page counts.**

**LongRAG*** — 8,030 input tokens

Hard Choices: 656; Lower Back Tattoo: 336.
Other three counts: not specified in the context.
**2/5 page counts.**

**DOS RAG*** — 8,127 input tokens

My Life: 1,008; The Woman in Me: 288;
Hard Choices: 656.
Born to Run and Lower Back Tattoo: not specified.
**3/5 page counts.**

**Self-RAG* + retry** — 41,728 input tokens

All five page counts correct.
**5/5; matches CALIBER.**
1.95× CALIBER's total input tokens.

**CALIBER (standalone)** — 21,389 input tokens

My Life: **1,008 pages**
The Woman in Me: **288 pages**
Hard Choices: **656 pages**
Born to Run: **528 pages**
The Girl with the Lower Back Tattoo: **336 pages**

**5/5 complete page counts.**

### (b) Incomplete multi-hop answer

**Question**

Which borough is Taylor Swift born in and what are the birth dates of the town's other 5 notable people?

**Evidence needed**

Swift → West Reading → five biographies
7 source documents; borough + 5 full birth dates.

**Dense RAG** — 39,918 input tokens

West Reading; other birth dates not provided.
**0/5 complete birth dates.**

**LongRAG*** — 39,713 input tokens

West Reading; the context does not provide
the other people's birth dates.
**0/5 complete birth dates.**

**DOS RAG*** — 40,493 input tokens

West Reading. Chad Henne: July 2, 1985.
Michael W. Kirst: 1939; Andrew H. Knoll: 1951.
Fetterman and Gursky omitted.
**1/5 full dates; two year-only answers.**

**Self-RAG* + retry** — 241,574 input tokens

West Reading; Fetterman and Henne dates correct.
Gursky, Kirst and Knoll: birth years only.
**2/5 full dates; three year-only answers.**

**CALIBER (standalone)** — 131,746 input tokens

**West Reading, Pennsylvania.**

John Fetterman: **Aug. 15, 1969**
Al Gursky: **Nov. 23, 1940**
Chad Henne: **July 2, 1985**
Michael W. Kirst: **Aug. 1, 1939**
Andrew H. Knoll: **Apr. 23, 1951**

**5/5 full dates; fewer tokens than Self-RAG*.**

*Common-backbone adaptations; see the accompanying protocol. Self-RAG uses GPT-4o critics and one retrieval retry, not its trained reflection-token model. Recorded outputs are condensed; scores count complete attributes, not EQ. Input totals include all online critic/revision/aggregation calls, excluding offline indexing. Selected cases; compute is not matched. CALIBER outputs are reused from the standalone pilot.

Figure 3: Recorded responses condensed for two qualitative FanOutQA cases. Standalone CALIBER recovers complete attributes omitted by budgeted Dense RAG and the LongRAG/DOS RAG adaptations. The Self-RAG-style retrieval-retry variant ties CALIBER on book page counts but returns only two of five complete birth dates. All CALIBER–baseline input-token ratios are below 3.32; this is not a compute-matched evaluation.