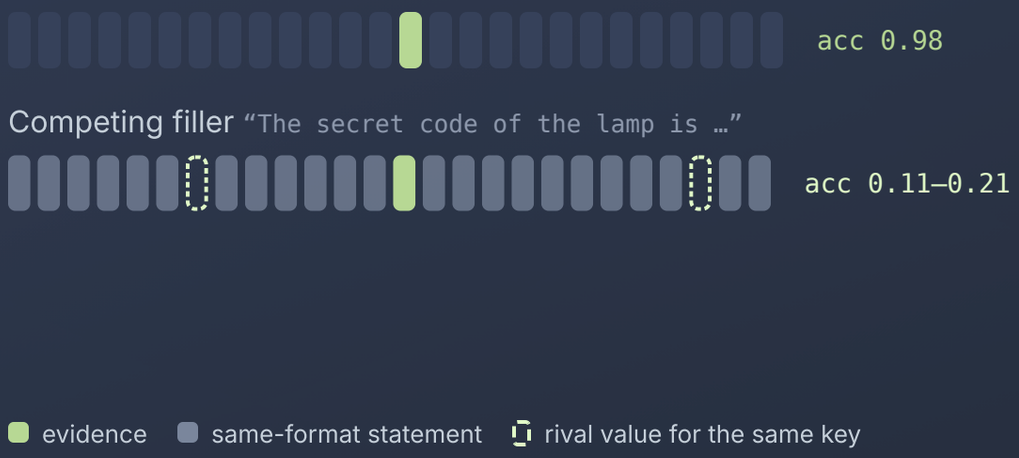
Context rot isn't about length. It's a fight for attention.

SAME EVIDENCE, SAME POSITION, DIFFERENT NEIGHBOURS



ACCURACY AT 32K TOKENS · QWEN2.5-1.5B

0.98

with neutral filler. Same-format statements drop it to **0.11–0.21**.

THE PAPER

Length Is Cheap, Competition Is Not: Attention Predicts Context Rot in Retrieval

Yinuo Liu · UC San Diego · Nov 2026

THE IDEA

Failure tracks how much attention is left on the evidence.



Mohammad Ashfaq Ur
AI Research Notes

ashfaqfardin.dev
arXiv 2611.04169 · preprint

Models fail on long inputs. Is it the length, or what fills it?

The common answer is length. Earlier work showed accuracy dropping even with filler masked out. This paper separates the two.

THE PROBE · KEY-VALUE RETRIEVAL

```
… The secret code of the river is 4821. …

Question: what is the secret code of the
river?
Answer: The secret code of the river is ▮
```

Evidence at a random depth. Score = log-prob of the right value.

The question

Hold the evidence fixed and change only its neighbours. Which change hurts?

THREE WAYS TO FILL THE CONTEXT

NEUTRAL

Random numbers, random words, or one sentence repeated

COMPETING

Statements in the evidence's format with other keys, some giving the river a rival code

MASKED

Filler hidden from attention; position indices still advance (Du et al., 2025)



Mohammad Ashfaq Ur
AI Research Notes

Change only the neighbours, then look inside the model

1

Plant

One key-value fact at a random depth

2

Fill

Neutral, competing or masked filler, 512 to 32K tokens

3

Score

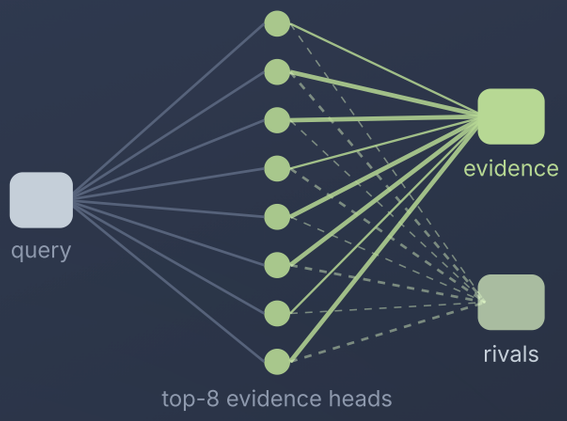
Loss Δ = log(p_short / p_filled), in nats

4

Read

Attention a* from the last token to the evidence

ATTENTION MASS A*



FIVE OPEN MODELS · FOUR FAMILIES

`Qwen2.5-0.5B-Instruct`
`Qwen2.5-1.5B-Instruct`
`SmolLM2-1.7B-Instruct`
`OLMo-2-1B-Instruct`
`Llama-3.2-3B (base)`

Plus Qwen2.5-7B for the MMLU replication.

SCALE

126,473

trial-level measurements, all released

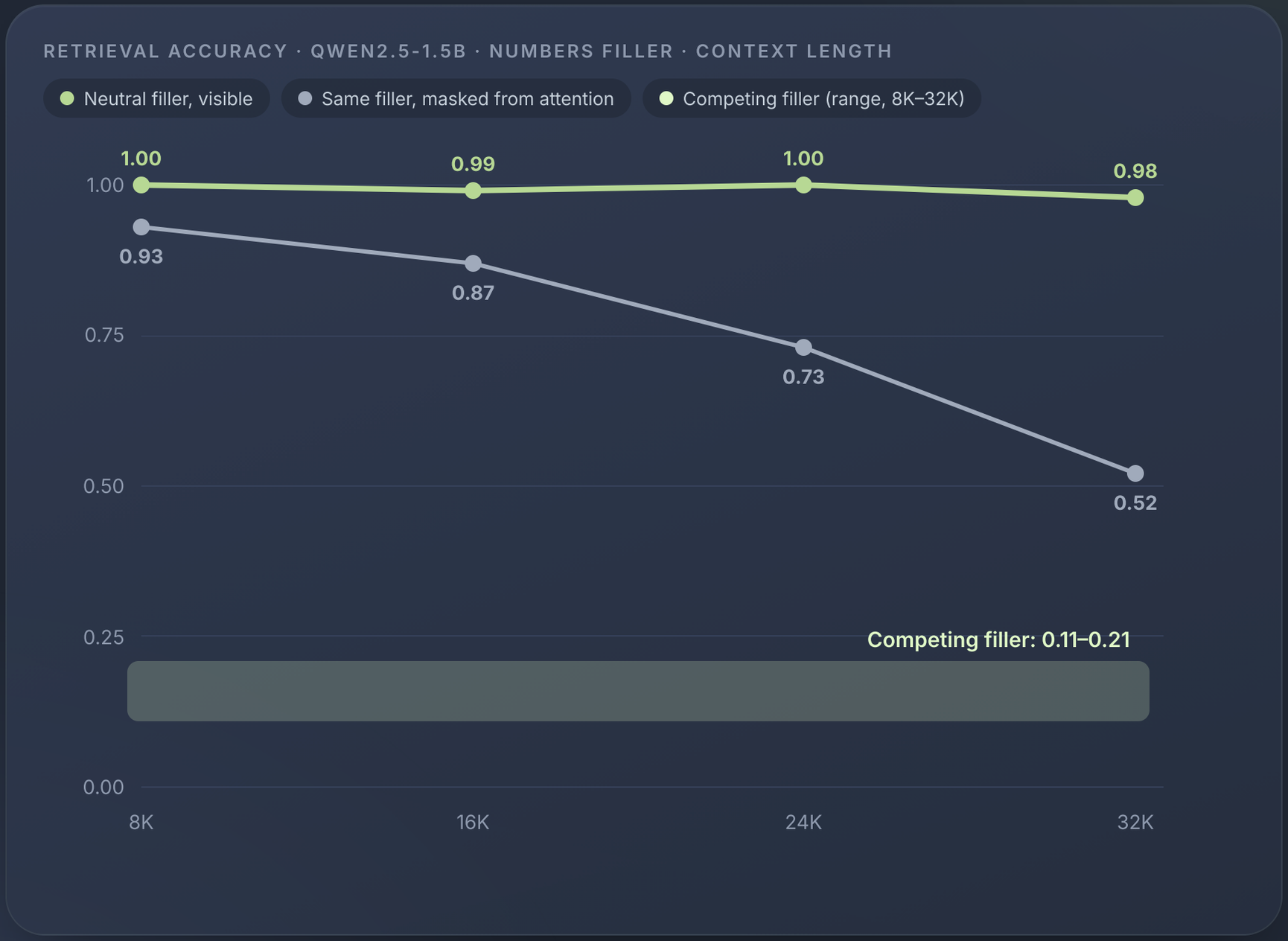


Mohammad Ashfaq Ur
AI Research Notes

ashfaqfardin.dev

arXiv 2611.04169 · preprint

Neutral filler is free. Hiding it is what hurts.



Masked-filler protocol from Du et al. (2025) still reproduces on MMLU (Qwen2.5-7B 65.8 → 48.1 at 30K). Source: Tables 2 & 5, arXiv 2611.04169.

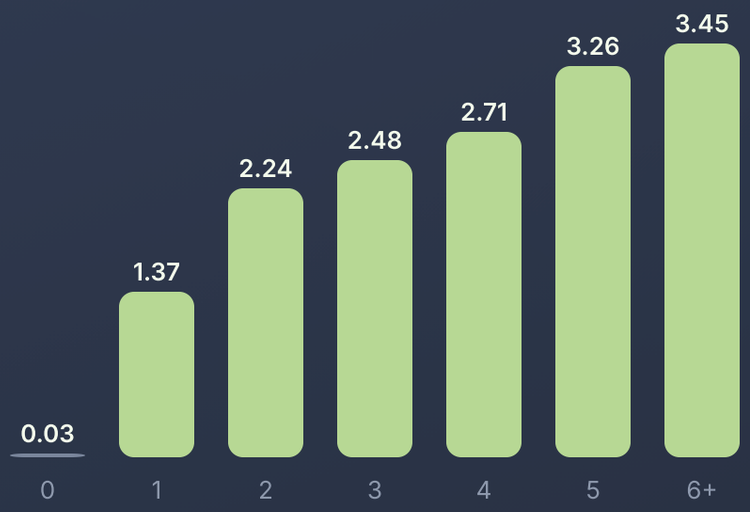


Mohammad Ashfaq Ur
AI Research Notes

ashfaqfardin.dev
arXiv 2611.04169 · preprint

Rival values drive the loss, and attention predicts it

LOSS (NATS) BY NUMBER OF RIVAL VALUES · QWEN2.5-1.5B · 2,048 TOKENS



One rival multiplies mean loss 9× to 110× across the five models. Format-only lookalikes with other keys: $R^2 \leq 0.003$.

VARIANCE IN LOSS EXPLAINED (R^2)



126,473 trials at 512–2,048 tokens (Table 4).

Inside the model: attention on the evidence falls from 0.51–0.77 in short contexts to 0.17–0.31 with competing filler.



Mohammad Ashfaq Ur
AI Research Notes

ashfaqfardin.dev
arXiv 2611.04169 · preprint

For retrieval, worry less about how much context and more about what competes.

CRITICAL PERSPECTIVE

- Correlational: attention read in the same pass, no intervention
- One synthetic task; retrieval models ≤ 3B parameters
- Rival count grows with length, so it isn't isolated above 2K
- Single-author preprint, not yet peer-reviewed

WHAT IT SUGGESTS FOR RAG

→ Deduplicate near-identical passages

→ Resolve conflicting versions of a fact before the prompt

→ Use visible-filler controls in long-context evals

My reading of the results, not experiments in the paper.

REFERENCES

arxiv.org/abs/2611.04169

github.com/yliusci/length-is-cheap

huggingface.co/datasets/yliusci/length-is-cheap



Mohammad Ashfaq Ur

Follow for daily AI research breakdowns

ashfaqfardin.dev

