Context Length Alone Hurts LLM Performance Despite Perfect Retrieval

Reading of arXiv:2510.05381v1

Yufeng Du, Minyang Tian, Srikanth Ronanki, Subendhu Rongali,
Sravan Bodapati, Aram Galstyan, Azton Wells, Roy Schwartz,
Eliu A Huerta, Hao Peng

Research group discussion 3 November 2026

arXiv:2510.05381v1, 18-page PDF. Page references in this talk use that PDF.

One sentence: retrieval is not the whole task

Takeaway

A model can recite the relevant evidence from a long input yet solve the same task much less accurately than with a short input.

- Controlled tests vary input length, filler content, evidence position, and attention visibility.
- A two-call retrieve-then-solve workflow improves selected GSM8K and RULER results.

Published in Findings of EMNLP 2025. Public implementation not found as of 3 November 2026.

Paper, Abstract and Secs. 3–5, PDF pp. 1, 4–9; ACL Anthology.

The question: does perfect recall imply effective use?

Common decomposition

Find evidence in the input $\longrightarrow$ use it to answer.

Retrieval-only tests measure the first step.

This paper's test

Hold the underlying problem fixed. Increase surrounding input. Measure recall and solution accuracy separately.

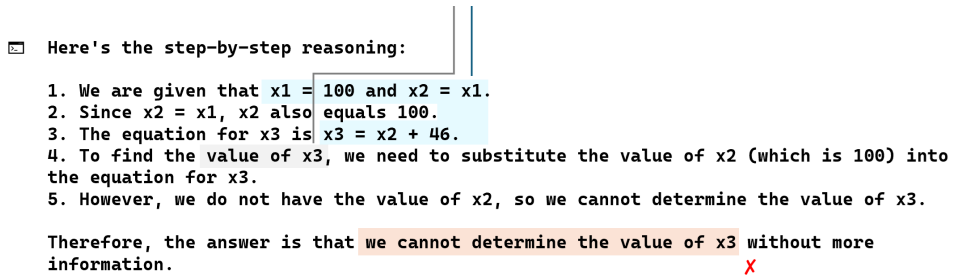
Research question: can length hurt solving even when evidence remains retrievable?

Paper, Secs. 1 and 3.1, PDF pp. 2, 4–5. Retrieval here means reciting context, not external search.

A motivating failure: the conditions are present

Short task

$x_1 = 100, \quad x_2 = x_1, \quad x_3 = x_2 + 46$. Correct answer: $x_3 = 146$.



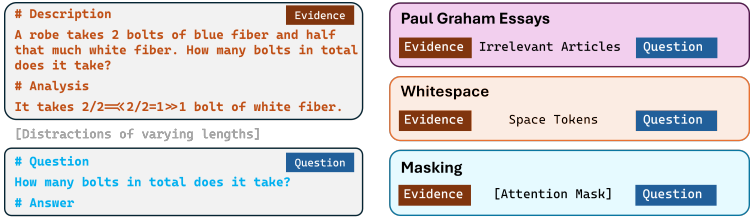
After about 25K spaces, the model states $x_2 = 100$, then says x_2 is unknown.

Crop of original paper Fig. 1, PDF p. 2. One illustrative case; aggregate evidence follows.

The benchmark inserts length into an unchanged problem

$$\underbrace{E}_{\text{evidence}} + \underbrace{D_L}_{\text{filler of length } L} + \underbrace{Q}_{\text{question}}$$

- Evidence is one contiguous span, initially at the beginning.
- Question is at the end; the short task supplies the $L = 0$ baseline.
- Filler progresses from essays to whitespace to masked positions.



Paper, Sec. 3.1, Table 1 and original Fig. 2, PDF pp. 4–5.

Four tasks probe different uses of evidence

Task	Evidence E	Question Q
VarSum	Values of 50 integers	Sum three named variables
GSM8K	Problem and reasoning steps	Math question
MMLU	Problem description	Question and four options
HumanEval	Function and docstring	Coding instruction

Design intent

Make retrieval relatively easy so solution accuracy is visible as a separate bottleneck.

Paper, Sec. 3.1 and Table 1, PDF p. 4; prompt examples, PDF pp. 13–14.

Exact-match recall and accuracy are separate tests

$$R_L = \frac{\#\text{ exact recitations of } (E, Q)}{N} \times 100\%, \qquad A_L = \frac{\#\text{ correct answers}}{N} \times 100\%.$$

$$\Delta A_L = A_L - A_0, \qquad \Delta R_L = R_L - R_0 \quad \text{(percentage points)}.$$

- Any mismatch in the recitation scores zero for that retrieval test.
- The answer prompt does not require recitation before solving.
- These are two task-level rates, not a joint per-answer conditional rate.

Teaching notation for Paper, Sec. 3.1, PDF pp. 4–5; appendix Table 6, PDF p. 16.

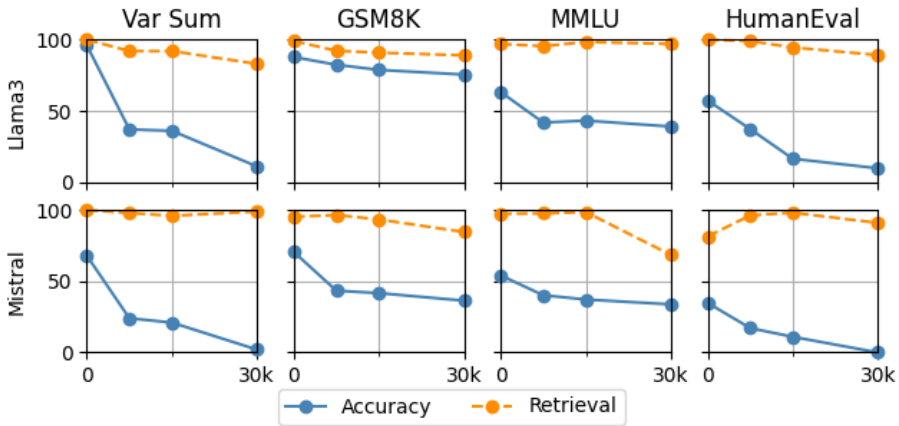
The controls separate candidate explanations

Condition	Input pattern	Main factor tested
Essay filler	$E\ D_L\ Q$	Recall versus answer
Whitespace	$E\ D_L\ Q$	Semantic distraction
Adjacent evidence	$D_L\ E\ Q$	Evidence–question distance
Masked filler	$E\ [\mathrm{mask}]_L\ Q$	Attention to filler content

Masking still shifts positions; adjacency still shifts the absolute position of E, Q.

Paper, Secs. 3–4, original Figs. 2 and 4 and Table 3, PDF pp. 4–7.

Essay filler: accuracy falls faster than recall



Original paper Fig. 3, PDF p. 6. Blue: answer accuracy; orange dashed: exact-match retrieval. Horizontal axis is total input tokens.

At 30K, the recall–answer gap is substantial

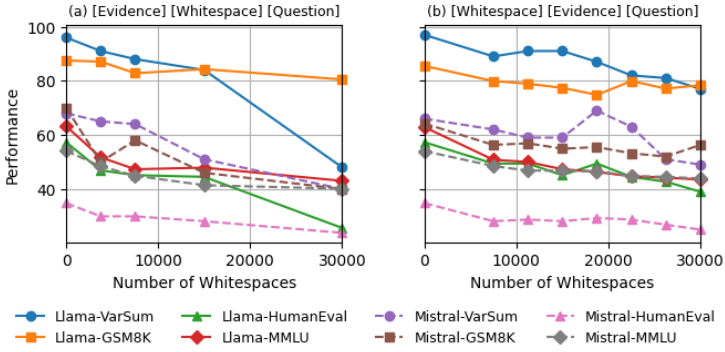
Model / task	$A_0 \to A_{30K}$	$R_0 \to R_{30K}$	ΔA
Llama / VarSum	96.0 → 11.0%	100.0 → 83.0%	−85.0 pp
Llama / MMLU	63.2 → 39.0%	97.0 → 97.0%	−24.2 pp
Mistral / HumanEval	34.8 → 0.0%	81.5 → 91.0%	−34.8 pp

Cleanest contrast

Llama/MMLU keeps 97% exact-match retrieval while losing 24.2 percentage points of accuracy.

Paper, appendix Table 6, PDF p. 16. 30K entries computed from each 0-token baseline plus the reported delta.

Whitespace reduces distraction, not the effect



- Left: E, whitespace, Q. At 30K, Llama/VarSum: 96% $\rightarrow$ 48%.
- Right: whitespace, E, Q. At 30K, Llama/VarSum: 97% $\rightarrow$ 77%.

Original paper Fig. 4, PDF p. 7; appendix Tables 7–8, PDF p. 17. Horizontal axis counts inserted whitespace tokens.

Closed models vary, including clear exceptions

Model / task	0 filler	30K filler	Change
GPT-4o / VarSum	100.0%	100.0%	0.0 pp
GPT-4o / GSM8K	87.8%	80.8%	−7.0 pp
Claude row / MMLU	82.2%	14.6%	−67.6 pp
Gemini-2.0 / GSM8K	83.2%	89.4%	+6.2 pp

No exact-match retrieval results are reported for these three models. The paper prose says Claude-3.7-Sonnet; Table 2 labels that row Claude-3.5.

Paper, Sec. 4.1 and Table 2, PDF p. 6; Limitations, PDF p. 9. Changes are percentage points.

Attention masking removes filler visibility

Model / task	0 masked	30K masked	Change
Llama / HumanEval	57.3%	7.3%	$-50.0\,\mathrm{pp}$
Llama / MMLU	62.8%	41.7%	$-21.1\,\mathrm{pp}$
Mistral / VarSum	66.0%	32.0%	$-34.0\,\mathrm{pp}$
Mistral / HumanEval	34.8%	26.9%	$-7.9\,\mathrm{pp}$

The model attends to evidence and question, while their positions are farther apart.

Paper, Sec. 4.2 and Table 3, PDF p. 7. 30K accuracies computed from baselines and deltas.

The controls support a length-related failure

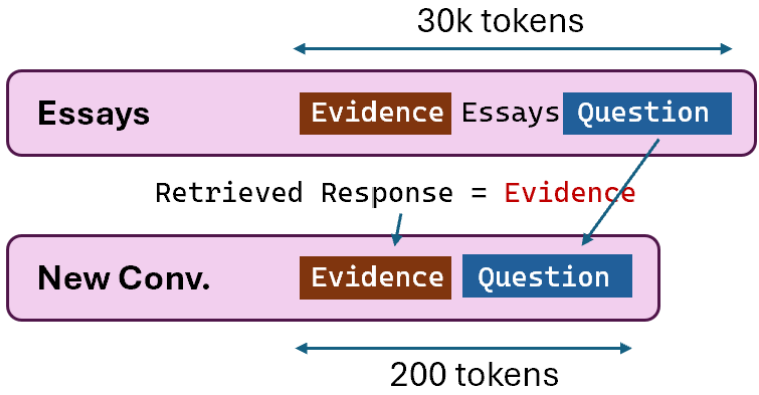
1. Essay filler: recall can remain high while accuracy drops.
2. Whitespace: meaningful distractor content is not required.
3. Adjacent evidence: a growing E-to-Q distance is not required.
4. Masking: attending to filler content is not required.

Interpretation

The study isolates length/position-related degradation. It does not identify a unique internal mechanism.

Paper, Secs. 3–4 and Discussion, PDF pp. 5–9. Interpretation follows the reported controls.

Retrieve then Solve makes the second call short



- First call: recite relevant evidence from the long input.
- New call: answer from that evidence plus the original question.

Original paper Fig. 5 and Sec. 5, PDF p. 8; RULER prompt, PDF pp. 14–15.

On GSM8K, the two-call workflow recovers accuracy

Inserted essay tokens	0	3,750	7,500	15,000	26,250
Direct answer	70.6	49.3	43.4	41.6	35.5
Retrieve then Solve	76.2	71.4	66.7	69.1	66.7

At 26,250 inserted tokens

Mistral-v0.3-7B-Instruct gains 31.2 percentage points: 35.5% $\rightarrow$ 66.7%.

The method gains 5.6 points with no filler, so the entire long-input gain cannot be assigned solely to shortening.

Paper, Sec. 5 and Table 4, PDF p. 8. Scores are answer accuracy in percent.

On RULER, the largest reported gain is 4 points

GPT-4o task / input	Direct	Two-call	Gain
QA2 / 32K	68.4%	72.4%	$+4.0\,\mathrm{pp}$
QA2 / 8K	71.4%	74.0%	$+2.6\,\mathrm{pp}$
QA1 / 4K	90.4%	92.2%	$+1.8\,\mathrm{pp}$
QA1 / 128K	88.2%	88.2%	$0.0\,\mathrm{pp}$

First-step retrieval must be accurate; no open-model RULER result is reported.

Paper, Sec. 5 and Table 5, PDF p. 9; Limitations, PDF p. 9.

Evidence, scope, and a useful next experiment

- **Evidence:** accuracy can fall despite near-stable exact-match recall, sparse filler, or masked filler.
- **Protocol:** recall and answering use separate prompts; closed-model recall is not reported.
- **Method:** the two-call strategy depends on retrieval and adds a model call.
- **Next experiment:** report item-level answer accuracy conditional on verified retrieval under a paired protocol.

Practical evaluation: measure evidence access and final task success at the target input length.

Paper, Secs. 3–7 and Limitations, PDF pp. 4–9. Next experiment is our proposed extension.

Reference and publication record

Yufeng Du, Minyang Tian, Srikanth Ronanki, Subendhu Rongali, Sravan Babu Bodapati, Aram Galstyan, Azton Wells, Roy Schwartz, Eliu A Huerta, and Hao Peng. 2025. *Context Length Alone Hurts LLM Performance Despite Perfect Retrieval.* Findings of the Association for Computational Linguistics: EMNLP 2025, pp. 23281–23298.

Official ACL Anthology entry DOI Read arXiv v1

Publication verified on 3 November 2026. Public implementation not found as of that date.

Paper identity: arXiv v1, PDF p. 1; publication: ACL Anthology official record.

Backup: reported deltas are percentage points

- Appendix Tables 6–9 give absolute accuracy at zero filler and differences thereafter.
- Llama/MMLU with essay filler: $A_0 = 63.2\%$, $\Delta A_{30K} = -24.2\,\mathrm{pp}$.
- Therefore $A_{30K} = 39.0\%$; the reported 24.2 is not a relative 24.2% reduction.
- Essay plot: total input tokens. Whitespace and masking tables: inserted-token counts.

Paper, original Fig. 3, PDF p. 6; appendix Tables 6–9, PDF pp. 16–17.

Backup: three discrepancies in arXiv v1

- Table 2 labels the Claude row Claude-3.5; Sec. 4.1 prose says Claude-3.7-Sonnet.
- Sec. 4.1 attributes Gemini/GSM8K $+8.6\,\mathrm{pp}$ to 30K. Table 2 gives $+8.6\,\mathrm{pp}$ at 15K and $+6.2\,\mathrm{pp}$ at 30K.
- Sec. 4.2 compares Llama/HumanEval masking $-50.0\,\mathrm{pp}$ with whitespace $-19.4\,\mathrm{pp}$; the corresponding 30K between-evidence-and-question whitespace row gives $-31.7\,\mathrm{pp}$.

Paper, Secs. 4.1–4.2 and Table 2, PDF pp. 6–7; appendix Tables 7–8, PDF p. 17. Numeric slides follow the tables.