أكاديمية كاوست
KAUST ACADEMY

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

ReadEnough: Training-Free Early Stopping for Long-Context Reading

Almoayyad Abuljdail, Muath Alyobi, Riyadh Almutawa,
Mohamed Eltahir, Tanveer Hussain

almoayyad.abuljdail@kaust.edu.sa, muath.alyobi@kaust.edu.sa, riyadh.almutawa@kaust.edu.sa,
mohamed.hamid@kaust.edu.sa, hussaint@edgehill.ac.uk

Abstract

Long-context understanding in LLMs is bottlenecked by how much context a model can process reliably without hallucination. A common approach manages long contexts by replacing the growing context with a compact state that preserves the information needed for answering. However, such methods come at substantial computational cost, implicitly assuming that improving accuracy requires more computation—an assumption that does not always hold. We propose ReadEnough, a training-free early-stopping framework for long-context reasoning based on a measured stopping signal. It sequentially processes the context and directly measures how the model's answer-space belief evolves as evidence accumulates. Reading terminates when the model becomes highly confident and its belief remains stable for a sufficient number of consecutive steps. This creates an adaptive accuracy–compute trade-off without model training or modification, allowing long contexts to be handled without unnecessary computation. On S-NIAH, where the position of the supporting evidence is known precisely, ReadEnough achieves 0.948 accuracy versus 0.991 with full-context reading, while reducing computation by 34%. In 95.7% of cases, it avoids stopping before reaching the supporting evidence, showing that the measured signal tracks when sufficient information has been observed. Compared with asked stopping, Compared with verbalized stopping, ReadEnough reduces premature stops from 9.6% to 4.3% on the same trajectories, indicating that direct measurement can be more reliable than the model's own verbalized judgment of when to stop. ReadEnough also outperforms fixed stopping policies: compared with Fixed@75%, it improves accuracy by 0.026 while using 23% less computation, and achieves a better accuracy–compute balance than both Fixed@25% and Fixed@75%. Finally, ReadEnough points toward a new direction for long-context handling, shifting the focus from how to read more to how to read efficiently.

Date: August 25, 2026
Code: https://github.com/lncch/ReadEnough

Contents

1 Introduction 2

2 Related Work 3

3 Method 3

3.1 Setting . . . 3

3.2 The measured signal . . . 4

3.3 The stopping rule . . . 4

3.4 Oracle-referenced evaluation of the stopping decision . . . 5

3.5 Experimental setup . . . 5

4 Results 6

4.1 Only the signal changes . . . 6

4.2 Does it stop once the evidence has been read? . . . 7

4.3 Five models, two families: consistency is the claim . . . 7

4.4 Beyond the model window . . . 8

4.5 Ablations and attribution . . . 9

5 Discussion 10

6 Limitations and Future Work 10

7 Conclusion 11

References 11

1 Introduction

Large language models are increasingly asked to answer questions about inputs far longer than their context windows. A common approach is to read the input in pieces: the model consumes one chunk at a time while maintaining a compact running state, and produces an answer from that state at the end. This recurrent reading pattern underlies agent-based readers [**zhang2024chainofagents**], RL-trained memory agents [1], and programmatic runtimes that decompose inputs with typed operators [2].

Reading everything is wasteful whenever the evidence needed to answer arrives early. Recent approaches converge on the same fix, early termination, and on a common way of obtaining it: learning a stopping signal. Dynamic context cutoff [3] discovers that specific attention heads encode sufficiency and monitors them with lightweight classifiers trained on data labeled by the position of the gold evidence span. GRU-Mem [4] adds an exit gate to a recurrent memory agent and teaches it through an RL reward, having identified that the plain loop "lacks an exit mechanism". MemAgent [1] trains the memory update itself. Collectively, these works frame the stopping decision primarily as a prediction problem: the model must determine when it has read enough. But the signals used to make this decision differ in where they come from.

A taxonomy of stopping signals. . Placing these alternatives side by side requires distinguishing how each stopping signal is obtained. Existing approaches obtain the signal in two ways: *latent*—inferring sufficiency from internal activations using a trained classifier [3]; or *asked*—explicitly prompting the model to declare whether it has enough information, as in the self-prompting variant of DCC [3] and the exit gate of GRU-Mem [4]. This paper introduces a third mechanism: *measured*. After every chunk, a frozen model already produces a distribution over the answer space. We read this distribution directly—without training or self-report—and halt once it is sufficiently confident and stable, yielding a training-free stopping criterion. The distinction between asking and measuring is substantive: the work that introduced latent probing also reports that self-prompted sufficiency detection reaches an F1 of 52.6 on a 1B model, compared with 88.3 for trained probes [3]. We likewise find that the asked gate stops before the evidence on up to 58% of questions, while the measured signal does so on fewer than 13% of questions across every model we test.

Measure the decision, not just the cost. Prior work evaluates early stopping by its downstream effects: accuracy retained and tokens saved. Both are confounded—a policy can ”save” tokens by stopping before the evidence and guessing. On needle benchmarks, the position of the evidence is known by construction, which provides, for free, the labels used to train latent methods. We use this known evidence position for evaluation: we report the rate of stopping before the evidence, the number of chunks over-read after it, and two oracle-normalized quantities—regret (accuracy lost relative to full reading) and saving capture (the fraction of the savings achievable by an oracle that stops exactly at the evidence).

Contributions..

- **A confidence measure**. We measure confidence from the posterior distribution over answers after each chunk.
- **A stability measure**. We measure stability using the JS divergence between answer distributions across successive chunks.
- **A training-free, tuning-free stopping rule**. We combine confidence and stability of the answer distribution to determine when to stop, with a single fixed operating point (θ=0.995, ε=0.05, w=3).

2 Related Work

Reading long inputs in pieces.. Several approaches address long inputs by processing them incrementally while maintaining information from previously read chunks. Chain-of-Agents [**zhang2024chainofagents**] distributes the input across worker agents that pass intermediate summaries to a manager, while MemWalker [**chen2023memwalker**] and ReadAgent [**lee2024readagent**] navigate long documents through iterative summarization. MemAgent [1] follows a recurrent formulation in which a fixed-length memory is updated after each chunk and trained with RL. These approaches differ in how they process and maintain information from the input, but they do not use a direct signal to decide when further reading is unnecessary.

Existing stop signals. Dynamic context cutoff [3] processes cumulative prefixes and halts when a sufficiency classifier fires. The classifier is a trained ensemble over selected attention heads, trained on data labeled by the position of the gold evidence span, and requires the growing prefix to fit within the model context window, limiting it to inputs substantially shorter than those handled by chunk-based readers. GRU-Mem [4] adds an exit gate to the MemAgent loop, represented by a text token and trained with an RL reward. The gate is therefore an *asked* signal by construction, and their no-RL ablation shows the untrained version is weak. However, existing approaches do not provide a reliable training-free stopping signal.

Stopping on other axes.. Early stopping has also been studied along other computational axes: CALM [**schuster2022calm**] exits transformer *depth*; adaptive self-consistency [**aggarwal2023asc**] stops *sampling* when additional samples are no longer needed; and FLARE [**jiang2023flare**] triggers *retrieval* based on token-level confidence. These methods reduce computation by stopping their respective processes, rather than deciding whether further input reading is necessary. Moreover, they do not evaluate the stopping decision against the known position of ground-truth evidence. Related uncertainty measures—including verbalized confidence [**xiong2024llmconfidence**], self-consistency [**wang2023selfconsistency**], and semantic entropy [**kuhn2023semantic**]—characterize uncertainty in answers but do not use it to steer a reading loop.

	DCC† [3]	GRU-Mem† [4]	MemAgent† [1]	CoA [zhang2024chainofagents]	ES/ASC [schuster2022calm, aggarwal2023asc]	**ReadEnough**
Training-free	×	×	×	✓	✓	✓
No labeled evidence spans	×	✓	✓	✓	✓	✓
Black-box (logprobs only)	×	×	×	✓	×	✓
Stops *reading*	✓	✓	×	×	×	✓
No prefix-fit ceiling	×	✓	✓	✓	n/a	✓
Calibrated confidence at stop	×	×	×	×	×	✓
Timing evaluated vs. Oracle	×	×	×	×	×	✓

Table 1 What each method provides. † = trained.

Table 1 summarizes the gaps addressed by ReadEnough: it combines training-free reading termination with output-probability-only access and explicitly evaluates the stopping decision rather than measuring only downstream accuracy or computational cost.

3 Method

3.1 Setting

A question q is paired with a context P that is too long for a single call should consume. The reader splits P into fixed-sized chunks of c characters (c=24K throughout) and processes the

context sequentially: at step t the model receives the question, the running notes (capped at 6K characters), and chunk t, and outputs updated notes—one call per chunk. The final answer is produced from the notes. We assume only that the serving stack exposes top-k token log probabilities. All stopping policies in this paper are replayed over identical recorded trajectories, so every comparison is paired by construction.

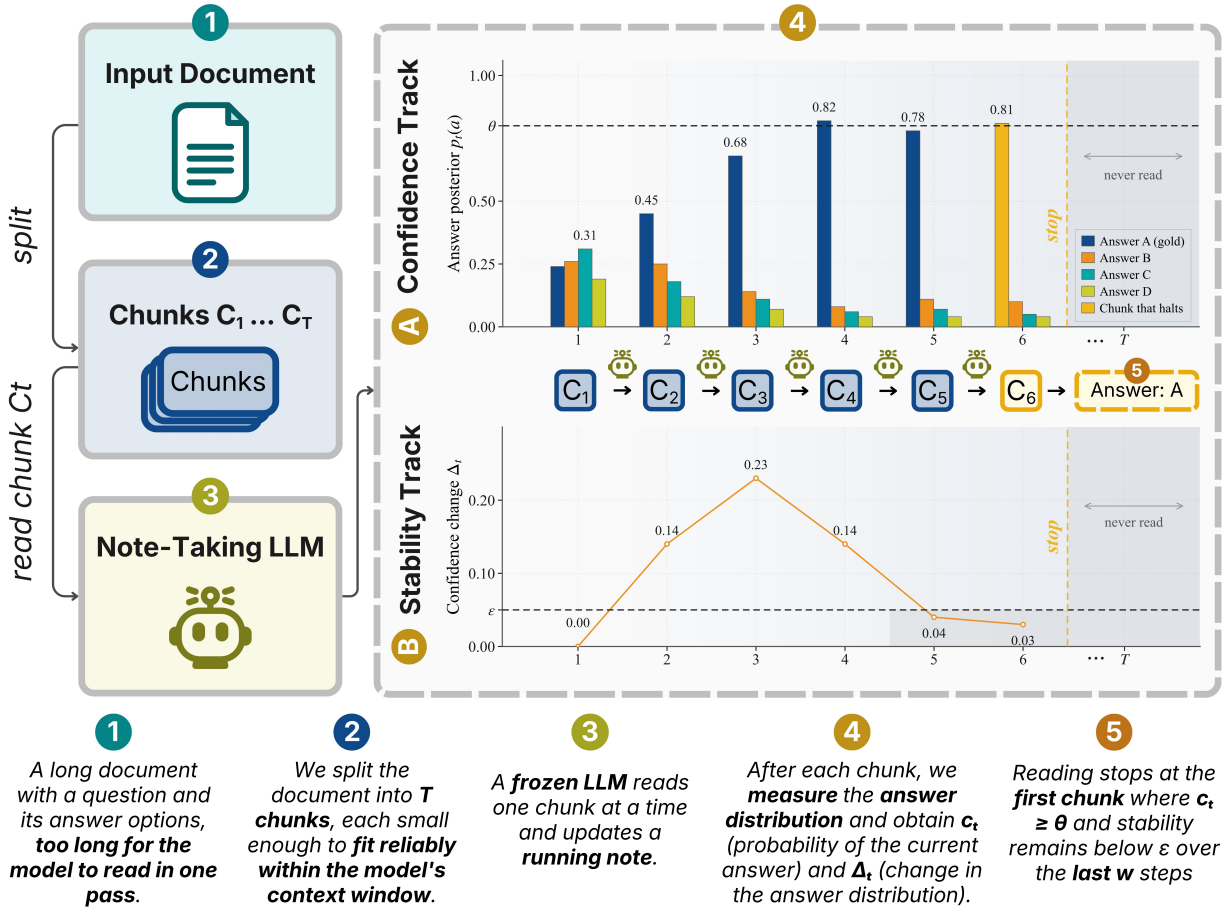


Figure 1 ReadEnough Training-Free Early Stopping for Long-Context Reading.

3.2 The measured signal

After processing each chunk, a short probe call is issued using the question, the current notes, and a fixed minimal template. For multiple-choice tasks, generation is constrained to the option letters and we read the renormalized distribution over options from the first-token log probabilities, obtaining an answer posterior $p_t \in \Delta(\mathcal{A})$. For open-ended tasks the probe instead decodes a short *draft answer* d_t with its sequence-level confidence. The template is identical at every step and across all experiments.

3.3 The stopping rule

Reading process halt at the first step t that passes a *confidence* test and a *stability* test:

$$\text{conf}_t \geq \theta \quad \text{and} \quad \frac{1}{w-1} \sum_{i=t-w+1}^{t-1} D(b_i, b_{i+1}) \leq \varepsilon, \tag{1}$$

For multiple choice, $b_t = \tilde{p}_t$, $\text{conf}_t = \max_a \tilde{p}_t(a)$, and D is the Jensen–Shannon divergence. For open-ended tasks, $b_t = d_t$, conf_t is the draft confidence, and $D(d_i, d_{i+1}) = 1 - \text{F1}(d_i, d_{i+1})$ over tokens—the stability measure used for open-ended drafts, with exact string equality as its $\varepsilon \to 0$ limit; abstaining drafts ("I don't know") never count as converged. The answer at the stop is the argmax of b_t for multiple choice, or the draft itself for open-ended, and its stated confidence is conf_t—reported, not discarded, which is what makes selective prediction possible downstream.

Requiring both tests separates confident-and-settled beliefs from transient spikes: a distractor can spike confidence for one step, but rarely holds the belief still for a window. This sopping rule is fixed as θ=0.995 (confidence threshold), ε=0.05 (JSD threshold), and w=3 (stability window), with the operating point selected using LOMO on the development split and never using test results.

3.4 Oracle-referenced evaluation of the stopping decision

On needle-style tasks the evidence position is known by construction: the needle's location in the source context is available as an evaluation label. Let e be the first chunk containing the evidence and s the chunk at which a policy stops. We report:

- **Premature-stop rate** $P(s < e)$: the fraction of questions where the policy halted before the evidence could have been read—a direct measure of premature stopping.
- **Accuracy conditioned on timing**, acc | $s<e$ and acc | $s\geq e$: decomposes errors into timing versus reasoning.
- **Over-read**: mean chunks consumed after e among late stops.
- **Regret**: oracle-stop accuracy minus the policy's accuracy.
- **Savings capture**: chunks saved divided by chunks oracle stopping saves (ceiling 100%). A policy can exceed 100% only by stopping illegally early, so savings capture must always be read jointly with the premature-stop rate.

Oracle stopping (halt exactly at e) upper-bounds every rule and anchors both normalized quantities. This protocol evaluates the *stopping decision* itself, rather than only its downstream accuracy and computational cost, and is independent of the particular stopping signal used.

3.5 Experimental setup

Benchmarks.. **S-NIAH** (RULER [5]): 250 single-needle instances, 50 per context length in 8K, 16K, 32K, 64K, 128K tokens, with needle depth sampled uniformly (measured mean 0.55) and substring scoring. This is the core benchmark for timing evaluation: we split the instances by ID hash (135/115), with all headline numbers reported on the held-out half.

Models.. Qwen2.5-7B-Instruct, Qwen3-14B, Qwen3-32B-FP8 [6]; Gemma-3-12B-it, Gemma-3-27B-it [7] — five models, two families, served with vLLM [8], thinking disabled, temperature 0.

Baselines.. All run inside the identical fold, so the underlying reading trajectory is held fixed and only the stopping policy differs: *full reading*; *fixed-fraction* stopping at 25/50/75%; *random* stopping (signal-free control, averaged over draws); the *verbalized gate* (asked), using the published exit prompt of GRU-Mem [4] verbatim at its 80% confidence threshold; *confidence-only* (ablating the stability test); and an *accumulator* variant of the measured rule. At the system level: *direct* single-call reading at a fair character cap covering the model window. Trained methods [1, 3, 4] appear as published reference values only.

Token accounting.. Each policy is charged for the tokens consumed by its own fold and probe calls: answer probes for ReadEnough, verbalized-gate calls for the asked baseline, and no probes for policies that do not use them. A policy is never charged for probes that another policy would have issued.

Metrics and statistics.. We report accuracy, token usage, ECE for confidence calibration, and AUROC for the ability of confidence to distinguish correct from incorrect answers; AUROC is uninformative when accuracy approaches ceiling because the negative class becomes too sparse.

4 Results

All numbers use the experimental pipeline described in Section 3.5: constrained probes, separate token accounting, a fixed operating point θ=0.995, ε=0.05, w=3, and headline statistics computed on the held-out half. Sections 4.1–4.4 use the 250-instance S-NIAH benchmark with Qwen3-14B (test n=115); Section 4.3 replicates the analysis across five models.

4.1 Only the signal changes

Table 2 compares stopping policies under the same reading procedure: every policy runs over identical chunks with identical prompts and model, and tokens include each policy's own probe cost. Thus, differences in accuracy and cost arise from the stopping policy rather than from the underlying reading trajectory.

Table 2 Stopping policies under a single reading procedure, S-NIAH, Qwen3-14B, held-out n=115. Tokens include each policy's own probe cost. ECE = expected calibration error (lower is better); AUROC is omitted because full reading is near-ceiling and every policy's AUROC degenerates to 1.0.

Family	Stopping signal	Acc	Tokens	ECE
none	full reading	0.991	54,881	0.009
none	fixed at 25%	0.409	17,857	0.545
none	fixed at 50%	0.635	31,891	0.339
none	fixed at 75%	0.922	46,881	0.074
none	random stop	0.665	33,015	—
asked	verbalized gate, prompt of [4]	0.904	28,632	0.088
measured	confidence test only	0.922	28,049	0.078
measured	accumulator variant	0.730	27,035	0.262
measured	**ReadEnough (confidence + stability)**	**0.948**	36,097	**0.052**

Three comparisons establish the main result. Against full reading, ReadEnough achieves 0.948 accuracy versus 0.991 while using 34% fewer tokens. Against fixed and random stopping, it shows the comparable savings from arbitrary stopping come with substantially larger accuracy losses: fixed stopping at 50% spends comparable tokens but achieves only 0.635 accuracy. Against the verbalized gate—the published prompt GRU-Mem [4] method, used here without its training—ReadEnough achieves +0.044 accuracy on identical trajectories, with a factor-of-two difference in timing errors. ReadEnough also has the best calibration among the stopping policies (ECE 0.052), with stopping confidence directly derived from the measured answer distribution.

4.2 Does it stop once the evidence has been read?

Table 3 evaluates the decision itself against the free oracle (Section 3.4). Evidence depth averages 0.55 of the document; the first evidence chunk ranges 1–21 (median 3).

Table 3 Oracle-referenced stopping on S-NIAH, held-out n=115. Savings capture is meaningful only when the premature rate is low; policies that "save" by stopping before the evidence are marked ‡ and their capture is not comparable to the oracle ceiling.

Stopping signal	P(stop<ev)	Over-read	Acc	Regret	Sav. capture
fixed at 25%	59.1%	0.60	0.409	+0.583	‡
random stop	32.2%	2.15	0.678	+0.313	‡
verbalized gate (asked)	9.6%	0.00	0.904	+0.087	‡
confidence test only	7.0%	0.01	0.922	+0.070	‡
ReadEnough (measured)	**4.3%**	1.35	**0.948**	**+0.043**	**79.3%**
full reading	0.0%	4.00	0.991	+0.000	0.0%
oracle stop	0.0%	0.00	0.991	+0.000	100.0%

Two structural findings. First, **On this benchmark, every observed error is a timing error**: for every policy, accuracy conditioned on stopping after the evidence equals full-reading accuracy (0.991), while accuracy conditioned on stopping before the evidence is 0.000. Thus, early stopping's failure mode is entirely *when* to stop, rather than *whether* the model can answer once the evidence has been read. Second, the stop position tracks the evidence position rather than a fixed fraction of the document: within fixed-length buckets, where document length is held constant, the Spearman correlation between stop chunk and evidence chunk is 0.79–0.93 for the 16K–128K buckets (the 8K bucket has 1–2 chunks and no variance to rank). ReadEnough over-reads 1.35 chunks on average past the evidence reflecting the additional reading required to establish stability, and captures 79.3% of the savings achievable by the oracle at +0.043 regret.

4.3 Five models, two families: consistency is the claim

Table 4 applies the *same fixed constants* to five models spanning two families and a 7B–32B range (n=24 held-out per model; the smaller replication build predates the n=250 headline build and is labeled as such).

Table 4 Fixed operating point across models, S-NIAH, n=24 held-out per model. Premature = P(stop<ev). The asked gate uses the same fixed 80% threshold everywhere. Per-model tuned column: dev-half tuning, superseded by the fixed constants.

	ReadEnough (fixed constants)			Verbalized gate (asked)	
Model	Acc	Premature	Tuned acc	Acc	Premature
Qwen2.5-7B	0.875	12.5%	0.792	0.500	50.0%
Qwen3-14B	0.958	4.2%	0.958	0.917	8.3%
Qwen3-32B	**1.000**	**0.0%**	0.917	0.500	50.0%
Gemma-3-12B	0.917	4.2%	0.917	0.417	58.3%
Gemma-3-27B	0.875	12.5%	0.875	0.750	25.0%
mean	**0.925**	**6.7%**	0.892	0.617	38.3%

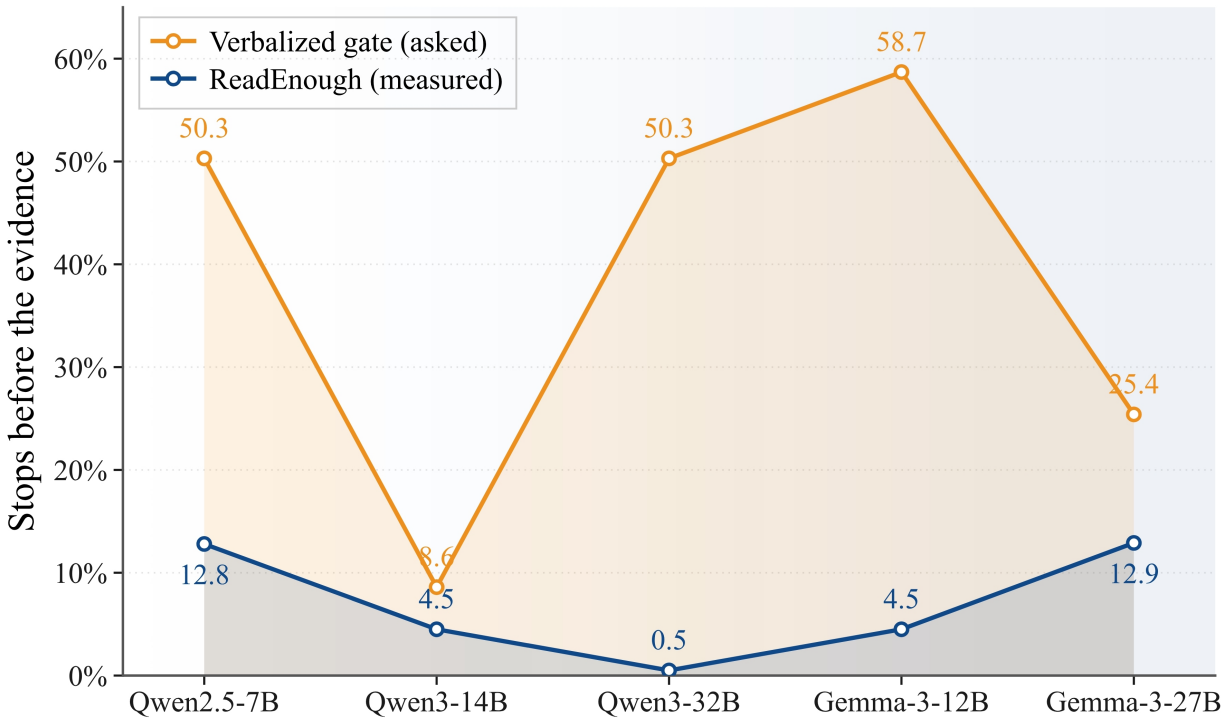


Figure 2 Oracle scatter showing stop position versus evidence position for 115 points. The premature region is shaded.

Three readings. (1) **Measured beats asked on 5 of 5 models**, but the deployment-relevant difference is the range: measured stays within a 0–12.5% premature-stop band, while the asked gate ranges from 8–58%. The asked gate also varies substantially across models, including a 50% premature rate on the strongest Qwen model, whereas the measured signal remains consistently low. (2) **Fixed constants beat per-model tuning** (0.925 vs. 0.892 mean): per-model tuning on the small development halves was unstable, and removing it costs nothing on any model. (3) The 14B row (4.2% premature) closely matches the independent n=115 build (4.3%), supporting the consistency of the result beyond the small replication sample.

4.4 Beyond the model window

Table 5 breaks the headline build down by document length against a *fair* direct baseline: a single call given as much of the document as the model window fits (cap 150K characters, above the window).

Table 5 Accuracy by context length, held-out questions, Qwen3-14B. Direct's token column plateaus at the model window (~40K): beyond it, direct sees a shrinking fraction of the document and its accuracy tracks that fraction.

Context	n	Direct	(tokens)	**ReadEnough**	(tokens)	Full fold
8K	22	1.000	8,096	1.000	8,744	1.000
16K	23	1.000	16,279	1.000	15,538	1.000
32K	20	1.000	32,672	0.950	30,175	0.950
64K	29	0.586	40,077	**0.931**	46,741	1.000
128K	21	0.238	40,073	**0.857**	73,714	1.000

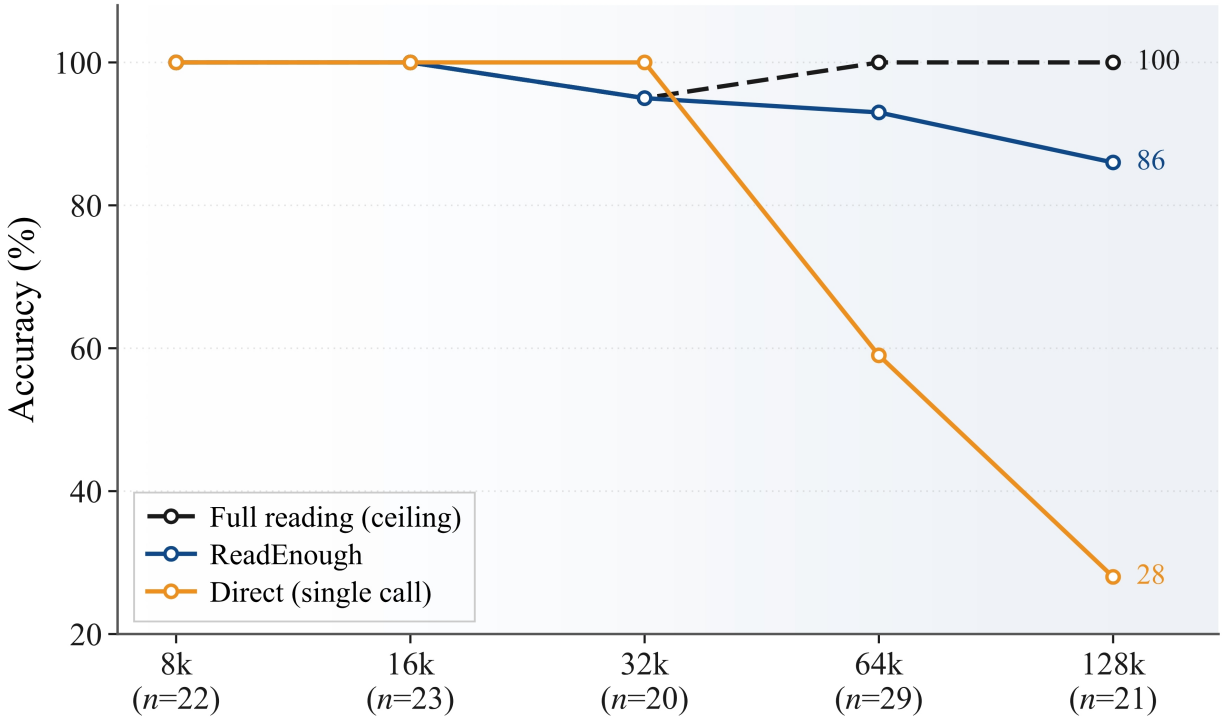


Figure 3 Five-model consistency figure generated: the measured band is shaded, while the asked line is shown as erratic.

Within the window, chunked reading buys nothing—direct ties everything, and honesty requires saying so. Beyond it, direct degrades sharply as the visible fraction of document decreases (0.586 at 64K, 0.238 at 128K), while ReadEnough retains 0.931 and 0.857. ReadEnough's remaining gap at 128k (0.857 vs. full's 1.000 at 128K) is associated with premature stopping; we report this rather than average it away. The regime where a stopping signal matters is exactly the regime where single-call reading is impossible.

4.5 Ablations and attribution

The stability test is load-bearing.. Removing it (confidence-only) costs accuracy and timing on the headline build (0.922 vs. 0.948; 7.0% vs. 4.3% premature) and on the five-model mean (0.817 vs. 0.925, +0.108). On Gemma-3-27B, it is the difference between a reliable and unreliable stopping policy: 45.8% premature / 0.542 accuracy without it, versus 12.5% / 0.875 with it. It changes the stop decision on most questions, with as few as 4/24 questions receiving the same stop as confidence-only, at a cost of 1.35 chunks of over-read.

ε **is empirically inert on extractive tasks..** Across all 1,900 stability-window measurements in the headline build, the statistic is exactly bimodal: 81.5% are exactly 0 (consecutive drafts identical once the needle is read), 18.5% exceed 0.05, and **0.0% fall in** $(0, 0.05]$. Every $\varepsilon \in (0, 0.05]$ therefore yields identical behavior, confirmed by the sensitivity grid, whose ε columns are identical at every θ. On extractive tasks, the stability test thus degenerates to draft equality; ε becomes a genuine threshold only on graded tasks, and we do not present it as a tuned parameter.

θ **is a monotone dial..** Sweeping the single active knob traces the accuracy–cost frontier directly: 0.530 @ 17K tokens (θ=0.5) $\rightarrow$ 0.730 @ 26K (0.95) $\rightarrow$ 0.896 @ 35K (0.99) $\rightarrow$ 0.948 @ 36K (0.995). The final step to the operating point costs only ~1K tokens, placing θ=0.995 near the high-accuracy end of the frontier without test-set tuning.

5 Discussion

Why the trained alternatives are trained. The two closest alternatives learn their stopping signal for different but understandable reasons. Latent probing [3] treats sufficiency as a property to infer from internal activations, requiring labels tied to the position of the gold evidence span; the authors also note that such labels impose a universal cutoff despite models potentially requiring different amounts of context. Our measured answer signal is model-specific by construction: it directly reads the model's answer distribution at the current reading state, rather than learning a separate predictor of sufficiency. Gate-based exit [4], in contrast, asks the model to emit an end marker; making such a self-reported signal reliable requires training, as their no-RL ablation and our five-model comparison suggest. The key difference is therefore not whether a model contains information about sufficiency, but where that information is read: ReadEnough measures the answer distribution the model already produces rather than learning or eliciting a separate stopping signal.

Consistency is the deployment argument. The five-model table's most useful fact is not that measured wins 5 of 5, but that the asked gate's failures are unpredictable. It fails worst on a mid-size Gemma (58% premature) and on the strongest Qwen (50%), while performing substantially better on Qwen3-14B (8%). Model capability alone does not predict when the asked gate will fail. In contrast, the measured signal remains within a narrow 0–12.5% premature-stop band at one fixed operating point, providing a behavior that system builders can plan around.

What measurement buys beyond efficiency. A stopping mechanism typically reduces the decision to whether to continue reading, whereas the answer distribution also provides a confidence estimate for the answer produced at that point. This makes the same signal useful beyond stopping: ReadEnough can evaluate confidence calibration (ECE 0.052, best among the policies in Table 2) and support selective prediction at the stopping point. Thus, the measured signal provides information about both *when* to stop and *how certain* the resulting answer is, rather than reducing its benefit to token savings alone.

Where the value concentrates. The context-length breakdown is candid about where a stopping signal matters: within the model window, a single call performs as well as chunked reading, making the latter unnecessary overhead. The regime that motivates this paper—and the memory-agent literature generally—is beyond the model window, where direct reading degrades with the fraction of the document it can see (0.238 at 128K), while the fold with measured stopping retains 0.857–0.931 accuracy. Early stopping in this regime is therefore not merely cheaper than exhaustive folding; it allows the reading cost to depend more on where the relevant evidence appears than on the full document length.

Timing is the whole failure mode.. On the headline S-NIAH build, accuracy conditioned on stopping after the evidence equals full-reading accuracy. Thus, once the evidence has been read, the fold and its compressed state preserve the information needed to answer; the observed failures come from stopping too early. This sharpens what a stopping signal must accomplish—and what future evaluations should measure: premature stopping is the direct failure mode, while downstream accuracy alone does not reveal it.

6 Limitations and Future Work

Replication scale. The five-model consistency experiment uses only n=24 held-out questions per model, substantially fewer than the n=115 questions in the headline evaluation. The

agreement between the Qwen3-14B replication (4.2% premature) and the larger build (4.3%) is encouraging, but the small per-model sample limits the precision of cross-model comparisons. Larger evaluations across models and datasets would provide stronger evidence for the observed consistency.

Dependence on log-probability access. ReadEnough requires access to top-k token log probabilities to construct the measured answer distribution. This limits applicability to serving systems that expose such information and prevents direct use with APIs that provide only generated text. Extending the measured-signal approach to settings without log-probability access is an important direction for future work.

Savings depend on when confidence emerges. ReadEnough does not guarantee a fixed computational saving. When the model's answer distribution become sufficiently confident only late in the reading process, the stopping rule appropriately continues reading and the resulting savings can be small. The method therefore provides an adaptive accuracy–cost trade-off rather than a guaranteed reduction in computation for every input.

7 Conclusion

Long-context readers do not always need to process an entire document to answer a question. When relevant evidence appears early, continuing to read is unnecessary. Existing approaches address this problem by learning a stopping signal, either from internal model representations or through a trained exit mechanism. ReadEnough takes a different approach: it measures the answer distribution that the model already produces after each chunk and uses its confidence and stability to decide when to stop.

Across five models from two model families, ReadEnough uses the same fixed operating point without training, labeled evidence spans, or per-model tuning. On the held-out S-NIAH evaluation, it stops before the evidence on only 4.3% of questions and capture 79.3% of the savings achievable by stopping exactly at the evidence, with only 0.043 accuracy regret relative to full reading. Beyond the model's context window, ReadEnough retains 0.857–0.931 accuracy, while direct single-call reading drops to 0.238–0.586. Because the stopping signal comes directly from the answer distribution, it also provides a confidence estimate at the point of stopping, making calibration and selective prediction possible.

These results suggest that early stopping does not necessarily require learning a separate predictor of when to stop. On the evaluation needle task, the remaining errors are primarily caused by stopping before the evidence rather than by failure to answer after the evidence has been read. This makes stopping timing the central problem for future improvement. More broadly, ReadEnough shows that measuring a model's existing state can provide a practical, training-free signal for deciding when it has read enough.

References

[1] Hongli Yu et al. "MemAgent: Reshaping Long-Context LLM with Multi-Conv RL-based Memory Agent". In: *arXiv preprint arXiv:2507.02259* (2026).

[2] Alex Zhang, Tim Kraska, and Omar Khattab. "Recursive Language Models". In: *arXiv preprint arXiv:2512.24601* (2025).

[3] Roy Xie et al. "Knowing When to Stop: Dynamic Context Cutoff for Large Language Models". In: *arXiv preprint arXiv:2502.01025* (2025).

[4] Leheng Sheng et al. “When to Memorize and When to Stop: Gated Recurrent Memory for Long-Context Reasoning”. In: *arXiv preprint arXiv:2602.10560* (2026).

[5] Cheng-Ping Hsieh et al. “RULER: What’s the Real Context Size of Your Long-Context Language Models?” In: *COLM*. 2024.

[6] Qwen Team. “Qwen3 Technical Report”. In: *arXiv preprint arXiv:2505.09388* (2025).

[7] Gemma Team. “Gemma 3 Technical Report”. In: *arXiv preprint arXiv:2503.19786* (2025).

[8] Woosuk Kwon et al. “Efficient Memory Management for Large Language Model Serving with PagedAttention”. In: *SOSP*. 2023.