Language Models Can Control Their Own Attention

Declarative Attention (DA) — paper discussion

Paper: Namgyu Ho, Huzama Ahmad, Woosung Koh,
Se-Young Yun, Tal Schuster, Cicero Nogueira dos Santos

arXiv:2609.02737v1 (2 September 2026)
Reading: 18 September 2026

The model declares what its next steps will read

Main finding

A fixed zero-shot protocol reduces cumulative decode attention reads, with a small average accuracy loss on the two largest models.

- Gemma-4-31B: -52.0% attended tokens; -1.27 percentage points accuracy.
- Qwen-3.6-27B: -31.1% attended tokens; -2.75 percentage points accuracy.
- These are attention-work reductions. Wall-time gains are modeled, not measured.

Table 2, PDF p. 8; Table 3, PDF p. 11.

Sparse attention weights still require expensive reads

- At each decode step, global attention reads the growing KV cache.
- The important tokens are not known before attention is computed.
- Proxy scorers reduce scan cost, but still inspect a context-dependent representation.
- DA asks the model to name its scope in generated text, then reuses that scope across a span.

Sec. 1, PDF pp. 2–3; Appendix A, PDF pp. 23–25.

Three modes separate navigation, extraction and synthesis

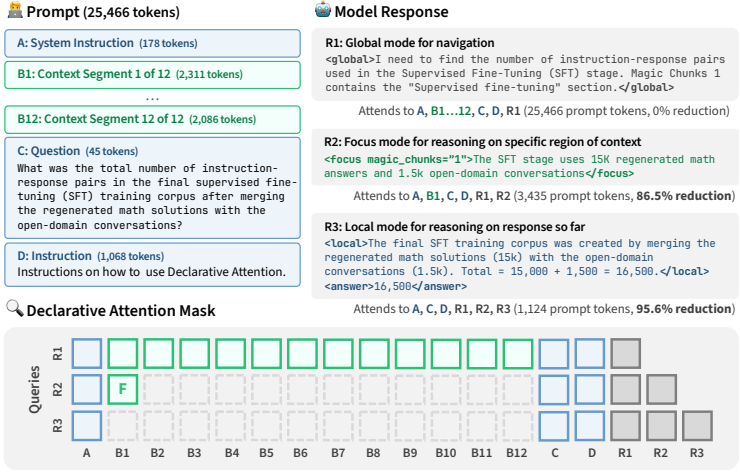
Mode	Visible document	Intended use
`global`	All chunks	Navigate
`focus`	Named chunk(s)	Extract facts
`local`	No chunks	Combine known facts

Always visible

System scaffold, question, DA instruction, and the entire generated response so far.

Sec. 2, PDF pp. 3–4; Appendix B, PDF p. 25.

The declarations become an attention mask



Original Fig. 1, PDF p. 1. Example percentages are not benchmark averages.

Chunks are prepared before decoding begins

- Segment text near semantic boundaries: target 2,048 tokens, usual cap 2,560.
- Prefer paragraphs, lines, sentences, clauses, then words; preserve the exact original text.
- Render numbered chunks as a simulated tool-use transcript.
- No retrieval tool is actually executed. All chunks already reside in context.
- One fixed prompt asks for navigation, verbatim extraction, then local synthesis.

Sec. 2.1–2.2, PDF p. 4; Appendix F, PDF pp. 37–40.

A small state machine controls real KV block reads

- Default: global. A complete focus/local opening tag changes the state; its closing tag restores global.
- Round retained spans outward to KV block boundaries (typically 16–32 tokens).
- Rewrite the request block table through vLLM metadata hooks; retain existing attention kernels.
- Apply only to global attention layers. SWA/GDN layers remain unchanged.
- Full KV cache stays resident: masking is reversible, but does not free cache capacity.

Sec. 2.3, PDF pp. 4–5; Appendix B, PDF p. 25.

Extracted facts make local computation possible

1. Global: locate the founding year in chunk 2.
2. Focus on chunk 2: extract “2003”.
3. Global: locate the IPO year in chunk 7.
4. Focus on chunk 7: extract “2011”.
5. Local: compute $2011 - 2003 = 8$ using the response history.

Interpretation

The model pays for navigation, then avoids repeatedly reading the full document while combining extracted evidence.

Teaching explanation adapted from Sec. 2, PDF p. 3; not an additional experiment.

Total work depends on both scope and output length

$$A = \sum_{t=1}^{D} \text{attended}(t)$$

- A: cumulative KV positions read during one response; D: decode steps.
- It is neither input length nor generated-token count.
- A narrow scope can lose its benefit if the model writes much more.
- Reported means: average within each source, then equally across 15 sources.

Sec. 4, PDF p. 7. Lower A is better; M denotes million positions accumulated over steps.

The experiment separates prompting from masking

Arm	Prompt	Attention
Vanilla	Raw context + reasoning	Full causal
DA-no-mask	Chunked DA protocol	Full causal
DA	Chunked DA protocol	Dynamic mask

- Two main backbones; six models for capability scaling.
- 15 sources: 10 single-span and 5 multi-span tasks; 4 use synthetic QA.
- Up to 128 examples/source; filtered at 244K (E4B: 116K).
- Native thinking disabled; 8K generation limit; B200 + vLLM.

Sec. 4, PDF pp. 6–7; Table 1, PDF p. 6.

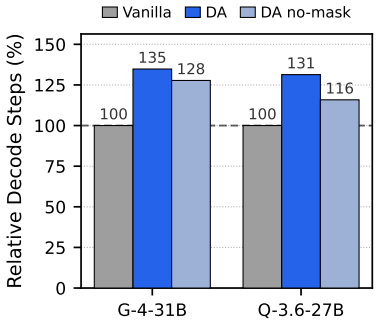
Headline savings come with measurable accuracy loss

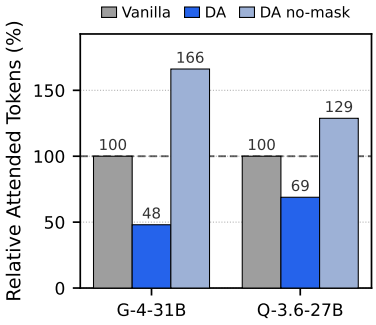
Model	Arm	Accuracy (%) ↑	Attended M ↓
Gemma-4-31B	Vanilla	87.01	13.43
	DA-no-mask	87.01	22.31
	DA	85.74	6.45
Qwen-3.6-27B	Vanilla	85.31	22.54
	DA-no-mask	84.62	29.02
	DA	82.56	15.52

Source-macro averages; accuracy losses are 1.27 and 2.75 percentage points.

Selected rows from Table 2, PDF p. 8; values transcribed, not a new experiment.

Masking offsets the protocol's extra decode steps





DA uses more steps, yet fewer cumulative attention reads.
Without masking, the protocol increases attention work.

Original Fig. 2(b,c), PDF p. 9. Both axes normalized to model-specific Vanilla.

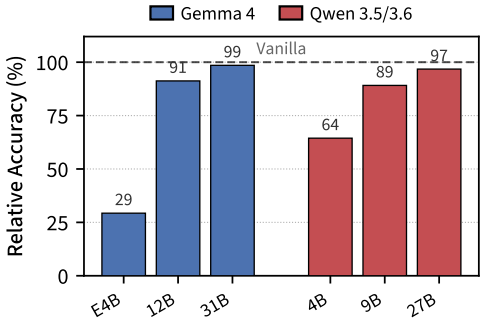
Multi-span tasks lose more accuracy on average

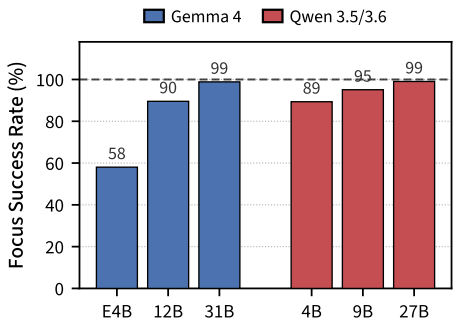
Accuracy loss (percentage points)	Gemma	Qwen
Single-span category	0.78	2.34
Multi-span category	2.28	3.59

- Gemma matches or improves on Vanilla in 7/15 tasks; Qwen in 5/15.
- Qwen attention work increases on five sources.
- Example: Qwen single-document QA rises from 12.39M to 18.00M attended tokens.
- Average savings do not guarantee savings on an individual task.

Sec. 5.1, PDF pp. 7–8; Table 2, PDF p. 8. Category losses use author-reported precision.

Larger models follow the protocol more reliably



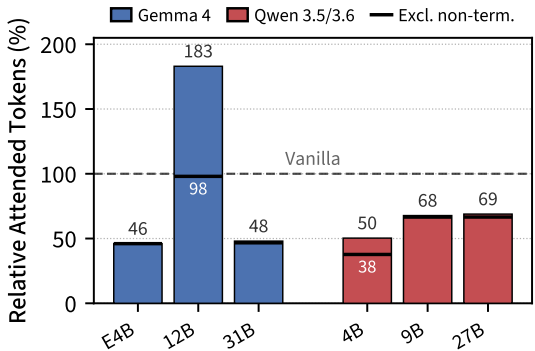


Relative accuracy and valid focus-reference rates improve with model size.

Valid chunk references do not establish correct evidence selection.

Original Fig. 3(a), PDF p. 9; Fig. 6(a), PDF p. 12.

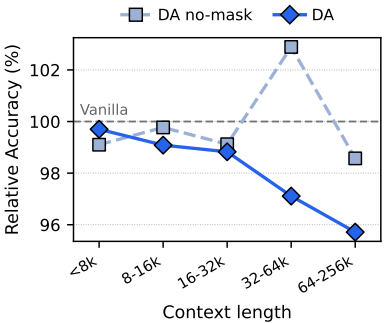
Non-termination can erase the cost benefit

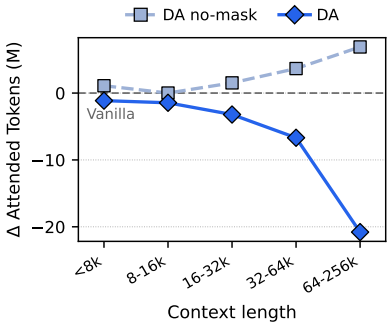


Gemma-4-12B: 183% of Vanilla work; about 6% of DA responses hit the 8K cap.
Excluding non-terminating responses from both arms gives 98%.

Original Fig. 3(b), PDF p. 9. Filtered values diagnose the tail; the tail remains a deployment cost.

Longer contexts increase absolute savings on Gemma

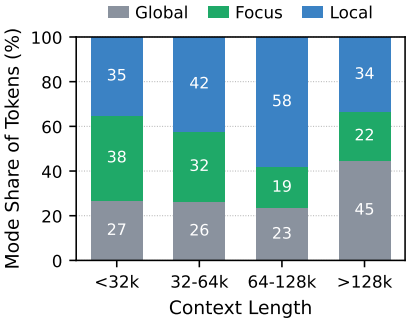


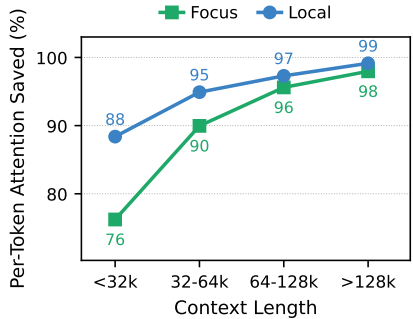


Longest bin: about 96% relative accuracy and 21M fewer attended positions per response. Pooled task bins are not a controlled length-only experiment.

Original Fig. 4, PDF p. 10. Left: relative accuracy; right: DA minus Vanilla (M/response).

Cheap modes dominate tokens, not attention cost





Gemma: focus/local account for about 73% of generated tokens.
Global rises to about 45% in the longest bin, limiting total savings.

Original Fig. 5, PDF p. 12. Left: token shares; right: per-token attention saved.

The roofline model charges three different costs

$$T \approx \underbrace{\frac{2PD}{C_{\text{peak}}\, u_{\text{FLOPs}}}}_{\text{matrix multiplies}} + \underbrace{\frac{Ab_{\text{kv}}}{B_{\text{peak}}\, u_{\text{BW}}}}_{\text{global KV reads}} + \underbrace{\frac{Ds_{\text{local}}}{B_{\text{peak}}\, u_{\text{BW}}}}_{\text{SWA/GDN reads}}$$

- P: active parameters; D: steps; A: total attended positions.
- b_{kv}: global KV bytes/position; s_{local}: fixed bytes/step.
- B200 bf16: 2.25 PFLOPS, 8 TB/s; MFU 40%, MBU 70%.
- Assumes saturated, large-batch serving: matmuls compute-bound, KV reads bandwidth-bound.

Combined expression from Appendix C.8, PDF pp. 28–29; no original equation number.

Estimated time falls less than global attention work

Model	Arm	Matmul	Global	Local	Total
Gemma	Vanilla	22.9	196.5	49.7	269.1
	DA	30.9	94.4	67.1	192.3
Qwen	Vanilla	34.4	263.8	8.0	306.2
	DA	45.1	181.6	10.5	237.3

All values: modeled ms per response. Ratios: 0.71× / 0.77×.

Evidence boundary

Not measured latency. Excludes prefill and block-rounding overhead; fixed local costs and extra generation dilute savings.

Table 3, PDF p. 11; Appendix C.8, PDF pp. 28–29. Totals may differ slightly due to rounding.

Additional tasks expose substantial failures

Gemma task	Accuracy V → DA (%)	Attended M: V → DA
Global word counting	100.0 → 70.3	26.50 → 25.12
Structured data	68.4 → 47.4	133.88 → 165.93
Word-frequency output	98.4 → 99.2	3.22 → 6.16
Book-summary ordering	20.3 → 13.3	6.64 → 7.22

- Six failure sources are outside the headline 15-source average.
- Local evidence decomposition can fail; long output can outweigh per-step savings.
- Lossless text segmentation does not guarantee preservation of usable cross-chunk structure.

Selected rows from Table 10, PDF p. 35; Appendix D.4, PDF pp. 33–35.

Promising control interface; incomplete systems evidence

- Established: zero-shot textual declarations can drive reversible attention masks on capable models.
- Unresolved: real serving latency, long-tail failures, native thinking compatibility, and direct sparse-baseline comparisons.
- Future work: trained mode policies, natural tool boundaries, cheaper global navigation, and KV offloading.
- Reader takeaway: account for control overhead and failures alongside every saved KV read.

Sec. 8, PDF pp. 13–14. Training, agent integration, and offloading are proposals, not demonstrated results.

Reference and discussion

Namgyu Ho, Huzama Ahmad, Woosung Koh, Se-Young Yun, Tal Schuster, Cicero Nogueira dos Santos.

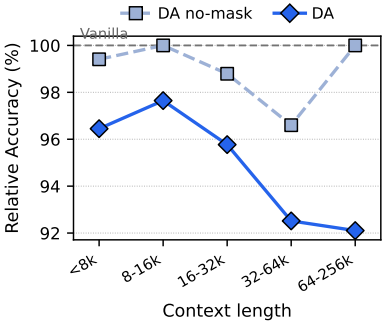
Language Models Can Control Their Own Attention.
arXiv:2609.02737v1, 2 September 2026.
`https://arxiv.org/abs/2609.02737v1`

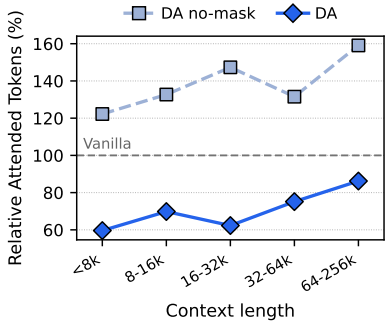
Discussion question

For a target workload, is the bottleneck repeated global KV reads, or the extra reasoning steps needed to avoid them?

Prepared as a paper-reading presentation; the presenter is not a paper author.

Backup: Qwen loses more at the longest contexts





Longest bin: about 92% relative accuracy; attended-token savings shrink. Global token share reaches about 55%.

Original Fig. 8, PDF p. 35; global share from Fig. 9, PDF p. 36.

Backup: judge agreement supports, but does not prove, validity

- Gemini-3-Flash constructs binary rubrics from context and reference answers.
- Thinking-enabled Qwen-3.5-4B grades answers; format failures count as wrong.
- 2,993 responses re-judged by Gemini-3.1-Pro: 98.53% agreement; 95% CI [98.03, 98.90].
- Cohen $\kappa = 0.940$; cell-level Pearson $r = 0.992$.
- Shared-rubric agreement is not human validation. Main performance gaps lack corresponding uncertainty estimates.

Appendix D.2, PDF pp. 32–34; Fig. 7 and Table 9.

Backup: what to measure in a replication

- End-to-end latency and throughput at fixed concurrency, context distribution, and hardware.
- Include prefill, parser/block-table overhead, output length, and block rounding.
- Keep failed and non-terminating requests in workload-level accounting.
- Compare against established sparse attention at a matched accuracy/cost budget.
- Test real multi-turn agent traces and report uncertainty across repeated runs.

Reader-proposed validation agenda; not experiments reported by this paper.