EARLY CONTEXT STARVATION IN CONVERSATIONAL LLMS: CAUSAL EVIDENCE FROM ATTENTION INTERVENTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models degrade over long conversations, but standard evaluation offers no mechanistic account of why. On human multi-turn dialogue rather than synthetic probe tasks, we show that phase-dependent degradation persists across three generations of two proprietary model families despite real capability gains. On two open-weight models, Gemma 4 31B and Qwen3.6-27B, we then measure and intervene on attention directly. In Gemma, attention to a conversation's earliest content collapses as the dialogue grows, because its sliding-window layers lose sight of early tokens and only its global layers still reach them. A dose-response intervention that biases attention toward or away from early context shows this attention is causal for late-turn accuracy. The effect is position-specific: the same intervention on middle-context keys reverses direction, and on late-context keys produces a sharp threshold collapse. Layer-band interventions localize the early-context effect to Gemma's deep layers. On Qwen3.6 the intervention moves attention just as much but leaves behavior unchanged, because its shallow and deep layers push in opposite directions and cancel. Whether attention to early context matters is therefore a property of the architecture, which motivates phase-resolved evaluation and architecture-aware accounts of multi-turn failure.

1 INTRODUCTION

Large language models are deployed as conversational interfaces for extended interactions—tutoring systems, social companions (Park et al., 2023), customer service agents (Chen et al., 2021), and clinical assistants. These deployments assume stable performance across turns, yet standard LLM evaluation relies on short-context, single-turn benchmarks that are structurally incapable of detecting temporal failure modes (Liu et al., 2024; Laban et al., 2026).

Three gaps motivate this work. **No mechanism**: behavioral benchmarks document *that* performance degrades over turns (Zheng et al., 2023; Kwan et al., 2024; Bai et al., 2024a), including on simulated multi-turn versions of single-turn tasks (Laban et al., 2026), but not *why*. **No human dialogue**: concurrent mechanistic work attributes multi-turn failure to declining attention to goal tokens, but on experimenter-constructed probe tasks (scripted fact-recall and persona-compliance episodes) (Dongre et al., 2026). **No cross-architecture causal test**: that work's causal ablation is run within a single model; whether a causal attention–behavior link transfers to another architecture – and if not, why – has not been tested. We address all three on human–robot and human–human customer-service dialogue; Figure 1 summarizes the approach.

Contributions.

1. **A generation-stable behavioral failure mode.** Across three generations of two proprietary model families, phase-dependent degradation persists despite per-generation capability gains.
2. **An architectural account of starvation.** In Gemma 4, early-context attention collapses because its sliding-window layers lose sight of early tokens as dialogues grow; its global layers do not decay, and Qwen3.6, without sliding windows, barely does.
3. **Causal, position-specific, layer-localized evidence.** A dose-response intervention on Gemma 4 shows early-context attention is on the causal path to accuracy; the same intervention on middle

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.

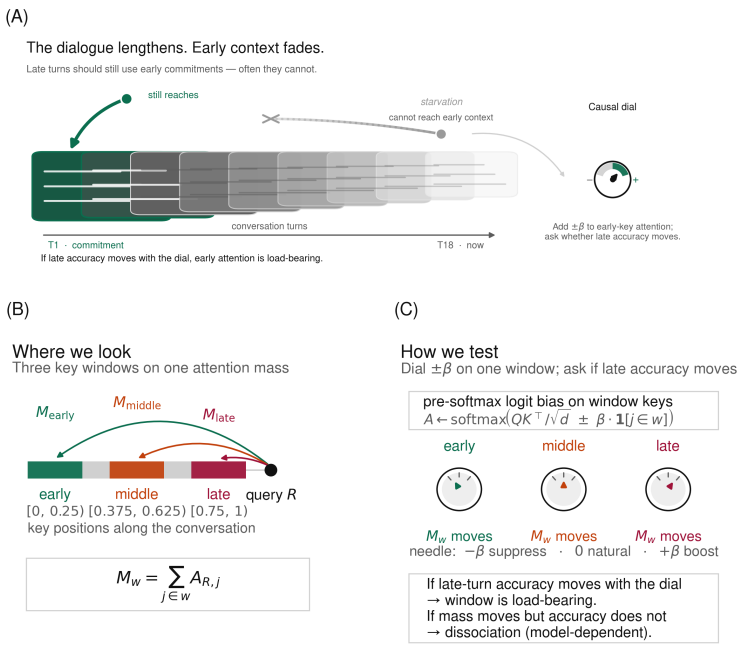

Figure 1: Overview. (A) As a dialogue lengthens, late turns can lose access to early commitments (starvation); a causal $\pm\beta$ dial tests whether restoring early attention moves late-turn accuracy. (B) Three probe windows over key positions – early $[0, 0.25)$, middle $[0.375, 0.625)$, late $[0.75, 1)$ – separated by gaps so that dialing one window does not spill into its neighbors; the measured quantity is window mass $M_w = \sum_{j \in w} A_{R,j}$ from the final query position R. (C) A pre-softmax $\pm\beta$ bias on one window's keys. If late-turn accuracy moves with the dial, the window is load-bearing; if mass moves but accuracy does not, mass and behavior dissociate.

and late windows reveals three distinct regimes (graded, reversing, threshold), and layer bands localize the early effect to deep layers.

4. **A cross-architecture dissociation.** On Qwen3.6 the intervention moves attention mass comparably but not behavior; layer bands explain this as *cancellation* between opposing shallow and deep effects, not indifference.

2 Related Work

Long-context and multi-turn failures. Liu et al. (2024) show that retrieval degrades for information in the middle of long inputs, and long-context benchmarks probe how much of a nominal context window models actually use (Bai et al., 2024b; Hsieh et al., 2024). Multi-turn benchmarks measure conversational ability across turns (Zheng et al., 2023; Kwan et al., 2024; Bai et al., 2024a), and Laban et al. (2026) find that models "get lost" when a task is revealed gradually over simulated turns. These works are behavioral; we study fully specified human dialogue and give a mechanistic, causal account.

Concurrent mechanistic work. Dongre et al. (2026) independently propose that attention to goal-defining tokens declines over turns (the Goal Accessibility Ratio), measured on four primary open-weight models with the trend replicated on six more, together with a sliding-window ablation that closes attention to the goal tokens and residual-stream probes. Their tasks are constructed probes (fact retention, persona and policy compliance), and their causal ablation is run in one model. We differ in three ways: human dialogue rather than constructed probes; a graded dose-response bias rather than binary closure, which reveals graded, reversing, and threshold regimes; and the same causal test on two architectures, with early, middle, and late windows and layer bands. We do not probe the residual stream.

Attention mechanisms and interpretability. Attention heads show systematic positional preferences (Clark et al., 2019), including attention sinks (Xiao et al., 2024) and recency biases built into position encodings (Press et al., 2022); residual-aware rollout (Herasimchyk et al., 2026; Abnar & Zuidema, 2020), graph-theoretic positional bias (Wu et al., 2025), and initial-saliency decompositions (Qiang et al., 2025) document structural recency tendencies in transformers (Vaswani et al., 2017). Early context starvation is conceptually distinct from attention sinks: sinks concentrate mass on specific initial tokens (typically position 0 or the beginning-of-sequence (BOS) token) regardless of semantic content, whereas starvation refers to the progressive loss of attention to the semantically meaningful first 25% of conversation tokens as the dialogue lengthens. The window is not dominated by special tokens: BOS makes up $\approx 1.2\%$ of it and all special tokens $\approx 2.3\%$ on ABCD prompts (Appendix I). Our attention intervention relates to causal mediation (Vig et al., 2020), causal tracing and activation patching (Meng et al., 2022; Zhang & Nanda, 2024), and circuit-level analyses (Elhage et al., 2021; Wang et al., 2023); our use of attention mass as an intervention target follows prior cautions about raw attention weights (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019; Abnar & Zuidema, 2020).

Efficient attention. Sliding-window attention (Child et al., 2019; Beltagy et al., 2020; Jiang et al., 2023) and linear or recurrent token mixers (Katharopoulos et al., 2020; Gu & Dao, 2023; Yang et al., 2025) trade full attention for efficiency; recent open models interleave them with a few full-attention layers (Gemma Team, 2025; Qwen Team, 2026). We show this design choice determines how early context is lost.

Reasoning limits and longitudinal evaluation. Chain-of-thought (Wei et al., 2022) and self-correction (Madaan et al., 2023) improve task performance, but Huang et al. (2024) found fundamental limits to intrinsic self-correction; we likewise find that longer reasoning does not offset phase-dependent degradation (Section 4.1).

3 METHODS

3.1 DATASETS

HRI-CUES and ABCD play complementary, deliberately different roles: HRI-CUES identifies the mechanism on richly naturalistic dialogue where labels are subjective; ABCD provides the objective labels a causal test with paired McNemar comparisons requires. MultiWOZ and ESConv are negative controls: short, locally-grounded conversations that should not require early-context integration, used to confirm the phenomenon is specific to tasks that need it rather than a generic property of long conversations.

HRI-CUES (Irfan et al., 2026) contains 590 turns from 25 older adults (mean age 73.4, SD 8.2) in open-domain dialogues with a social companion robot (Swedish; 12–29 turns, mean 23.6). Three expert annotators rated turn-level interaction quality on a 5-point Likert scale (ICC$= 0.44$). We report MAE.

ABCD (Chen et al., 2021) contains human–human customer-service dialogues collected by pairing trained crowdworker agents with crowdworker customers following scenario prompts; each turn's label is the action the agent actually logged, which we score as binary correct. We evaluate 753 instances from 278 conversations exceeding 20 turns (coupling analysis) and all 754 instances (causal intervention), split 20-60-20 into Early (n=63), Middle (n=325), and Late (n=366) phase.

Boundary condition datasets: MultiWOZ (Budzianowski et al., 2018) (1,103 turns; mean turn 10.4; 13% reach turn 18) and ESConv (Liu et al., 2021) (682 turns from 50 conversations; mean 13.6 turns) validate null behavioral patterns where long-horizon context integration is not required.

3.2 MODELS

We evaluate 16 conditions across four families, with three generations each for Claude and OpenAI (Gen 1: Q1–2 2025; Gen 2: Q3–4 2025; Gen 3: Q1–2 2026; full enumeration in Appendix M, Table 11), plus two open-weight architectures, Gemma 4 31B (Google DeepMind, 2026) and Qwen3.6-

27B (Qwen Team, 2026), used for all mechanistic and causal work since proprietary APIs do not expose attention weights.

Both open-weight models are *hybrid* architectures, which matters for interpreting every attention-level result. Gemma 4 31B has 60 layers: 50 sliding-window (local) attention layers with a 1,024-token window, interleaved 5:1 with 10 global attention layers, a design introduced with Gemma 3 (Gemma Team, 2025). Qwen3.6-27B has 64 layers: 48 Gated DeltaNet (Yang et al., 2025) linear-attention (recurrent) layers interleaved 3:1 with 16 softmax attention layers. Only softmax attention layers produce an attention matrix, so attention mass and the intervention below are defined on those layers. In Qwen, three-quarters of token mixing runs through a recurrent state that the intervention cannot reach. All mechanistic and causal experiments use standard mode: Qwen3.6's thinking mode, on by default, is disabled, and for Gemma no reasoning precedes the scored answer. Reasoning variants appear only in Section 4.1 and Appendix C.

Early-context attention mass is the proportion of last-position attention directed at a given window of tokens, pooled across all layers and heads by arithmetic mean unless otherwise noted (layer-band analysis, Section 5.5, restricts this pooling to a contiguous half of layers).

3.3 The Attention-Window Intervention

The intervention modifies the attention computation of each model's eager (non-fused) attention implementation, adding a scalar logit bias $\beta \in \{-2, -1, 0, +1, +2\}$ to query–key scores (pre-softmax) for key positions in a chosen window w, applied uniformly across all softmax attention layers and heads unless a layer-band restriction is specified (Figure 1):

$$A_{ij} \leftarrow \mathrm{softmax}_j\Big(q_i k_j^\top / \sqrt{d} \; + \; \beta \cdot \mathbb{1}[j \in w]\Big),$$

so negative β suppresses and positive β boosts attention to window w. We test three windows over the token sequence: **early** $[0, 0.25)$, **middle** $[0.375, 0.625)$, and **late** $[0.75, 1.0)$. Prompts, context, and turn position are identical across bias levels; only the attention distribution changes. For layer-band analysis (Section 5.5), the bias is restricted to either the shallow or deep half of layers. On Gemma's local layers the bias has no effect on keys already outside the sliding window, which are masked; on Qwen it acts only on the 16 softmax layers. Halves are split by layer index (Gemma layers 0–29 / 30–59; Qwen 0–31 / 32–63, i.e., 8 softmax layers per half on Qwen). We verified that the modified attention path enforces Gemma's sliding-window mask: on all 391 HRI-CUES prompts longer than the window, local layers place zero mass on keys outside it (Appendix B). In all conditions we verify empirically that the intervention moves measured attention mass in the intended direction before interpreting any behavioral result.

3.4 Statistical Analysis

We split each conversation 20-60-20 into Early, Middle, and Late phases; Appendix D shows the ABCD results hold under four other splits. For behavioral trends we fit linear mixed-effects models with a random intercept per conversation (REML), because turns from the same conversation are not independent. Mechanistic trends use Spearman correlation, paired t-tests, and a conversation-cluster bootstrap (B=2000). Every causal contrast compares the same instances with and without the intervention, so we use McNemar's exact test (McNemar, 1947) on the discordant pairs. Because the window and layer-band analyses produce many such contrasts, we apply Holm–Bonferroni correction (Holm, 1979) within pre-defined families and across all 24 all-layer contrasts (Appendix J); † marks results that are significant uncorrected but not after within-family correction.

4 Behavioral Results

4.1 Consistent Temporal Degradation

On HRI-CUES, error follows an inverted-U across phases, peaking mid-conversation (Figure 2); the pattern is the same in reasoning and standard variants (Mann-Whitney p=0.86, 0.68, 0.34 at Early/Middle/Late; all ns). **This mid-conversation peak is specific to HRI-CUES.** On ABCD, degradation is instead monotonic – error rises from 20.6% (Early) to 71.3% (Late) with no late-phase

recovery (Section 5.2) – so we scope the inverted-U claim to HRI-CUES throughout and refer to *degradation* more generally elsewhere.

Mixed-effects modelling across 7,080 observations: middle-phase error increase β=0.081, z=3.68, p<0.001; Phase×Family interactions non-significant (both p>0.74). Reasoning does not help: reasoning-trace length does not vary by phase ($F(2, 586)$=1.22, p=0.296) or predict accuracy (r=0.041; 590 traces).

4.2 Three-Generation Longitudinal Stability

Temporal patterns persist across three development generations (Figure 18 in Appendix L; Claude: ρ=0.96; OpenAI: ρ=0.94) despite 2–10% aggregate capability gains: Claude slump depth evolves 15.2%→14.3%→8.2%; OpenAI dips at Gen 2 (6.8%) before reverting (15.6%), consistent with an architectural rather than training-era property.

4.3 Confound Check: Is This Just Turn Difficulty?

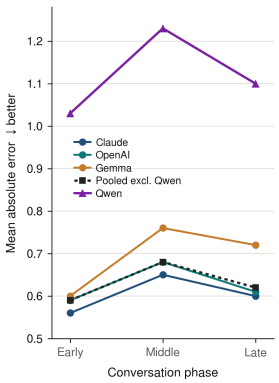

Figure 2: HRI-CUES MAE by phase and family.

A natural alternative explanation is that later turns are intrinsically harder, independent of any attentional mechanism. Two checks argue against this. On ABCD, the majority-class baseline accuracy itself falls sharply with phase (92% / 38% / 15%, Early/Middle/Late), confirming late turns are harder in general. Gemma (79% / 45% / 29%) falls below this baseline early, where one dominant action makes the majority guess strong, but exceeds it in the middle (+7pp) and late (+14pp) phases: its margin over a difficulty-only predictor *widens* as the conversation lengthens, so difficulty alone does not explain the pattern. On HRI-CUES, phase-level inter-annotator agreement (ICC(2,1), n=130/330/130) is 0.32 / 0.46 / 0.38 – *lowest* in early turns, the opposite of what a label-noise account of the MAE pattern would predict.

5 Mechanistic and Causal Analysis

5.1 Early Context Attention Decay (HRI-CUES)

On HRI-CUES, Gemma's early-context mass pooled over all layers halves from 0.183 at turn 4 to a plateau of about 0.093 reached around turn 10, while Qwen's barely moves (0.684 → 0.648, a 5% decline, though monotone in every participant: median per-participant ρ=−0.976, 25/25 negative, Wilcoxon p=5.96×10^{-8}; Figure 3).

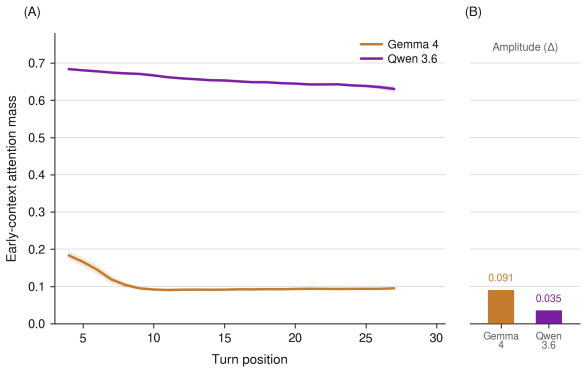

Figure 3: Early-context attention mass vs. turn (HRI-CUES, no intervention). (A) Mean ± 95% CI across 25 participants. (B) Amplitude Δ: mass at turn 4 minus the late-turn plateau, on the same scale: Gemma 0.091, Qwen 0.035. Standard (non-reasoning) mode; Gemma measured with its sliding-window mask enforced.

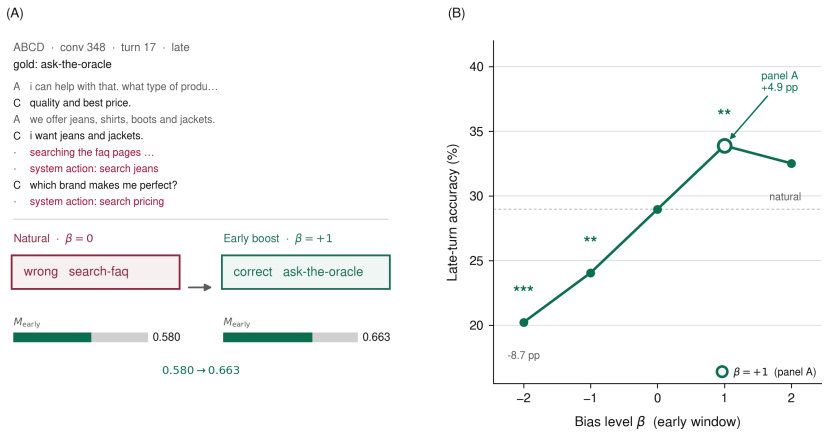

Figure 4: Worked example and early-window dose-response (Gemma 4 / ABCD). (A) Conversation 348, turn 17: at β=0 the model predicts search-faq (wrong; M_{early}=0.580); the early boost β=+1 yields ask-the-oracle (correct; M_{early}=0.663). (B) Late-turn accuracy vs. β (n=366); stars mark McNemar significance vs. β=0 (*** $p<0.001$, ** $p<0.01$). The open point is panel A's condition, the largest beneficial effect in the study (+4.9 pp).

Per-layer measurement shows exactly where Gemma's decline comes from. HRI-CUES prompts grow from ∼1,140 tokens at turn 4 to over ∼1,800 in later turns, passing ∼1,365 around turns 9–10. Beyond ∼1,365 tokens the early quarter lies entirely outside the 1,024-token window of Gemma's 50 sliding-window layers, which then place exactly zero mass on it (all 391 long prompts). Gemma's 10 global layers, which see the whole context, hold early mass flat (0.552 early vs. 0.557 late). Starvation in Gemma is therefore *architectural*: once a dialogue outgrows the local window, five-sixths of the network loses access to early context and only the global layers still reach it. Qwen, whose softmax layers are all full-attention, shows no comparable loss. On ABCD, prompts are short (mean 374 tokens) and nearly always fit inside the window, so the causal tests below probe the attention Gemma still pays to early context rather than the window cutoff.

5.2 Establishing Causality: The Dose-Response Intervention

Observational coupling between early-context mass and ABCD accuracy (ρ=0.267, p=9.3×10^{-14}) is confounded with turn index: a logistic regression controlling for turn position reduces this to ρ_{partial}=−0.020 (ns), so we rely on the intervention below for causal evidence.

On Gemma 4 31B / ABCD (n=754), the all-layer early-window intervention produces a near-monotonic dose-response (Figure 4; Table 1 in Appendix A): suppressing early attention (β=−2.0) reduces late-turn accuracy by 8.7 pp (p=9.1×10^{-6}); moderately boosting it (β=+1.0) improves accuracy by 4.9 pp (p=0.005); boosting further (β=+2.0) shows a smaller, non-significant gain (+3.6 pp, p=0.148), consistent with overshoot. Effects replicate with larger magnitude in the middle phase (suppress −17.8 pp, p=1.2×10^{-9}; boost +8.9 pp, p=5.7×10^{-5}), ruling out a late-turn-only ceiling artifact. All causal claims in this section are scoped to Gemma 4 31B on ABCD.

5.3 Position Specificity: Early, Middle, and Late Windows

The intervention above establishes causality but not *specificity* to early position. We apply the identical intervention to middle $[0.375, 0.625)$ and late $[0.75, 1.0)$ windows on the same model and dataset (Figure 5; both phases in Appendix E). The three windows behave qualitatively differently, not merely at different magnitudes:

Early (Section 5.2): graded, monotonic, suppress hurts and boost helps.

Middle: *reverses* the early-window pattern and is phase-dependent – suppressing middle-window attention *helps* middle-phase accuracy (+10.5 pp at β=−1.0, p=1.4×10^{-7}), while boosting it hurts

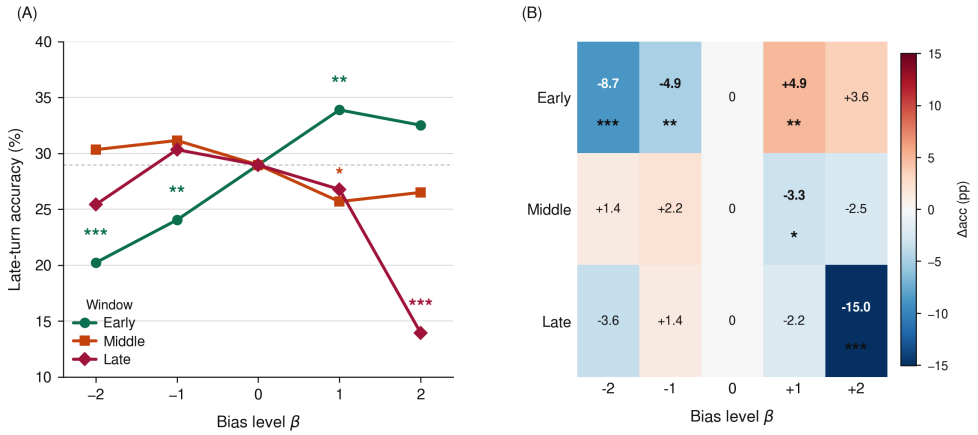

Figure 5: Position specificity (Gemma 4 / ABCD, late turns, n=366). (A) Late-turn accuracy vs. β for each key window. (B) Δaccuracy (pp vs. β=0); stars mark uncorrected McNemar significance (*** p<0.001, ** p<0.01, * p<0.05). All starred cells survive Holm correction except middle $+1$ (Appendix J).

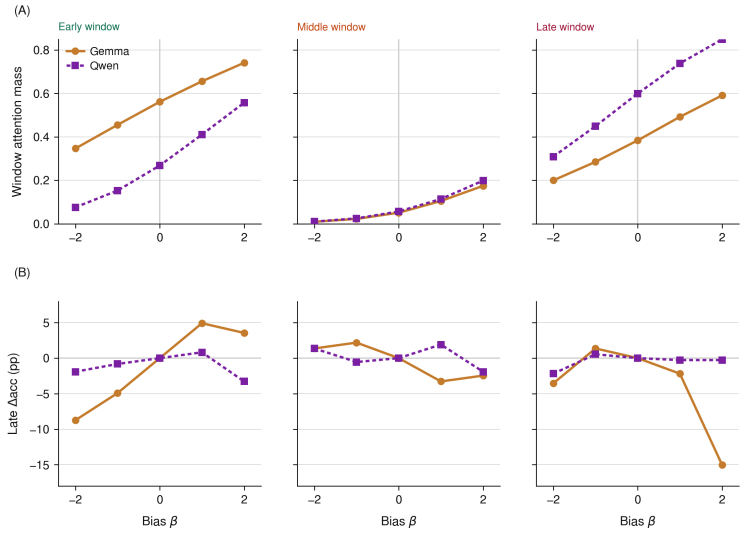

Figure 6: Cross-architecture transfer (ABCD, identical protocol). (A) Window attention mass vs. β: the dial moves mass on both Gemma and Qwen. (B) Late-turn Δaccuracy vs. β: Gemma's curves move; Qwen's stay flat in every window.

sharply (-17.2 to -19.1 pp, $p<10^{-11}$); late-phase effects point the same direction but are weaker and mostly non-significant at this sample size.

Late: a *threshold* rather than graded effect – little happens at $\beta = \pm 1.0$ (all $p>0.13$), but boosting to β=+2.0 produces the single largest effect across all nine window×bias conditions tested (-15.0 pp late, $p=2.4\times10^{-11}$; -31.7 pp middle, $p\approx0$).

In every window, we confirm the intervention moved measured attention mass as intended (middle-window mass $0.009\rightarrow0.051\rightarrow0.175$ suppress/natural/boost; late-window mass $0.200\rightarrow0.384\rightarrow0.591$) before interpreting the behavioral result. Early position is thus not interchangeable with the other windows, but not uniquely fragile either: each window has a distinct causal shape.

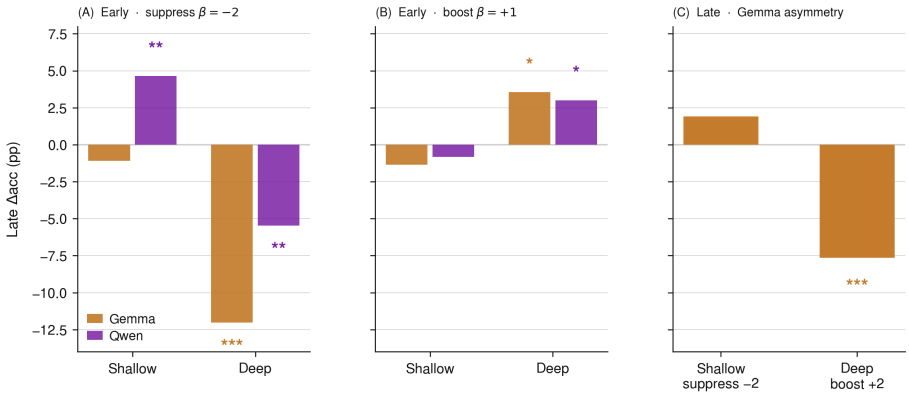

Figure 7: Layer-band localization (ABCD late turns; McNemar vs. natural; stars only where $p<0.05$). (A) Early suppress $\beta=-2$: Gemma's effect is deep (-12.0pp); Qwen's shallow ($+4.6$pp) and deep (-5.5pp) effects oppose and cancel under the all-layer probe. (B) Early boost $\beta=+1$: deep layers again carry the effect on both models. (C) Gemma's late window: suppression is null in both bands, while the harm from boosting ($\beta=+2$, -7.7pp) sits in the deep half. Full numbers in Appendix G.

5.4 Cross-Architecture Causal Transfer

We repeat the full three-window intervention on Qwen3.6-27B / ABCD ($n=754$; Figure 6; every late-turn contrast for both models in Appendix F). Qwen's natural late-turn baseline is 35.5%. In every window, the intervention moves attention mass by a magnitude comparable to or larger than on Gemma (early: mass $0.08/0.27/0.56$ at $\beta=-2/0/+2$; middle: $0.01/0.06/0.20$; late: $0.31/0.60/0.85$) – **but late-turn accuracy does not move in any window** (all $|\Delta| \lesssim 3$pp, all ns). Middle-window intervention does produce a significant *concurrent*-phase effect on Qwen (a reversal on middle-phase accuracy, $\approx \pm 11$pp, $p<0.001$) that does not reach the late phase, showing Qwen is not simply inert to the probe everywhere – strengthening rather than weakening the late-turn null.

This is a sharper claim than the original correlational dissociation: it is not that Qwen's attention fails to reallocate, but that reallocating it does not translate into late-turn behavioral change. Attention to early context is therefore **causally manipulable in both architectures, but behavior-determining in only one** – motivating the layer-level explanation below. One caveat is structural: on Qwen the dial reaches only the 16 softmax layers (Section 3.2), so the null means that biasing *softmax* attention does not move late-turn behavior, not that early context is unused.

5.5 Layer Localization

Where in the network does the early-window effect live? We restrict the bias to the shallow or deep half of layers only (Figure 7). On Gemma, the early-window effect is concentrated in deep layers in *both* directions: suppressing deep layers reproduces the harmful effect (deep -12.0pp***; shallow ns), and boosting deep layers reproduces the beneficial effect (deep $+3.6$pp*; shallow ns). Suppression localizes robustly (deep -12.0pp, $p=3.6\times10^{-8}$); the boost direction agrees but is weaker (deep $+3.6$pp, $p=0.024$).

On Qwen's early window, shallow and deep layers push in *opposite* directions: suppressing shallow layers helps ($+4.6$pp, $p=0.006$) while suppressing deep layers hurts (-5.5pp, $p=0.0045$); both survive Holm correction across all eight layer-band contrasts. Boosting shows a smaller, consistent asymmetry (shallow ns, deep $+3.0$pp, $p=0.043$). These opposing effects explain the all-layer null from Section 5.4 as **cancellation**, not indifference: Qwen's attention mass matters to its behavior in each direction individually, but a uniform bias averages the two away.

On Gemma's late window, suppression has no effect in either band (shallow $+1.9$pp, $p=0.28$; deep ≈0, ns), matching the null all-layer late suppression (-3.6pp, ns). The harm from boosting the late window is carried by the deep half ($\beta=+2$: deep -7.7pp***, shallow ns). Qwen's late-window bands are flat in both halves (shallow $+1.4$pp, deep $+0.5$pp, both ns), consistent with a genuinely flat, not

cancelling, late-window null. On Gemma, then, every layer-band effect we detect – early suppression, early boost, and late boost – is carried by the deep half of the network.

6 DISCUSSION

An attention trade-off. Early-context attention trades off against other positions: middle-window suppression *helps* middle-phase accuracy and late-window boosting collapses it, so more early attention is not always better.

Why does architecture matter? Architecture shapes both halves of our result. On Gemma, starvation itself is architectural: its sliding-window layers structurally lose early context as dialogues grow, leaving early commitments reachable only through 10 global layers and the residual stream. On Qwen, the most concrete candidate for the behavioral null is also architectural: 48 of Qwen's 64 token-mixing layers are Gated DeltaNet recurrences (Yang et al., 2025) that carry early-context information in a fixed-size state our softmax-level intervention never touches, so biasing the 16 attention layers can redistribute mass without cutting off the information. This makes a testable prediction: interventions on the recurrent state (e.g., scaling the DeltaNet gate for early positions) should move Qwen's late-turn behavior where the attention bias did not. Three further candidates remain: (1) long-context training – Qwen's 262K-token native context may favor extraction pathways less dependent on early-token attention magnitude; (2) redundancy – if Qwen has more redundant heads or pathways performing similar work, no single window's attention mass would be individually load-bearing; (3) a genuinely separate failure mechanism – Qwen still degrades substantially on ABCD (Appendix K), just not through this pathway. Distinguishing these requires value-path or residual-stream analysis.

Implications for evaluation and deployment. Standard single-turn benchmarks cannot detect a failure mode that only emerges over extended dialogue and depends on *where* attention is disrupted, not just whether it decays. Practically, our cross-architecture result argues against a one-size-fits-all mitigation: context-management interventions (summarization, retrieval, periodic refresh (Lewis et al., 2020; Packer et al., 2024)) target exactly the pathway that matters for Gemma-like architectures – for sliding-window models, keeping early commitments inside the local window is a direct remedy – but may do little for architectures where behavior does not track early-context attention mass in the same way.

7 LIMITATIONS

All causal tests are on ABCD; HRI-CUES is scored by absolute error on 1–5 ratings rather than binary correctness, so repeating them there needs a different paired test. Causal tests use non-thinking mode only. In both datasets the model reads and judges an existing dialogue rather than participating in it; whether the effects hold when the history contains the model's own turns is untested. Head-level localization is untested, and attention mass is an incomplete account of information flow (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019): our interventions test its causal role, not what the attended content is used for. Our measurements cover only softmax layers (Gemma's 60, of which 10 are global; Qwen's 16), not Qwen's 48 linear-attention layers; a fully-attentional model (e.g., dense Llama or Mistral) would separate architecture type from model family. Qwen's ABCD degradation, which does not track early-context attention through any pathway we test, remains open.

8 CONCLUSION

On real dialogue, Gemma 4's sliding-window layers lose sight of early context, and the attention that remains is causal for late-turn accuracy in a position-specific way: early is graded, middle reverses, late collapses at extremes. On Qwen3.6 the same dial moves mass but not behavior, because shallow and deep effects cancel. Whether early attention matters is thus architectural; whether unreachable early context persists in the residual stream (Dongre et al., 2026) is a natural next step.

AI USE STATEMENT

The research questions, ideas, and experimental design are the authors' own. We used LLM-based assistants for partial code assistance, for finding some of the relevant literature, and for drafting and editing the text. The authors checked all AI-assisted content, validating code through its outputs.

REPRODUCIBILITY STATEMENT

The attention intervention is specified in Section 3.3 and Appendix A: a pre-softmax additive bias on a window of keys, with no training. All datasets are public (links in Appendix M); both open-weight models are public checkpoints run with greedy decoding. Appendices A–M give the window definitions, layer splits, scoring rule, exact counts and p-values for every contrast, and the software environment.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, 2020.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. MT-Bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 7421–7454, 2024a.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 3119–3137, 2024b.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. MultiWOZ – a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5016–5026, 2018.

Derek Chen, Howard Chen, Yi Yang, Alexander Lin, and Zhou Yu. Action-based conversations dataset: A corpus for building more in-depth task-oriented dialogue systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3002–3017, 2021.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 276–286, 2019.

Vardhan Dongre, Joseph Hsieh, Viet Dac Lai, Seunghyun Yoon, Trung Bui, and Dilek Hakkani-Tür. When attention closes: How LLMs lose the thread in multi-turn interaction. *arXiv preprint arXiv:2605.12922*, 2026.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Google DeepMind. Gemma 4 model card. `https://huggingface.co/google/gemma-4-31B-it`, 2026.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Hanna Herasimchyk, Robin Labryga, Tomislav Prusina, and Sören Laue. A residual-aware theory of position bias in transformers. *arXiv preprint arXiv:2602.16837*, 2026.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*, 2024.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*, 2024.

Bahar Irfan, Jura Miniota, Sofia Thunberg, Erik Lagerstedt, Sanna Kuoppamäki, Gabriel Skantze, and André Pereira. Human-robot interaction conversational user enjoyment scale (hri cues). *IEEE Transactions on Affective Computing*, 17(2):1438–1453, April 2026. ISSN 2371-9850. doi: 10.1109/taffc.2025.3590359. URL http://dx.doi.org/10.1109/TAFFC.2025.3590359.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3543–3556, 2019.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. MT-Eval: A multi-turn capabilities evaluation benchmark for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 20153–20177, 2024.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs get lost in multi-turn conversation. In *International Conference on Learning Representations*, 2026.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 3469–3483, 2021.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. Memgpt: Towards llms as operating systems. 2024. URL https://arxiv.org/abs/2310.08560.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2022.

Zewen Qiang, Sendong Zhao, Haochun Wang, Bing Qin, and Ting Liu. Uncovering the role of initial saliency in U-shaped attention bias: Scaling initial token weight for enhanced long-text processing. *arXiv preprint arXiv:2512.13109*, 2025.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model. https://qwen.ai/blog?id=qwen3.6-27b, April 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 11–20, 2019.

Xinyi Wu, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. On the emergence of position bias in transformers. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2024.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving Mamba2 with delta rule. In *International Conference on Learning Representations*, 2025.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 36, 2023.

A Intervention Implementation and Full Early-Window Results

Why: the main text reports the early-window dose-response compactly; here we give the implementation and the full paired counts behind it, so each effect can be checked from its discordant pairs.

Implementation. The intervention patches the eager attention function in each model's Hugging Face Transformers implementation, adding a scalar logit bias β to the query–key score matrix (pre-softmax) for every key position in the target window, e.g. $[0, \lfloor 0.25 \times \text{seq_len} \rfloor)$ for the early window. Unless a layer band is specified, the bias is applied identically in every layer and head. No weights are changed and nothing is trained. We score the first action line of each output; Gemma often appends further text after it, which is ignored. Predicted actions are normalized before scoring (surrounding whitespace and trailing dash runs stripped) and compared with the gold action by exact match. On Gemma, the bias magnitudes $\{-2, -1, 0, +1, +2\}$ span a range that roughly halves to doubles early-context attention mass relative to the natural baseline (late phase: natural M=0.561; β=-2: M=0.347; β=$+2$: M=0.742). We verified this before the main run.

Table 1: Five-condition dose-response intervention, early window, full ABCD dataset (late phase, n=366). McNemar tests two-sided paired binomial.

Condition	**Late Acc.**	**Δ vs Nat.**	p	**Sig.**
Suppress (-2.0)	20.2%	-8.7 pp	9.1×10^{-6}	***
Suppress (-1.0)	24.0%	-4.9 pp	0.0014	**
Natural (baseline)	29.0%	—	—	—
Boost ($+1.0$)	33.9%	$+4.9$ pp	0.0051	**
Boost ($+2.0$)	32.5%	$+3.6$ pp	0.148	ns

Late phase (n=366). Table 1 gives accuracies and p-values. Figure 8 shows the ±2 triad with its McNemar discordant pairs (b: natural-wrong→intervention-correct; c: natural-correct→intervention-wrong). Suppress -2: b=10, c=42 (net $-32/366$=-8.7pp, p=9.1×10^{-6}). Boost $+2$: b=41, c=28 (p=0.15). The boost condition changes 69 of 366 decisions, so its null net effect reflects helped and hurt instances nearly offsetting, not an inert intervention.

Middle phase (n=325). Effects are larger in magnitude (Table 2), ruling out a late-turn ceiling artifact.

Table 2: Early-window intervention, middle phase (n=325), Gemma 4 31B / ABCD. b: natural-wrong→intervention-correct (helped); c: natural-correct→intervention-wrong (hurt).

Condition	**Mid Acc.**	**Δ vs Nat.**	p	(b, c)
Suppress (-2.0)	27.1%	-17.8pp	1.2×10^{-9}	$(18, 76)$
Suppress (-1.0)	37.2%	-7.7pp	0.0010	—
Natural	44.9%	—	—	—
Boost ($+1.0$)	53.8%	$+8.9$pp	5.7×10^{-5}	—
Boost ($+2.0$)	45.8%	$+0.9$pp	0.82	$(41, 38)$

Late-phase pairs for β= $\pm$ 1: suppress -1 $(6, 24)$, p=0.0014; boost $+1$ $(28, 10)$, p=0.0051. The headline beneficial effect is 28 instances helped against 10 hurt.

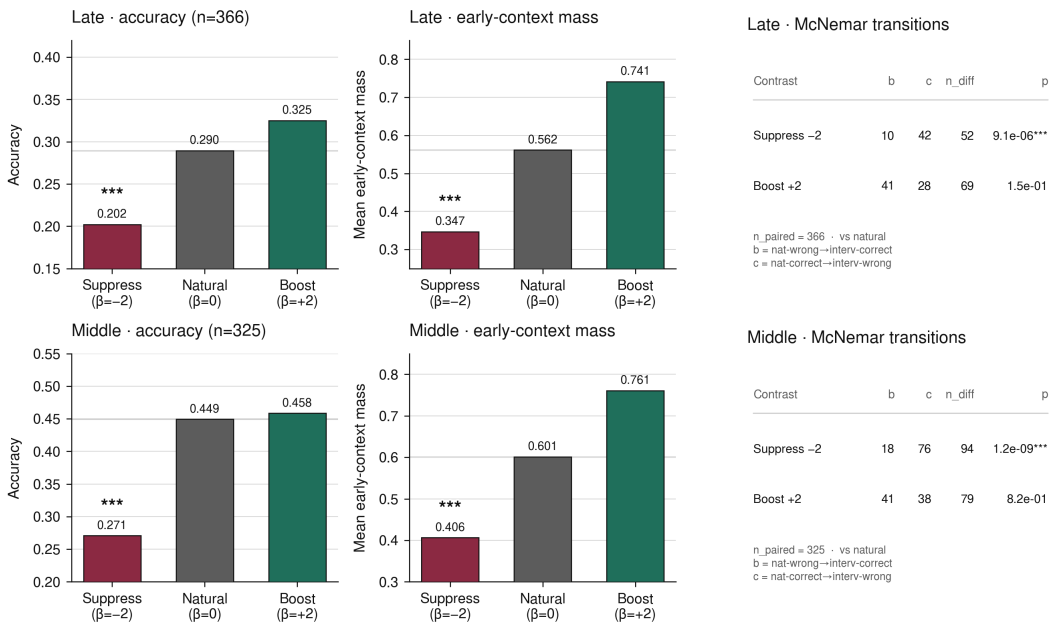

Figure 8: Early-window ± 2 triad (Gemma 4 / ABCD). Top: late phase (n=366); bottom: middle phase (n=325). Left: accuracy; centre: early-context attention mass; right: McNemar discordant pairs vs. natural. Stars: McNemar significance.

B HRI-CUES Per-Layer Measurement

Why: Gemma's sliding-window layers could make early-context decline a mechanical consequence of prompt length. This appendix tests that directly by measuring mass separately in local and global layers.

All HRI-CUES attention measurements are taken with Gemma's sliding-window mask enforced, i.e., on the model as deployed. A forward pass over 515 HRI-CUES prompts records early-context mass per layer. Among the 391 prompts longer than $\sim$1,365 tokens, where the early quarter lies outside every local window, local-layer early mass is zero in every case, and pooled mass never exceeds the bound $10/60 \approx 0.167$ implied by the 10 global layers. Pooled early mass falls from 0.183 to 0.093 between early and late turns; global-layer mass is flat ($0.552 \rightarrow 0.557$). The pooled decline is therefore produced by local layers losing sight of early context, not by global layers reallocating attention away from it.

Figure 9 separates the pools. Local-layer early mass falls from 0.110 at turn 4 to zero by about turn 10, when prompts outgrow the window; global-layer mass stays near 0.55 throughout; and the pooled curve is exactly their layer-weighted average, $(10\,M_{\text{global}} + 50\,M_{\text{local}})/60$ (e.g., 0.184 at turn 4, 0.093 on the plateau).

Reasoning-mode HRI-CUES runs and the attention-entropy trajectory were measured before the mask diagnostic and have not been remeasured, so we do not report them.

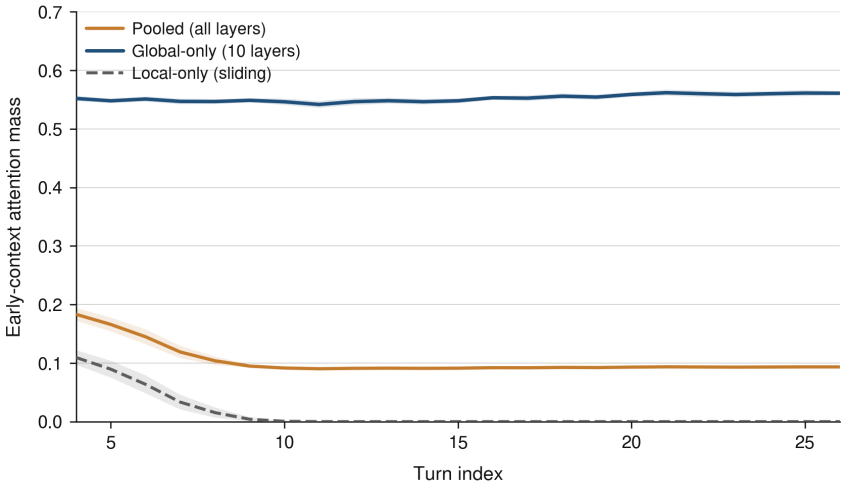

Figure 9: Mask-enforced early-context attention mass on HRI-CUES (Gemma 4 31B, standard mode; 515 turns, 391 above the $\sim$1,365-token bound; mean $\pm$ 95% CI). All three curves come from the same per-layer measurement: pooled over all 60 layers (0.183 at turn 4 to a 0.093 plateau), global layers 5, 11, …, 59 ($0.552 \rightarrow 0.556$), and the 50 sliding-window layers ($0.110 \rightarrow 0$).

C Per-Layer Attention Analysis (Descriptive)

Why: pooled attention mass could hide a decline driven by a few layers. This appendix shows how early-context mass is distributed across depth.

This appendix is *descriptive*: it reports where early-context mass sits across depth under natural inference. It is distinct from the *causal* layer-band intervention in Section 5.5, which changes attention in one half of the network and measures the behavioral consequence.

Figure 10 shows early-context mass for every Gemma layer and turn on a 120-instance ABCD subset. Mass differs strongly by layer: layers 0–3 hold $\approx$0.15–0.2, while layers $\approx$42–56 hold 0.8–0.9 in early turns. No layer is near zero at early turns, and darkening toward late turns is visible across most layers, most clearly in the deep band in later turns. The decline is therefore not produced by one outlier layer. Late-turn columns (turns 26 and 34) contain few instances and are noisy.

Figure 11 plots mean early-context mass per softmax attention layer on ABCD against relative depth (layer index / max), so models with different layer counts share an axis. Qwen is shown at its 16 softmax layers; its 48 Gated DeltaNet layers have no attention matrix. In thinking mode, Qwen keeps high early mass across depth ($\approx$0.55–0.75 until the final layer); in non-thinking mode, early mass falls with depth ($\approx 0.59 \rightarrow$ 0.23–0.27). Gemma's 60-layer profile is jagged, varying between $\approx$0.2 and $\approx$0.85; no layer is near zero. Dotted verticals mark Gemma's 10 global layers. In the shallow part of the network (relative depth $\lesssim$0.4) the global layers sit on local *peaks* of early mass; deeper, they sit on local *troughs* relative to the neighboring sliding-window layers. Several deep sliding-window layers hold early mass of 0.75–0.85, which is only possible if the early window lies inside their 1,024-token span. This matches the ABCD prompt lengths (mean 374, median 363 tokens): nearly all prompts fit inside the 1,024-token window, so on ABCD the sliding window does not cut off early context.

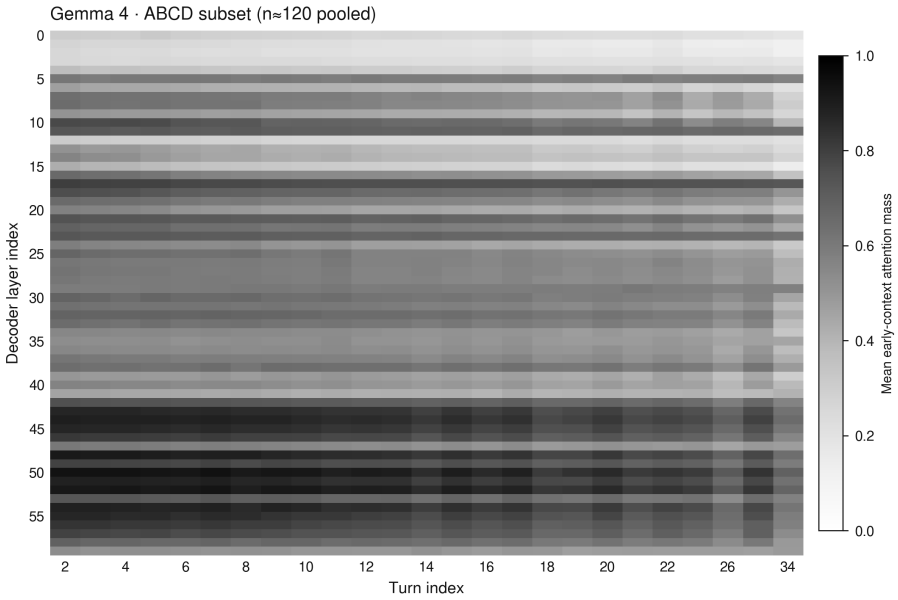

Figure 10: Early-context attention mass by decoder layer (rows) and turn (columns), Gemma 4 31B, ABCD subset (n≈120 pooled). Turn columns are the turn indices present in the subset.

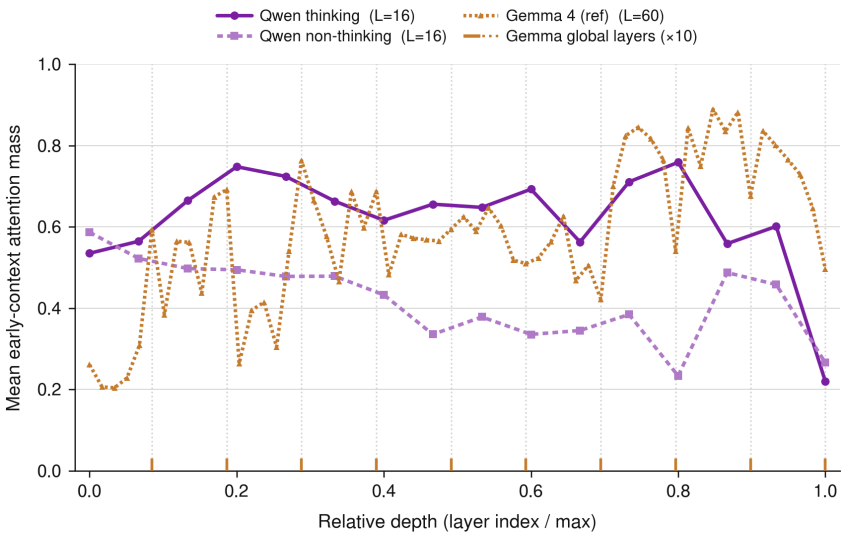

Figure 11: Per-layer early-context attention mass (ABCD, forward pass only, no intervention) vs. relative depth. Qwen thinking (solid) vs. non-thinking (dashed) at its 16 softmax layers; Gemma 4 (dotted, 60 layers) as reference, with its 10 global layers marked by vertical lines.

D Phase-Split Robustness

Why: the 20-60-20 phase split is a choice; this appendix checks that the ABCD degradation pattern does not depend on it.

We repeat the ABCD phase analysis under four phase-division schemes (Figure 12). Under 15-70-15, 20-60-20, and 30-40-30 the early phase is most accurate (0.67–0.82) and middle/late accuracy is lower (0.43–0.47 middle, 0.28–0.30 late). Under 10-80-10 the early bin holds a single instance (n=1), so only the middle/late contrast is informative (0.46 vs. 0.29). Middle and late accuracy are stable across all schemes ($\pm$0.02), so the degradation is not an artifact of the 20-60-20 boundary. This

analysis uses its own instance file, so its 20-60-20 early accuracy (22/27 = 0.82) differs from the main runs (Appendix L).

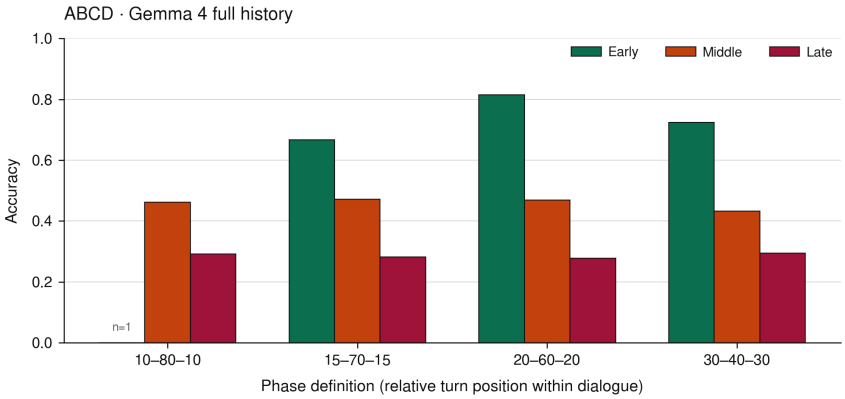

Figure 12: Phase-split robustness: ABCD accuracy (Gemma 4, full history) under four phase-division schemes. The 10-80-10 early bin has n=1 (marked); 20-60-20 early accuracy here is 0.815 (this analysis's instance file).

E Full Position-Specificity Results

Why: the main text shows late-turn effects only; this appendix gives both phases for the middle and late windows, verifies that each window's mass moved as intended, and shows unselected instance-level flips.

Early-window results for both phases are in Appendix A. Tables 3 and 4 give both phases for the middle and late windows (Gemma 4 31B / ABCD; natural baseline 0.290 late, 0.449 middle, identical across windows by construction). Before interpreting any new window, we checked that its β=0 condition reproduced the original run exactly.

Table 3: Middle window $[0.375, 0.625)$, both phases.

β	**Late acc.**	**Δ late (p)**	**Mid acc.**	**Δ mid (p)**
−2.0	0.303	+1.4pp (0.42)	0.511	+6.2pp (0.015)
−1.0	0.311	+2.2pp (0.12)	0.554	+10.5pp (1.4×10^{-7})
0	0.290	—	0.449	—
+1.0	0.257	−3.3pp (0.023)	0.277	−17.2pp (5.8×10^{-12})
+2.0	0.265	−2.5pp (0.18)	0.258	−19.1pp (4.7×10^{-12})

Table 4: Late window $[0.75, 1.0)$, both phases.

β	**Late acc.**	**Δ late (p)**	**Mid acc.**	**Δ mid (p)**
−2.0	0.254	−3.6pp (0.066)	0.351	−9.8pp (1.4×10^{-5})
−1.0	0.303	+1.4pp (0.42)	0.458	+0.9pp (0.664)
0	0.290	—	0.449	—
+1.0	0.268	−2.2pp (0.18)	0.418	−3.1pp (0.132)
+2.0	0.139	−15.0pp (2.4×10^{-11})	0.132	−31.7pp ($<10^{-15}$)

Mass verification. In every window, measured attention mass moves monotonically with β (Gemma, late turns; β=−2/−1/0/+1/+2): early 0.347/0.455/0.562/0.656/0.741; middle 0.009/0.023/0.051/0.105/0.175; late 0.200/0.285/0.384/0.492/0.591. The middle window receives far less mass than the others even at maximum boost (≈0.18 vs. 0.6–0.75), yet produces some of the largest behavioral effects. Its small share of attention is disproportionately load-bearing; we do not have a mechanistic explanation for this.

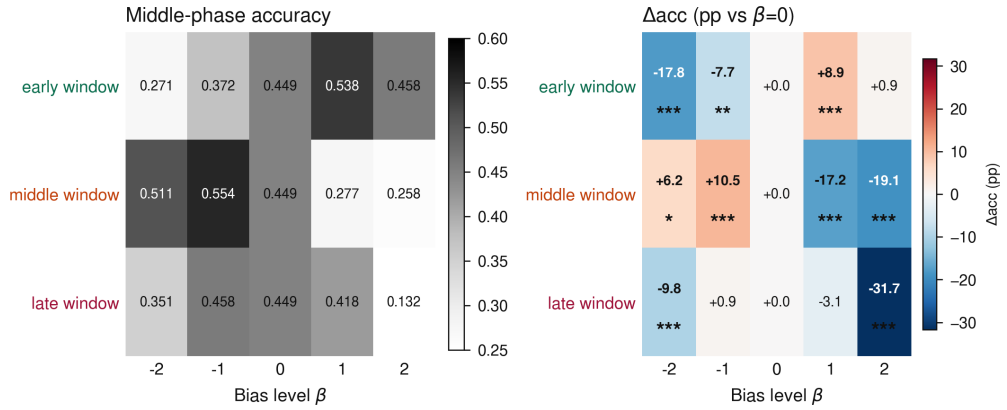

Figure 13: Middle-phase companion to Figure 5 (Gemma 4 / ABCD, n=325 middle turns). Left: accuracy per window and β. Right: Δaccuracy vs. β=0; stars: uncorrected McNemar.

Where the window differences come from. Most of the between-window spread comes from the boost direction. At β=+2 the three windows span ≈19pp of late-turn accuracy (0.325 early vs. 0.139 late); at β=−2 they lie within ≈10pp (0.202–0.303). The three-regime description is therefore driven chiefly by what happens when a window is *over*-weighted.

E.1 Unselected instance-level flips (middle window)

To show what the aggregate effects look like at the level of single decisions, Table 5 lists four middle-window flips taken *in file order* from the flip log: the first two suppress −2.0 flips and the first two boost +2.0 flips, with no reordering or selection. They are balanced two favorable / two unfavorable. Excerpts end at the turn where the model made its prediction.

Table 5: Unselected middle-window flips. Mass is middle-window attention mass, natural → intervention.

Conv.	β	Gold	Natural	Intervention	Mass
8309	−2.0	validate-purchase	verify-identity ✗	validate-purchase ✓	0.048 → 0.010
3048	−2.0	verify-identity	verify-identity ✓	update-account ✗	0.052 → 0.011
348	+2.0	ask-the-oracle	search-faq ✗	ask-the-oracle ✓	0.049 → 0.125
1884	+2.0	verify-identity	verify-identity ✓	pull-up-account ✗	0.035 → 0.170

Figure 14 shows the context for all four flips. Conversation 8309 (suppress, favorable): the account has already been pulled up and the open issue is a completed return, so the natural prediction (verify-identity) repeats a step the conversation has moved past. Conversation 3048 (suppress, unfavorable): only a name and phone number have been given, so verification must precede an account change, and suppression skips to update-account. Conversation 348 (boost, favorable): the latest question is subjective ("which brand makes me perfect?"), which an FAQ search cannot answer. This is the same instance as the main-text worked example (Figure 4), where the *early*-window boost β=+1 also corrects it, so the same error is fixed by two different interventions. Conversation 1884 (boost, unfavorable): the account was already pulled up and verification is the missing step, so boosting yields the redundant pull-up-account.



Figure 14: The four unselected middle-window flips (file order; key window $[0.375, 0.625)$). (A) 8309, suppress -2, favorable. (B) 3048, suppress -2, unfavorable. (C) 348, boost $+2$, favorable (also in Figure 4). (D) 1884, boost $+2$, unfavorable. Boxes: natural vs. intervention prediction; mass: middle-window attention mass, natural $\rightarrow$ intervention.

F Full Cross-Architecture Transfer Results

Why: the Qwen null is a central claim, so this appendix gives every Qwen contrast rather than a summary.

Qwen3.6-27B's natural baseline is 61.9% / 33.9% / 35.5% (Early/Middle/Late). The intervention moves window mass on Qwen by at least as much as on Gemma ($\beta{=}-2/0/+2$; early $0.08/0.27/0.56$, middle $0.01/0.06/0.20$, late $0.31/0.60/0.85$). Late-turn accuracy is flat in every window (all $|\Delta| \lesssim$ 3pp, all ns). The middle window produces a significant concurrent-phase reversal ($\approx \pm$ 11pp on middle-phase accuracy, $p{<}0.001$). Figure 15 shows every late-turn contrast for both models with 95% confidence intervals: all 12 Qwen intervals include zero, while Gemma's early-window suppression and boost, middle $+1$, and late $+2$ intervals exclude it.

Qwen's natural accuracy in the intervention run is 0.651 / 0.329 / 0.355 (Early/Middle/Late). Table 6 gives the $\beta{=}\pm 1$ late-turn contrasts with discordant pairs; the ± 2 contrasts appear in Figure 15. Window mass on Qwen ($\beta{=}-1/0/+1$): early $0.153/0.268/0.411$, middle $0.025/0.057/0.114$, late $0.449/0.600/0.738$.

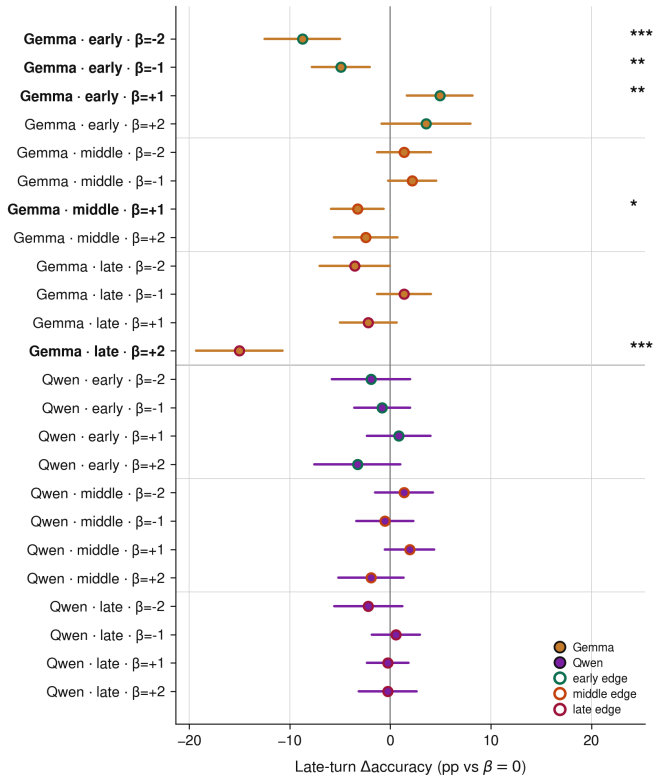

Figure 15: Evidence map: every late-turn causal contrast (window × β × model). Points are Δaccuracy (pp vs. β=0) with 95% CIs; stars mark uncorrected McNemar significance. Fill encodes model; marker edge encodes key window.

Table 6: Qwen3.6-27B, all-layer $\beta = \pm 1$, late turns (n=366). b: helped; c: hurt.

Window	β	**Δ late**	(b, c)	p
Early	-1	-0.8pp	$(12, 15)$	0.70
Early	$+1$	$+0.8$pp	$(19, 16)$	0.74
Middle	-1	-0.5pp	$(13, 15)$	0.85
Middle	$+1$	$+1.9$pp	$(14, 7)$	0.19
Late	-1	$+0.5$pp	$(11, 9)$	0.82
Late	$+1$	-0.3pp	$(7, 8)$	1.00

G Full Layer-Band Results

Why: the main text reports the layer-band contrasts the argument needs; this appendix gives all of them, including null and uninterpreted ones, so the selection can be checked.

Layers are split into a shallow and a deep half; the bias is applied only within one half, all else identical to the all-layer intervention. Figure 16 shows every layer-band contrast with 95% CIs, and Table 7 lists the early- and late-window values used in the main text.

Discordant pairs for the early-window deep band (b: helped, c: hurt; n=366): Gemma suppress $(11, 55)$, Gemma boost $(21, 8)$, Qwen suppress $(13, 33)$, Qwen boost $(18, 7)$. Natural late accuracy is 0.290 (Gemma) and 0.355 (Qwen) in every band run.

Output-format side effect. Shallow-only late-window suppression changed the *form* of Gemma's outputs: 97% of its late-turn predictions carried a trailing dash sequence (e.g., `validate-purchase--`), versus 0% in the deep-only and all-layer runs. Under our normalization (Appendix A) these score as the action they name; the chosen action is essentially unchanged

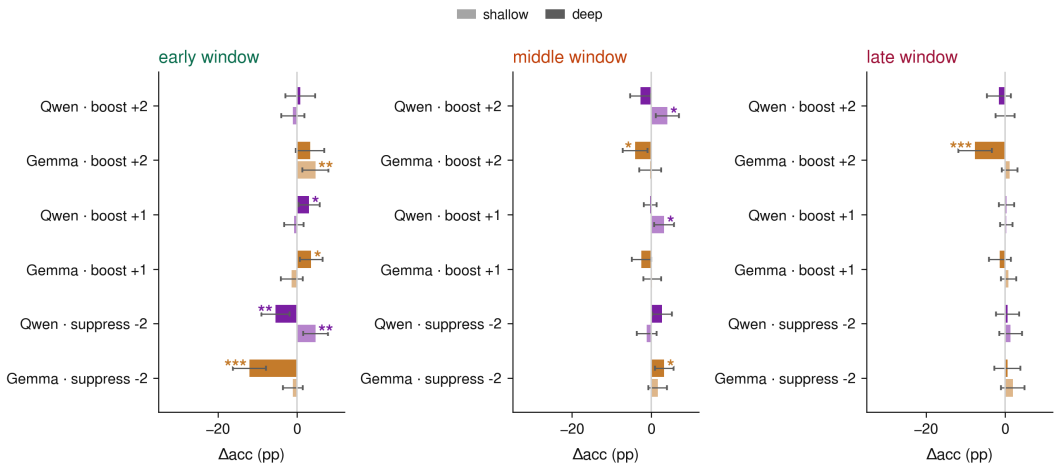

Figure 16: All layer-band contrasts (late-turn Δaccuracy vs. natural, 95% CIs, uncorrected McNemar stars), rows ordered boost +2, boost +1, suppress −2. Fill: model; dark: deep half (Gemma 30–59, Qwen 32–63); light: shallow half.

Table 7: Layer-band intervention, late-turn Δaccuracy vs. natural. ns: $p \geq 0.05$.

Model	Window	β	Shallow	Deep
Gemma	Early	−2.0	−1.1pp ns	−12.0pp (3.6×10^{-8})
Gemma	Early	+1.0	−1.4pp (0.44)	+3.6pp (0.024)
Qwen	Early	−2.0	+4.6pp (0.006)	−5.5pp (0.0045)
Qwen	Early	+1.0	−0.8pp (0.66)	+3.0pp (0.043)
Gemma	Late	−2.0	+1.9pp (0.28)	+0.5pp (0.87)
Gemma	Late	+1.0	+0.8pp (0.58)	−1.4pp (0.44)
Gemma	Late	+2.0	+1.1pp (0.42)	−7.7pp (6.2×10^{-4})
Qwen	Late	−2.0	+1.4pp (0.46)	+0.5pp (0.86)
Qwen	Late	+1.0	+0.3pp (1.00)	+0.3pp (1.00)
Qwen	Late	+2.0	0.0pp (1.00)	−1.6pp (0.38)

(+1.9pp, p=0.28). Shallow late-context attention thus affects output formatting without affecting the decision.

Boost +2. Figure 16 also shows β=+2 in every window. On Gemma's early window the +2 pattern differs from +1: the shallow half carries a significant gain (p<0.01) while the deep half is not significant. Together with the all-layer overshoot at +2 (Section 5.2), this suggests that strong early boosts act through a different depth than moderate ones; we report it but do not build on it. On Qwen's early window, β=+2 is null in both halves.

Middle window. Middle-window layer-band effects are small. Gemma deep suppression (+3.3pp, p=0.012), Qwen shallow boost +1 (+3.3pp, p=0.017), and single-star effects at β=+2 (Gemma deep, Qwen shallow) reach uncorrected p<0.05; none survives correction and we do not interpret them. Every layer-band contrast uses all 366 late turns; halves are Gemma layers 0–29 / 30–59 and Qwen 0–31 / 32–63.

H Length Stratification

Why: if early context matters more when there is less of it, the suppression effect should depend on prompt length. This appendix tests that.

We split the 366 late-turn instances into short, mid, and long context-length strata (token tertiles over all phases, cuts at ≈317 and ≈407 tokens; late-turn n=36/112/218) and recompute the early-window suppress (β=−2) effect within each (Figure 17). On Gemma the effect shrinks with length: −22.2pp in short contexts (n=36, p<0.05), −13.4pp in mid (n=112, p<0.01), and −4.1pp in long (n=218, ns). Qwen is near null in every stratum (−5.7, −0.9, −1.8pp; all ns).

All stratum boundaries lie far below Gemma's 1,024-token sliding window, so the window cutoff cannot explain the gradient. The more plausible reading is that in short contexts the early window holds a larger share of the information the next action depends on, so suppressing it costs more.

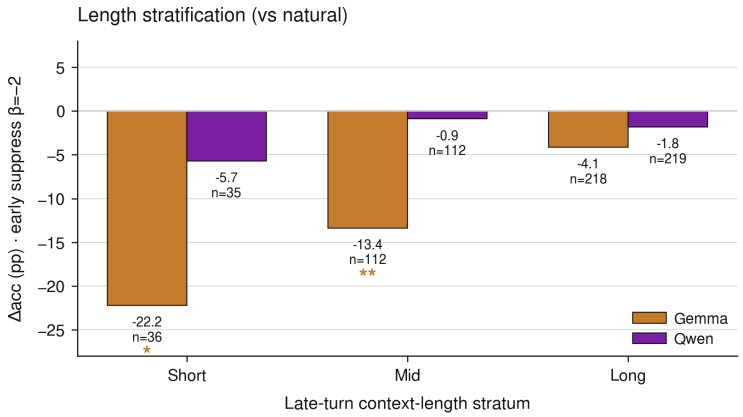

Figure 17: Early-window suppress (β=−2) late-turn Δaccuracy by context-length stratum. Stars: McNemar vs. natural within stratum.

I BOS/System-Token Composition Check

Why: if the early window consisted mostly of BOS or template tokens, early-context mass would measure attention sinks rather than conversational content. This appendix measures that share. We tokenize 200 randomly sampled ABCD prompts with Gemma's chat template and, for each, count the tokens in the early window (the first 25% of the prompt) that are BOS or any tokenizer special token (chat-template markers). BOS accounts for a mean 1.15% (median 1.11%) of the early window and all special tokens together for 2.31% (median 2.21%). The early window is therefore almost entirely conversational content, not structural tokens on which attention sinks typically form.

J Multiple-Comparisons Correction

Why: the paper reports many McNemar contrasts, so some uncorrected results will be significant by chance. This appendix shows which survive correction.

We apply Holm–Bonferroni at α=0.05 in two ways. **Family**: each window × phase is one family of four contrasts (the four non-zero biases vs. natural); this is the natural unit because each family answers one question ("what is the dose-response of this window in this phase?"). **Global**: all 24 all-layer contrasts on Gemma (3 windows × 2 phases × 4 biases) as a single family. The early-window late-phase family was the pre-existing confirmatory test; the middle and late windows were added afterwards as specificity tests.

Every qualitative claim in Section 5.3 survives global correction: early suppression hurts in both phases, middle-window suppression helps and boosting hurts in the middle phase, and the late-window +2.0 collapse holds in both phases. Two contrasts are fragile. The middle-window late-phase −3.3pp (p=0.023) does not survive either correction and should be read as directional only. The early-window late-phase boost +1.0 (p=0.0051) survives within its family but not globally; its middle-phase counterpart (+8.9pp, p=5.7×10^{-5}) survives both, so the claim that moderate early boosting helps rests on firm ground across phases.

Table 8: Holm–Bonferroni for all 24 all-layer contrasts (Gemma 4 31B / ABCD). ✓: significant after correction. Contrasts with uncorrected $p>0.05$ are omitted (none become significant under correction).

Window / phase	β	p (raw)	Holm (family)	Holm (global, 24)
Early / late	-2.0	9.1×10^{-6}	✓	✓
Early / late	-1.0	0.0014	✓	✓
Early / late	$+1.0$	0.0051	✓	–
Early / mid	-2.0	1.2×10^{-9}	✓	✓
Early / mid	-1.0	0.0010	✓	✓
Early / mid	$+1.0$	5.7×10^{-5}	✓	✓
Middle / late	$+1.0$	0.023	–	–
Middle / mid	-2.0	0.015	✓	–
Middle / mid	-1.0	1.4×10^{-7}	✓	✓
Middle / mid	$+1.0$	5.8×10^{-12}	✓	✓
Middle / mid	$+2.0$	4.7×10^{-12}	✓	✓
Late / late	$+2.0$	2.4×10^{-11}	✓	✓
Late / mid	-2.0	1.4×10^{-5}	✓	✓
Late / mid	$+2.0$	$<10^{-15}$	✓	✓

Layer band. We correct the early-window layer-band contrasts under two families. **All eight** contrasts (2 models $\times$ 2 biases $\times$ 2 bands): Gemma deep suppression ($p=3.6\times10^{-8}$) and Qwen's opposing shallow ($p=0.006$) and deep ($p=0.0045$) suppression effects survive, so both the Gemma deep localization and the Qwen cancellation account hold. The two deep boost effects (Gemma $p=0.024$, Qwen $p=0.043$) do not survive. **Deep band only** (four tests): all four survive. Since the eight-test family is the stricter and was specified first, the main text treats the boost-direction localization as consistent but not independently confirmed.

K Cross-Task Boundary Conditions and Qwen Dissociation

Why: this appendix places the Gemma/Qwen contrast in the context of all four datasets, including the two short-horizon controls.

Table 9 summarizes the observational picture across datasets. On HRI-CUES, Qwen shows monotonic but low-amplitude decay ($\Delta<0.06$; $\rho_{\text{med}}=-0.976$, 25/25 negative) and no behavioral degradation (Friedman $p=0.274$), despite the visible middle-phase rise in Figure 2. On ABCD, Qwen degrades substantially (61.9%→33.9%→35.5%) while its early-context mass is flat ($M\approx0.33$–0.35) and uncoupled from correctness. The causal-transfer result (Section 5.4) sharpens this: Qwen's attention *can* be moved, but moving it does not change late-turn behavior, and the layer-band analysis attributes this to cancellation. Qwen's ABCD degradation must therefore run through a pathway other than all-layer early-context attention mass. Both locally-grounded tasks (MultiWOZ, ESConv) show no behavioral degradation despite moderate decay, consistent with the phenomenon requiring long-horizon context integration.

Table 9: Cross-task and cross-architecture summary (observational).

Dataset	Model	Attn. decay	Degradation	Coupling
HRI-CUES	Gemma 4	Architectural (sliding window)	Yes ($p<.001$)	ns; ICC$=0.44$
HRI-CUES	Qwen 3.6	Low ($\Delta<0.06$)	No ($p=.274$)	$\rho=-.025$, ns
ABCD	Gemma 4	High (progressive)	Yes	$\rho=.267$, $p<10^{-13}$
ABCD	Qwen 3.6	Minimal (flat)	Yes	$\rho=-.015$, ns
MultiWOZ	Gemma 4	Moderate	No ($p=.784$)	$\rho=.011$, ns
ESConv	Gemma 4	Moderate	No ($p=.297$)	$\rho=.005$, ns

L Additional Statistical Details

Why: this appendix collects the statistical details behind the behavioral and observational results in the main text.

Mixed-effects model (HRI-CUES, 7,080 observations). Fixed effects: Middle vs. Early β=0.081, SE= 0.022, z=3.68, p<0.001; Late vs. Early β=0.030, SE= 0.027, z=1.11, p=0.202; Phase×Family both p>0.74. Random effects: conversation σ^2_u=0.030 (p<0.01); residual σ^2=0.156. Reasoning vs. standard variants: Mann–Whitney p=0.86/0.68/0.34 (Early/Middle/Late).

HRI-CUES mechanistic (mask enforced). Gemma pooled early mass $0.183 \rightarrow 0.093$; global layers $0.552 \rightarrow 0.557$; local layers zero on all 391 prompts above $\sim$1,365 tokens. Qwen (standard): $0.684 \rightarrow 0.648$ (5.2%); pooled Spearman ρ=−0.920; median per-participant ρ=−0.976, 25/25 negative, Wilcoxon p=5.96×10^{-8}.

ABCD coupling. n=753 instances from 278 conversations with ≥20 turns. ρ=0.267, p=9.3×10^{-14}; conversation-cluster bootstrap 95% CI $[0.20, 0.34]$ (B=2000). Robust across window fractions 0.20–0.30 (all $\rho\geq0.267$). Logistic regression: attention-mass coefficient -0.163, turn-index coefficient -0.770; partial Spearman controlling for turn ρ_{partial}=−0.020, CI $[-0.089, 0.048]$ (ns). This is why we treat the coupling as motivating only.

Consolidated phase-level results. Table 10. Three Gemma early-phase accuracies appear in the paper and come from different passes: 0.794 (50/63) from the full-history coupling run used here and in Section 4.3; 0.762 (48/63) from the natural (β=0) condition of the intervention run, which all causal contrasts are paired against; and 0.82 (22/27) from the phase-split analysis in Appendix D, which uses its own instance file. The two full runs agree within 3.2pp at every phase (intervention natural: 76.2% / 44.9% / 29.0%).

Table 10: ABCD phase-level accuracy and early-context mass (coupling run).

	Early (n=63)		**Middle** (n=325)		**Late** (n=366)		
Model	Acc	Mass	Acc	Mass	Acc	Mass	ρ
Gemma 4	.794	.638	.446	.601	.287	.562	.267***
Qwen 3.6	.619	.354	.339	.333	.355	.329	−.015 ns

Figure 2. The Gemma line (0.60 / 0.76 / 0.72) pools the standard and reasoning conditions. Qwen's middle-phase rise (1.03 / 1.23 / 1.10) is not significant (Friedman p=0.274; Appendix K).

Temperature sensitivity. Gemma re-run on a 200-instance ABCD subset at temperature 1.0: ρ=0.363 vs. greedy ρ=0.345.

Three-generation stability. Claude 3.7→4.5→4.6: ρ=0.96, p<0.001; slump depth 15.2%→14.3%→8.2%. GPT-4.5→5.2→5.5: ρ=0.94, p<0.001; slump depth 17.2%→6.8%→15.6%.

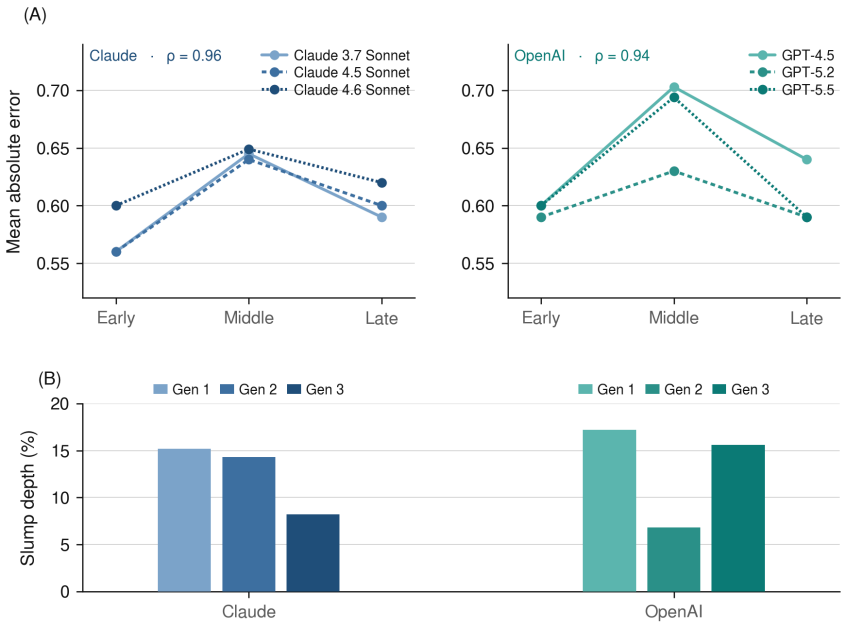

Figure 18: The pattern persists across generations (HRI-CUES). (A) Per-generation Early/Middle/Late MAE; ρ is shape stability across generations. (B) Slump depth by generation: Claude's shrinks, OpenAI's dips at Gen 2 and reverts, so temporal robustness does not track overall capability.

M Model and Dataset Specifications

Why: this appendix lists exactly which models, settings, and data were used, for reproduction.

Gen 1 (Q1–2 2025): Claude 3.7 Sonnet / 3.7 Extended Thinking; GPT-4.5 / o1. **Gen 2 (Q3–4 2025):** Claude 4.5 Sonnet / 4.5 Extended Thinking; GPT-5.2 / 5.2 Extended Thinking. **Gen 3 (Q1–2 2026):** Claude 4.6 Sonnet / 4.6 Extended Thinking; GPT-5.5 / 5.5 Extended Thinking. **Open-weight:** Gemma 4 31B, Qwen3.6-27B. Proprietary models at temperature 1.0; open-weight models with greedy decoding (`do_sample=False`). Total API calls: 7,080 (590 turns $\times$ 12 proprietary conditions) plus open-weight inference; API cost approx. $200. Environment: Python 3.11.13, PyTorch 2.5.1, transformers 5.7.0, statsmodels 0.13.2, scipy 1.9.0; one host with 8$\times$RTX 3090 (24 GB).

Table 11: Canonical enumeration of the 16 model conditions and which claims each subset supports.

Family	Conditions	Used for
Claude	3.7 / 4.5 / 4.6 Sonnet (+ Extended Thinking)	Longitudinal grid
OpenAI	GPT-4.5 / 5.2 / 5.5 (+ o1, Extended Thinking)	Longitudinal grid
Gemma	4 31B (standard + reasoning)	Longitudinal + mechanistic + causal
Qwen	3.6-27B (standard + reasoning)	Longitudinal + mechanistic + causal

HRI-CUES protocol. Each turn is scored by presenting the full conversation history up to and including the current turn as a numbered Swedish transcript and asking for a 1–5 quality rating. The same prompt is used for all families with no family-specific calibration. Ratings are extracted by regex; ties are broken toward the lower value ($<2\%$ of turns). Reasoning traces are excluded from the rating prompt.

Dataset links. HRI-CUES: `https://zenodo.org/records/12588810`. ABCD: `https://github.com/asappresearch/abcd` (8,034 conversations; 278 with $\geq$20 turns used). MultiWOZ and ESConv: public benchmarks.

Table 12: Dataset summary. Mean: mean turn index of evaluated instances. %≥18: share of instances at turn 18 or later.

Dataset	Dial.	Instances	Mean	%≥18	Label
HRI-CUES	25	590	12.6	29.3	5-pt quality; ICC= 0.44
ABCD	278	754	12.3	20.3	Logged next action (binary)
MultiWOZ	120	1,103	10.4	13.0	Token F1
ESConv	50	682	13.6	5.9	Empathy rating + F1