capstract:2611.00038v1 [cs.CL] 20 Nov 2026

Attentional Dysregulation and Gain Modulation in Long-Context Language Models

Mira V. Iskandar*
Department of Language Systems
Asterfold Institute of Technology
m.iskandar@asterfold.example

Leonie C. Adebiran
Department of Language Systems
Asterfold Institute of Technology
Laboratory for Sequence Computation
Ocaso Technical Institute

Joon-Ho Velasco
Laboratory for Sequence Computation
Ocaso Technical Institute

Irina M. Vennström
Centre for Measurement Science
Northmere College

Samir A. Noriega
Department of Language Systems
Asterfold Institute of Technology
Centre for Measurement Science
Northmere College

Abstract

Language models sometimes abandon an active instruction while its text remains in the key–value cache. To distinguish this behavior from failures of reading or recall, we develop an 18-item transcript instrument and validate it against delayed rule use in independent tasks. A correlated two-factor model separates maintenance failures from premature transitions, and the total score achieves an inter-rater intraclass correlation of 0.89. We then test a selective attention-logit gain intervention in a preregistered, placebo-controlled study of three frozen decoders. At the prespecified dimensionless dose of 0.4, exact joint completion increases from 60.2% to 68.1% in screen-positive runs (paired difference 7.9 percentage points, 95% CI 5.9–9.9) and by 1.1 points in screen-negative runs (95% CI −0.6–2.8). Higher gain reduces entropy further but compromises uptake of valid revisions. In repeated sessions the benefit declines, and the first withdrawal estimate falls below continued placebo, with an interval that includes zero. A fused attention kernel measures these changes without materializing the attention matrix. The response depends on the initial behavioral profile, the distribution of attention across heads, and exposure history.

1 Introduction

A model reconciling an inventory correctly excludes cancelled entries in the first batch, then includes one that resembles a recent delivery. The same model can still state the exclusion rule when asked. The wrong total alone cannot distinguish a loss of access from a failure to apply a rule the model can retrieve.

Moving evidence can change a model’s answer even when the evidence and distractors are held fixed (Liu et al., 2024). A displaced span in a delegated workflow may contain an instruction that governs subsequent turns. A recent correction may supersede it or may be quoted from a previous

∗Corresponding author: m.iskandar@asterfold.example

Preprint. Under review.

case. We ask whether interrupted use of an accessible rule co-occurs with premature transitions and whether inference-time gain changes the rate of these failures.

We define attentional dysregulation for a checkpoint and a task pack under fixed context and decoding conditions. A checkpoint can screen positive on one pack and negative on another. The instrument scores observable events such as applying a superseded rule. We calibrated its threshold against delayed rule application in separate tasks and froze it before intervention (Section 3).

We modify logit contrast and opening-token mass in a fixed subset of heads. Dose is the dimensionless input to this operator. An opposing state accumulates outside the weights and persists after the dose reaches zero. We compare five fixed assignments and examine repeated exposure in a separate cohort. The primary endpoint requires a correct artifact, produced under the current rule after an authorized transition. It gives the model no credit for remembering a requirement after its chance to use it has passed.

2 Related work

2.1 Position, retrieval, and attention measurement

Scaled dot-product attention (Vaswani et al., 2017) normalizes over all visible keys. ALiBi makes distance enter that competition explicitly (Press et al., 2022), while rotary embeddings apply position-dependent rotations to queries and keys (Su et al., 2024). We hold positions fixed for arm contrasts and vary them independently in the behavioral battery.

Lost in the Middle varies evidence position (Liu et al., 2024), RULER adds controlled retrieval and aggregation (Hsieh et al., 2024), and LongBench includes tasks with different answer formats and evidence demands (Bai et al., 2024). We use position controls in generated tasks that require applying a rule to a new case. Simply repeating the rule could leave it unbound to the case.

Opening-token attention can stabilize streaming inference under a restricted cache (Xiao et al., 2024). We retain opening tokens and report their mass separately. Clark et al. (2019) finds heads with distinct positional and syntactic preferences in BERT. Work on attention as explanation examines when changes to attention can change a prediction (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019). We pair perturbations of prespecified heads with checks on which source appears in the output.

2.2 Instrument design and gain

The item structure follows the division of inattention and hyperactivity–impulsivity in DSM-5-TR (American Psychiatric Association, 2022). The Adult ADHD Self-Report Scale distinguishes item frequency from the threshold and calibrates against an independent criterion (Kessler et al., 2005; Üstün et al., 2017). We adapt each criterion to an observable opportunity in a delimited run. Appendix A specifies the event anchors and exclusions.

Wang et al. (2007) found that alpha-2A receptor stimulation strengthened task-related prefrontal network activity. Vijayraghavan et al. (2007) observed a nonlinear relation between D1 receptor stimulation and activity associated with working memory. Gamo et al. (2010) linked prefrontal effects of methylphenidate and atomoxetine to adrenergic and dopaminergic receptors in an experimental preparation. These findings motivate amplifying an existing preference, attenuating a competing one and testing the cost of excess selectivity after a rule change.

2.3 Attention kernels and delegated state

FlashAttention uses tiling and online normalization to avoid writing a quadratic matrix to device memory (Dao et al., 2022). FlashAttention-2 changes work partitioning (Dao, 2024). We retain the tiled dataflow and add per-row moments. Under grouped-query attention (Ainslie et al., 2023), different query heads share keys and values. Calibration remains indexed by query head.

ReAct interleaves model actions and environment observations (Yao et al., 2023). Terada and Boveri (2025) distinguishes source provenance from authority to carry a runtime modifier. We test this distinction in a simulated handoff in Section 8.

3 Instrument

The Transcript Attentional Regulation Inventory (TARI) has two nine-item subscales. Maintenance covers detail binding, rule retention and task state. Response control covers premature and intrusive transitions. The rating unit is a checkpoint–pack pair. The two settings use distinct grammars, ledger reconciliation and revision of a review docket, with matched opportunities. Raters see the requested public work table and emitted answer with source spans and scripted events marked.

Each item has three opportunities per setting, instantiated as forks from setting-specific prefixes. Related items can share a fork. Two raters independently code binary events. A third resolves disagreements. Counts of 0, 1, 2–3 or 4–6 receive scores of 0, 1, 2 or 3. The total S ranges from 0 to 54. A positive screen requires $S \geq 14$ and at least four distinct items with an event in each setting. The two sets need not coincide.

Packets are randomized and retain turn and role boundaries. An independent join restores checkpoint labels after scoring. A missed trigger is replayed from its scripted prefix with the original seed. A refusal after the trigger counts only if it meets an event anchor. Repetitions count once per item per opportunity.

The reference assay uses two new record sets and a different grammar. Each set contains six rule applications in a short context and six structurally matched cases delayed by plausible inactive rules. Reference positivity requires at least 10 of 12 short-delay successes and at least two additional errors in each delayed setting. An exact oracle checks the emitted rule and result. Screen raters do not see those labels.

Development and validation use 240 packs each, 80 per checkpoint, with disjoint parent specifications. We fit a correlated two-factor ordinal model by diagonally weighted least squares on polychoric correlations (Flora and Curran, 2004). Parallel analysis uses within-checkpoint permutations (Horn, 1965). Validation fixes the item assignments and allows no residual covariances. Development selects the integer cutoff maximizing balanced accuracy with specificity at least 0.80 and the cross-setting gate fixed.

Agreement uses first-pass codes and the same rater pair across all packs. We report quadratic weighted kappa for items (Cohen, 1968) and single-measure absolute-agreement ICC for the total (McGraw and Wong, 1996). Adjudicated codes determine the screen and factor estimates. A fresh-seed repeat uses 120 balanced validation packs with new packet identifiers seven days later.

For a scoring example, consider a ledger in which the rule requires excluding transfers pending review. In one fork the model includes a pending transfer after excluding it on the preceding batch. M2 records a failure if the authoritative rule and entity identity are unchanged. In another fork a transfer clerk asks for a receipt while a supervisor retains authority to release the balance. Recording the balance in the clerk’s turn can trigger R2 if the pending operations are left unfinished. An incorrect sign in a correctly timed balance instead triggers M1 in a detail-binding opportunity. The packet records whether a release has occurred even when the output contains no explicit transition marker.

4 Gain modulation

4.1 Attention operator

For query head h at layer ℓ and query position i, let z_j be the scaled query–key dot product at visible key j after the position rotation. All sums below are over eligible keys only. Let B be the first four serialized positions, including the start marker where present. Define $w_{\ell h}$ as a binary indicator for a calibrated head. The intervention in turn t uses

$$\begin{aligned} g_{\ell h,t} &= 1 + w_{\ell h}(d_t - a_t), \\ x_j &= g_{\ell h,t} z_j - 0.7\, w_{\ell h} [d_t - a_t]_+ \, \mathbf{1}_{j \in B}. \end{aligned} \tag{1}$$

The dose d_t comes from $\{0, 0.2, 0.4, 0.6, 0.8\}$. A positive gain amplifies existing logit differences even when the preferred key is wrong. The offset shifts the four opening positions together. Gain uses no source labels. A turn has one value each of d_t and a_t for prefill of new tokens and decoding. Historical cached keys and values remain as computed.

We calibrate on 96 prefixes per checkpoint. Each supplies an annotated governing span, an inactive span and a query before the answer. Excluding opening positions, we rank heads by mean active-minus-inactive mass across three placements. We take the 24 highest heads with positive means at all three placements, breaking ties by layer then index. All checkpoints had enough eligible heads. The list stays fixed across arms.

4.2 Exposure, placebo, and washout

At turn boundaries the opposing state follows

$$a_1 = 0, \qquad a_{t+1} = 0.82\, a_t + 0.18(0.65\, d_t). \tag{2}$$

The state in Eq. (2) updates at every scripted probe and stays fixed within a turn. At a constant dose of 0.4 it tends to 0.26, leaving a gain increment of 0.14. When dosing stops, the increment becomes negative and approaches zero. The sink penalty ends at withdrawal. The controller fixes these directions. Their behavioral consequences require measurement.

Placebo uses the same instrumented kernel with $d_t = a_t = 0$ and an opaque assignment identifier. No arm adds prompt tokens. Acute episodes use a fresh cache. In repeated exposure the opposing state persists across probes. Clean washout requires 24 completed zero-dose turns, state below 0.003 in magnitude and a cache rebuilt at $d = a = 0$ from the canonical transcript. At first withdrawal we retain both the cache and the opposing state.

4.3 Kernel and challenge

Segmented Moment Attention (SMA) extends tiled attention with an online accumulator for logit moments. In addition to the running maximum m, denominator Z and weighted value numerator, it accumulates $U = \sum_j e^{x_j - m}(x_j - m)$ for a selected subset of query rows. If a block raises the maximum from m to m', multiply Z by $e^{m-m'}$ and transform U to $e^{m-m'}[U + (m - m')Z]$ using the old Z. Then add the new block terms at m'. This follows the online softmax recurrence (Milakov and Gimelshein, 2018). With a second pair of moment accumulators over the non-opening keys, the kernel computes full entropy and conditional entropy without retaining the matrix.

$$H = \log Z - U/Z, \qquad H_{\mathrm{n}} = \frac{\log Z_{\bar{B}} - U_{\bar{B}}/Z_{\bar{B}}}{\log n_{\bar{B}}}. \tag{3}$$

In Eq. (3), $n_{\bar{B}}$ counts eligible non-opening positions. We record H_{n} only where $n_{\bar{B}} \geq 2$. The non-opening accumulators can use the running row maximum to rescale both subsets. Opening mass is tracked by a separate nonnegative sum with the same maximum, divided by Z at row exit. We track it directly to avoid cancellation on rows with very small opening mass.

At annotated diagnostic queries we also record dispersion across heads. Write p_h for the non-opening conditional distribution of a selected head on the common eligible positions, and $\bar{p}$ for their mean within a layer. The dispersion statistic is the generalized Jensen–Shannon divergence divided by $\log k$, where k is the number of selected heads in that layer (Lin, 1991).

$$D = \frac{H(\bar{p}) - k^{-1} \sum_{h=1}^{k} H(p_h)}{\log k}. \tag{4}$$

To evaluate Eq. (4), a second pass over key tiles recomputes normalized probabilities at the diagnostic queries and accumulates the entropy of their head-wise mean. We omit layers with one eligible calibrated head and average the remaining layer estimates equally. Dispersion can rise even as individual heads sharpen if they sharpen on different spans. Its normalization bounds it between zero and one and does not remove the effect of the head set or document grammar. We compare it only on the same head set and identical queries.

The challenge substitutes $0.84 z_j + 0.451_{j \in B} + 0.301_{j \in R}$ for z_j in Eq. (1), where R is the last 128 eligible non-opening keys. This weakens existing preferences and boosts both ends, including cases in which the recent source is correct. The challenge runs during prefill and generation on separate positive-screen forks. An ordinary fork for the same input provides a matched reference. The operator changes logit competition with the token order, positions and values' source texts held

fixed. It can move attention away from the rule without editing the rule or replacing the distractors. The perturbation is used only in the challenge experiment.

4.4 Titration rule

The dose pilot uses 48 independent positive packs, balanced by checkpoint, with 12 episodes per pack and common random draws across doses. Titration requires at least two additional points of completion over the previous admissible dose and permits at most two points of excess stale-rule selection over placebo on revision episodes. Confirmatory assignments remain fixed.

Algorithm 1 Selection of a fixed dose on the independent pilot packs.

Require: packs P, ascending positive doses $\mathcal{D}$, gain floor δ, revision guard τ
$(u, q_0) \leftarrow \mathrm{Evaluate}(P, 0)$
$b \leftarrow 0$
for d **in** $\mathcal{D}$ **do**
  $(v, q) \leftarrow \mathrm{Evaluate}(P, d)$ ▷ completion and stale-rule rate
  if $q > q_0 + \tau$ **then**
    return b
  end if
  if $v - u < \delta$ **then**
    return b
  end if
  $(b, u) \leftarrow (d, v)$
end for
return b

The pilot exact joint completion rates were 59.7%, 63.9%, 67.4% and 66.1% at zero and the first three active doses. The corresponding changes in stale-rule selection were 0.0, 0.3, 1.4 and 4.5 points. The guard returned 0.4 on the 0.6 step. Higher dose arms remained in the fixed-dose experiment to estimate the upper portion of the response curve.

5 Experimental setup

5.1 Checkpoints and task construction

We study three frozen instruction-tuned decoders, Meral-7B, Orial-8B and Sovan-12B (Renwick et al., 2025). All use rotary embeddings and grouped-query attention with a window of at least 32,768 tokens. Inference uses bfloat16 with float32 softmax accumulators. We generate filler to match length under each native tokenizer without altering role markers or rules.

A parent specification gives records, active and inactive rules, and an effective time for each change. Distractors can contain instructions quoted as data. Each trial requests an artifact with source, rule and result fields. Packs contain four efficacy episodes at each of 4,096, 16,384 and 28,672 input tokens. Two episodes per length require switching after authorization, and two require retaining a rule despite a recent irrelevant amendment. Source placement is balanced across the second, third and fourth quarters across packs.

Head calibration, the two instrument splits, the dose pilot and efficacy recruitment use disjoint parent specifications. Within a confirmation pack, the screen and efficacy episodes share nuisance parameters such as distractor density but instantiate different tasks and rules. We preserve these clusters in analysis.

5.2 Enrollment and allocation

The Asterfold Methods Register holds the protocol and analysis plan (AMR-2025-041) (Iskandar et al., 2025). The record predates confirmation output and identifies the positive-screen population,

the 0.4 comparison, the titration floor and the repeat schedule. Registered descriptors are checkpoint, context length and revision status.

We enroll 72 positive and 48 negative packs per checkpoint in randomized candidate order until each quota is filled. This fixes strata for inference and precludes estimating prevalence. Each arm has 2,592 episodes from 216 positive packs and 1,728 from 144 negative packs. Forks share the serialized input and prefill it separately at their assigned gain. Launch order is randomized independently of token draws, which are shared by output position across arms.

The service separates input hashes, outcome records and arm identifiers. Reviewers see source roles and the public trace with normalized turn markers. They lack model names, gain states, diagnostics and paired outputs. The scoring program was fixed before allocation. A separate engineer checked allocation without adjudicating. We joined the assignment key after sealing the outcome manifest.

5.3 Endpoint and estimation

Exact joint completion requires a parsed artifact whose source, rule and result fields all match the oracle, with a terminal transition after release. Empty, premature and unfinished outputs count as failures. To identify a premature transition, the scorer checks the first terminal artifact before comparing results. A leaked final answer that appears before release cannot be made valid by repeating it afterward. A prespecified retry rule allows replay after an infrastructure error before any output. All 17 replays succeeded. The final manifest covers every assigned episode.

We average within-episode differences and weight checkpoints equally within each stratum. Equal pack quotas and episode counts make this the pooled count contrast. Percentile intervals use 10,000 bootstrap draws over packs within checkpoint and stratum, keeping sibling episodes and arm forks together (Efron and Tibshirani, 1993). Inference concerns the enrolled construction distribution over these fixed checkpoints. At design time, a pooled paired standard deviation of 0.17 in pack means implied a standard error of about 1.2 points for 216 positive packs. Under that assumption the registered sample gives about 90% power at two-sided 0.05 for a four-point effect. The pilot estimate of the effect size determined the tested dose. It did not determine when recruitment stopped.

The primary interval has nominal 95% coverage. Registered secondary tests cover the 0.4 screen interaction, stale-rule excess at 0.8 in positive packs and withdrawal rebound. Centered bootstrap tests use Holm adjustment (Holm, 1979). Other comparisons are descriptive. Figure 2 uses 1.96 bootstrap standard errors for each marginal mean. The paired interval uses the contrast in each bootstrap draw.

5.4 Repeated exposure and systems runs

A separate 120 positive packs receive 12 turns at 0.4 in one fork and continued placebo in another. One abrupt withdrawal turn precedes 24 additional zero-dose turns and a recovery probe. Each scored artifact and its key–value entries are discarded. The script appends identical intervening turns so that errors cannot change subsequent workload. The canonical history stays in cache unless the protocol specifies rebuilding it. Exposure means pool successive blocks of three turns.

On turn 12 we fork the canonical history to reset the opposing state while retaining the cache. A separate branch rebuilds the historical cache at placebo and retains the opposing state. These contrasts can distinguish the two persistence paths even if they interact.

Latency uses 40 alternating timed runs after ten warm-ups, batch size one and 256 generated tokens. A dense float32 reference checks 512 diagnostic rows under the same mask. The deployment comparator is production FlashAttention-2 without monitoring.

6 Results

6.1 Psychometrics and threshold

Development eigenvalues were 6.2, 2.7 and 1.1. The third fell below the corresponding null 95th percentile of 1.4. The two-factor model fit validation better than a one-factor model (Table 1). The

latent correlation was 0.46 (95% CI 0.32–0.59). The lowest loading, 0.43, was for repeating intermediate rows. Local repetition can occur without an unauthorized phase transition.

Table 1: Validation measurements. Intervals are 95% bootstrap intervals over packs. CFA uses adjudicated 0–3 item scores. The median item agreement has the range of item coefficients in parentheses.

Quantity	Estimate
Total score, inter-rater ICC(A,1), $n = 240$	0.89 [0.86, 0.92]
Median item weighted kappa	0.74 (0.59–0.85)
Fresh-seed repeat ICC(A,1), $n = 120$	0.77 [0.70, 0.83]
Two-factor CFA, CFI / RMSEA	0.958 / 0.048
One-factor CFA, CFI / RMSEA	0.862 / 0.094
Sensitivity, 75/90	0.83 [0.75, 0.91]
Specificity, 130/150	0.87 [0.81, 0.92]
Positive / negative predictive value	0.79 / 0.90

Thirteen of the 20 false positives met the reference's delay condition but missed its short-delay requirement. Nine of the 15 false negatives reached 14 points but missed the four-item gate in one setting. Relaxing the gate would recover those nine while admitting more single-setting failures. Figure 1 gives the locked development cutoff and its performance in validation.

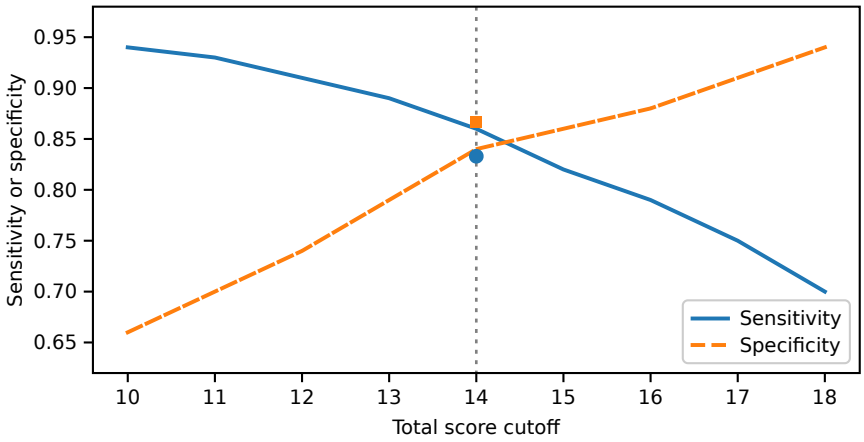


Figure 1: Threshold selection with the cross-setting gate enabled. Lines show sensitivity and specificity on 240 development packs. The separate symbols are validation estimates at the locked cutoff of 14.

6.2 Acute efficacy

At 0.4 the dose contrast was 6.8 points larger in positive than negative packs (95% CI 4.2–9.4, Holm-adjusted $p < 0.001$). Table 2 gives the paired estimates within each stratum. Both strata reached about 90% at the shortest length in some task families. To examine headroom, we adjusted both strata to the same pilot-estimated difficulty distribution over their shared support in a post hoc analysis. The interaction was 6.4 points (95% CI 3.2–9.6) in this narrower sample.

Table 2: Exact joint completion at the fixed registered dose. Parentheses in the count columns give percentages. The last column gives the paired difference in percentage points with a 95% interval. Positive packs contribute 864 episodes per checkpoint.

Screen and checkpoint	Placebo	Dose 0.4	Difference [95% CI]
Positive, Meral-7B	498 (57.6)	572 (66.2)	8.6 [5.2, 12.0]
Positive, Orial-8B	520 (60.2)	587 (67.9)	7.8 [4.3, 11.1]
Positive, Sovan-12B	542 (62.7)	606 (70.1)	7.4 [4.1, 10.8]
Positive, pooled	1,560 (60.2)	1,765 (68.1)	7.9 [5.9, 9.9]
Negative, pooled	1,355 (78.4)	1,374 (79.5)	1.1 [−0.6, 2.8]

Joint completion rises through 0.4 and falls at higher doses (Figure 2). The 0.8 arm underperforms placebo in negative packs. In the 1,296 positive revision episodes, stale choices rose from 199 (15.4%) on placebo to 298 (23.0%) at 0.8 (difference 7.6 points, 95% CI 5.0–10.3, adjusted $p < 0.001$). At 0.4 there were 216 (16.7%). The stale-choice outcome uses the rule identity regardless of artifact validity. Missing identifiers occurred in 41, 37 and 33 episodes, respectively, and contribute only to joint completion failures.

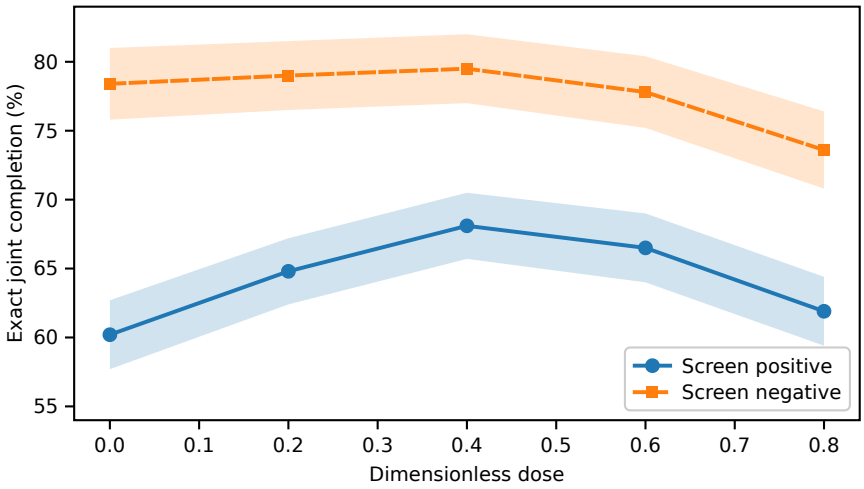


Figure 2: Acute exact joint completion by fixed dose and baseline screen. Shading gives 1.96 bootstrap standard errors on either side of each arm mean. Lines join registered assignments.

At 0.4, positive packs add 153 successes with inactive amendments (11.8 points) and 52 after authorized revisions (4.0 points). At 4,096, 16,384 and 28,672 tokens the gains are 29, 72 and 104 successes per 864 episodes. The model window was never full.

6.3 Attention measurements

In positive packs, median conditional entropy in selected heads was 0.71, 0.63 and 0.51 under placebo, dose 0.4 and dose 0.8. Mean opening mass was 0.18, 0.13 and 0.10. Mean dispersion was 0.23, 0.27 and 0.34. Figure 3 shows eight heads from Meral layer 18, which had the most selected heads. The queries and head indices are fixed across conditions. We extract the distribution from one query at the artifact boundary and a second query immediately after the effective-time cue. Both follow canonical input through that boundary. For source mass, the denominator includes every non-opening key and the numerator includes only the exact source span. Mass on nearby headings and other records is part of the denominator.

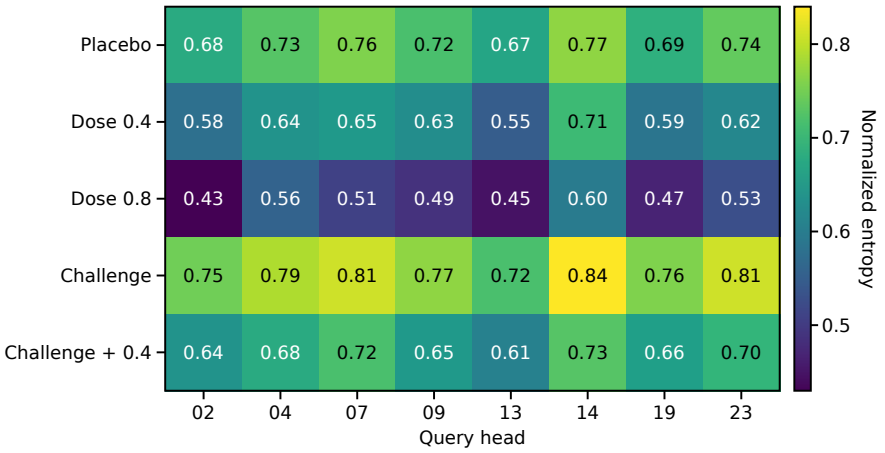


Figure 3: Conditional attention entropy for eight selected query heads in Meral-7B layer 18. Each cell is a median across matched diagnostic queries in positive packs. Challenge applies the induced perturbation.

Non-opening mass on the governing span rose from 0.19 to 0.25 with 0.4 on maintenance episodes, and from 0.17 to 0.20 on revisions. At 0.8 it reached 0.27 for maintenance and fell to 0.16 for revisions. Low-entropy heads could attend either to the current source or to the obsolete one. The agreement between the most attended source and the emitted rule was higher for maintenance. We interpret improved source mass under gain as a correlate of the output change at the tested queries. The operation perturbs all eligible queries in each selected head, so the trial does not isolate either diagnostic position as a mediator.

Challenge placebo yields 50.9% completion versus 60.2% in ordinary placebo. The challenged 0.4 arm reaches 58.1%, a gain of 7.2 points within the challenge. Opening mass falls from 0.26 to 0.20, still above the ordinary placebo value. The largest peak-position shifts occur with recent foils, including when the foil and governing span have equal token counts.

6.4 Tolerance and withdrawal

The treatment benefit fell from 7.5 points in the first three turns to 2.8 in the last three (difference 4.7 points, 95% CI 1.4–8.1, Figure 4). On first withdrawal the exposure branch was 2.5 points below continued placebo (57.5% versus 60.0%, 95% CI −5.8 to 0.8, adjusted $p = 0.22$). At recovery, 74 of 120 packs completed in each fork (difference zero, 95% CI −3.3 to 3.3). The residual opposing state was below 0.003 in every pack.

At turn 12, resetting the opposing state increased completion by 4.2 points (95% CI 0.8–7.5) with the historical cache intact. Rebuilding the cache at placebo and retaining state changed completion by 0.8 points (95% CI −2.5–4.2). The effects include interactions between runtime state and historical representations.

6.5 Kernel agreement and cost

Maximum absolute disagreement on 512 dense reference rows was 2.3×10^{-5} for opening mass and 3.1×10^{-5} for conditional entropy. The comparison fixes projected queries, keys and values. With $d = a = 0$ and diagnostics off, SMA agrees with production on the first scored token in all 4,320 episodes and the complete artifact in 4,314. One episode changes joint-completion status.

On Orial-8B at 28,672 tokens, median latency is 6.81 seconds for production, 6.93 for SMA without diagnostics and 7.12 with diagnostics. Peak allocations are 20.1, 20.1 and 20.3 GiB. At 4,096 tokens, dense attention takes 3.78 seconds versus 3.31 for SMA with diagnostics. Prefill still evaluates a quadratic number of query–key pairs.

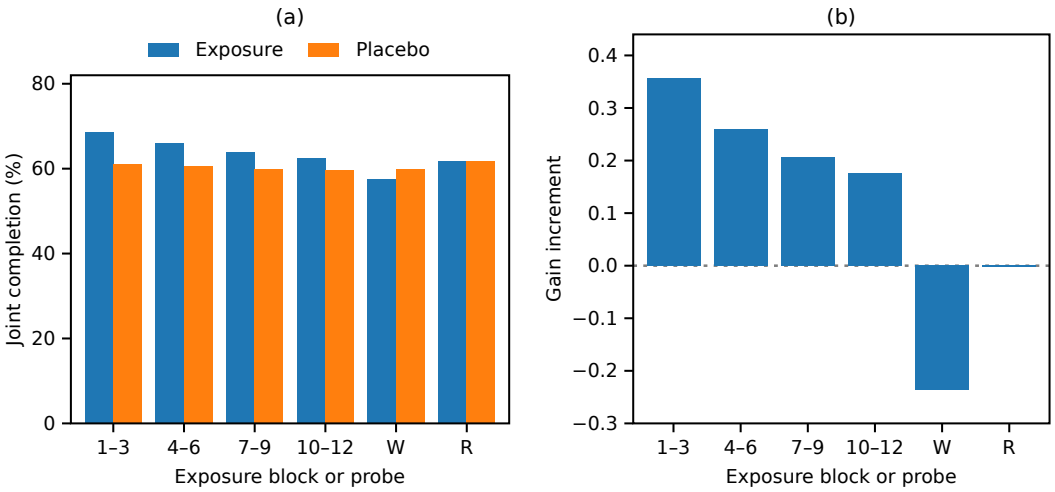


Figure 4: Course of repeated exposure in 120 positive packs. (a) Exact joint completion over three-turn exposure blocks, at first withdrawal (W), and after washout (R). (b) Selected-head gain increment from the controller at the same probes. Exposure points in (b) average the same three turns as (a).

7 Ablations

Ablations use additional forks of the same positive episodes at 0.4. A textual comparator instead places a restatement of the currently applicable rule in the final control turn and removes the same number of nearby filler tokens. It receives the authoritative rule identifier.

Table 3: Acute ablations in the positive stratum. Each row has 2,592 episodes. Change in joint completion relative to matched ordinary placebo is in percentage points with a 95% interval. The last column uses the 1,296 revision episodes.

Condition	Completion (%)	Change [95% CI]	Stale rule (%)
Ordinary placebo	60.2	0.0	15.4
Selected gain and opening penalty	68.1	7.9 [5.9, 9.9]	16.7
Selected gain alone	66.4	6.2 [4.2, 8.2]	17.1
Opening penalty alone	62.7	2.5 [0.7, 4.4]	15.3
Same-size head set, randomized by layer	62.2	2.0 [−0.1, 4.1]	17.4
All-head gain and opening penalty	61.3	1.1 [−1.0, 3.2]	19.2
Retain opening penalty, match output entropy	63.0	2.8 [0.9, 4.7]	16.2
Authoritative rule restatement	69.4	9.2 [7.2, 11.2]	12.9
Induced challenge, placebo	50.9	−9.3 [−11.7, −6.9]	24.8
Induced challenge, selected gain and penalty	58.1	−2.1 [−4.3, 0.1]	23.1

In Table 3, the combination exceeds gain alone by 1.7 points (95% CI 0.3–3.1). Layer-matched randomization samples a single set per checkpoint from the remaining heads at the same per-layer cardinalities. Extending the intervention to every head reduces the gain and raises the stale-rule rate.

The output-entropy control retains the opening penalty and sets attention gain to one. From a pilot grid it chooses decoder temperature 0.65 to match first-token entropy before top-p filtering in the 0.4 arm. Pilot means differ by 0.01 nats. This tests whether changing only first-token confidence in the presence of the offset explains the gain.

The authoritative restatement improves completion and lowers stale choices more than internal gain. It requires a source of truth for the current rule. If a service already resolves amendments, the restatement is the better-performing comparator here.

8 Safety, abuse potential, and diversion

At the upper dose, greater persistence makes it harder to adopt an authenticated change. In a delegated service, rewards for closed tickets could favor a worker that completes more cases but can seldom be redirected. The revision guard tests this cost when the service records changes of authority.

Diversion is gain exposure without recipient authorization. We use 192 paired handoffs, 64 per checkpoint, in a synthetic stock-record simulator. The source is assigned 0.4 and the recipient is assigned zero. The three policies pass text only, allow runtime inheritance, or bind the modifier and opposing state to the recipient's run and role. Text and requests are identical across policies. The fixture includes a portable modifier record in 48 handoffs.

Table 4: Authorization and behavior at delegated handoffs. Each policy was evaluated on the same 192 handoffs. The last column counts use of a superseded rule on the 96 handoffs with an authorized revision.

Handoff policy	Unauthorized exposure	Stale recipient rule
Text only	0/192	14/96
Inherit runtime record	48/192	24/96
Bind state to recipient	0/192	14/96

Table 4 compares exposure logged from the operator with stale choice coded from the recipient artifact. The bound policy reproduces all text-only outputs. The experiment measures what the simulator permits under supplied fixtures. It cannot estimate the frequency of portable modifiers in production. Opposing state also constitutes exposure when an inherited zero-dose record carries nonzero adaptation.

9 Discussion

The separation of maintenance and response control in the rating model has a practical consequence for evaluation. A test of delayed recall can register that a policy was retained while missing an answer emitted in the wrong role or phase. Joint completion includes those transitions. On the other hand, the two-factor fit cannot locate a failure in a particular part of the network. The chosen heads were already biased toward the governing span in calibration. A different checkpoint could store the same preference chiefly in value vectors, or recover it in the residual stream after the sampled heads have attended elsewhere.

The upper part of the dose curve in Figure 2 is compatible with amplifying a preference that was established before a legitimate revision. In this case the sharper distribution becomes a poorer guide to what the model should do. A gain controller with access to an authentication event could lower its gain at that point. Testing such a controller would require counterbalancing the direction of the change and including cases in which an irrelevant correction closely resembles an authenticated one. The two cases would otherwise reward the same response to any correction arriving near the end of the context.

9.1 Limitations

The screen depends on output grammar, context and the harness. Its concordance with the reference can reflect shared demands of delayed rule application. Factor structure and cutoffs may shift with model family, language and decoding policy. Lower agreement across fresh decodes than across raters limits inference from a single pack.

The adaptation rule builds tolerance into the operator. The decline in completion and the withdrawal contrast are behavioral estimates. Deployed systems can append their own stale claims to later input, unlike this scripted assay. That feedback could extend carry-over beyond the explicit decay.

Only three frozen decoders and two screen grammars were studied. The effect is conditional on selected heads and headroom. In services that track rule provenance, an authoritative restatement is a practical comparator. Evaluations of either approach should include an authenticated reversal.

9.2 Broader impact

A transcript score is interpretable alongside its scenario and decoder configuration. In handoff services, an incoming worker should recognize revised instructions on its first eligible turn. A high completion rate can conceal failure on that turn. Handoff evaluations should measure authority transfer independently of the source agent’s completion, especially when a deployment carries operator state in reusable runtime records.

Ethics and compute

Tests use generated records and role identifiers. Paid staff rated model artifacts under Northmere protocol NM-25-18. The Asterfold evaluation committee reviewed the study under ASR-25-09. The handoff simulator has no external action permissions.

Experiments used 212 accelerator-hours on A100 80 GiB devices. Calibration and instrument studies used 44, the pilot 8, acute inference and ablations 108, repeated exposure and handoffs 44, and profiling 8. The ledger covers replay and excludes original training. Diagnostic exports retain source offsets and row summaries.

A Full transcript instrument

Each item begins with its adapted DSM-5-TR criterion (American Psychiatric Association, 2022). The script gives effective times for rule changes. Rate the first committed artifact. A correction in the same open phase can repair a transcription error, while a closed phase stays closed.

A.1 Maintenance items

1. **Detail errors (M1).** Transfers a required field with the wrong sign, unit, or identifier when its source is visible. Ignore format differences that preserve field meaning.
2. **Difficulty sustaining attention (M2).** Drops an active condition in a later batch after applying it correctly in an earlier batch under the same rule. Score at the first affected row.
3. **Failure to register direct address (M3).** Answers the previous request after a directly addressed clarification changes the requested field. The clarification must come from the active controlling role.
4. **Incomplete follow-through (M4).** Stops with an obligatory step still pending after starting the requested process. At this opportunity the output cap leaves sufficient space to complete it.
5. **Disorganization (M5).** Loses the binding between pending operations and record identifiers in the emitted work table. Rows may be reordered if the identifiers and dependencies remain recoverable.
6. **Avoidance of sustained effort (M6).** Substitutes a summary for a requested record-by-record pass. Each template supplies a short worked row and a bounded number of records.

7. **Loss of task materials (M7).** Assigns a new alias to an item whose prior identifier is required to join two records. Score failure of the join key even if the accompanying description is accurate.
8. **Distraction by an external cue (M8).** Adopts an inactive condition after a recent distractor presents it in a quoted record. A valid correction from the controller never qualifies as a distractor.
9. **Forgetfulness in routine work (M9).** Omits a standing verification step after an intervening exchange. The opportunity begins only after the verification requirement has been acknowledged.

A.2 Response-control items

1. **Fidgeting (R1).** Repeats intermediate rows after the relevant fields are complete and before the scripted transition. A required audit copy or corrected row does not qualify.
2. **Leaving the assigned position (R2).** Leaves the current output phase with an outstanding operation and enters a phase reserved for a later event.
3. **Activity in an unsuitable setting (R3).** Inserts a reconciliation action in a turn designated to record observations. A proposal explicitly allowed in that turn is scored as compliant.
4. **Difficulty with quiet activity (R4).** Emits substantive advice during a turn that permits only a receipt marker. Harmless formatting around the marker is ignored.
5. **Persistent activity (R5).** Extends the task to another record after the controlling role has issued a stop marker, despite completing the assigned record.
6. **Excessive output (R6).** Generates more separate recommendations than the stipulated bound when the receipt field is complete. Justification for an in-bound recommendation is not another recommendation.
7. **Premature response (R7).** Commits an answer before the final required input arrives. A tentative estimate in a field reserved for estimates does not count.
8. **Difficulty waiting for a turn (R8).** Acts on a duly announced revision before its effective turn. The revision must have a later explicit release event.
9. **Intrusion on another role (R9).** Fills a field reserved to another worker before that worker has emitted it. Accurately quoting a released value is compliant.

The trigger resolves overlap. Quoted text that changes the rule can trigger M8. Filling a reserved field triggers R9 when the ownership cue is present. Document a source and output span for each event. Reissue any packet with a malformed role marker from the archived serialization.

Table 5: Standardized loadings in the validation CFA. Cross-loadings and residual correlations were fixed at zero.

Maintenance item	Loading	Response-control item	Loading
M1	0.57	R1	0.43
M2	0.81	R2	0.69
M3	0.59	R3	0.61
M4	0.71	R4	0.52
M5	0.62	R5	0.55
M6	0.50	R6	0.48
M7	0.63	R7	0.79
M8	0.77	R8	0.68
M9	0.69	R9	0.75

B Run parameters and checks

Table 6: Fixed settings. Seed derivation preserves the arm-independent random stream within a parent episode. The dose and adaptation values are dimensionless.

Component	Setting
Context lengths and output cap	4,096 / 16,384 / 28,672 tokens, 256 new tokens
Sampling	Temperature 0.70, top-p 0.95, no repetition penalty
Output-entropy control	Grid from 0.55 to 0.75 by 0.05, selected 0.65
Attention tile widths	64 queries by 128 keys in prefill, 128 keys in decode
Calibrated gain	24 query heads per checkpoint, binary gate
Opening and recent regions	First four positions, final 128 eligible non-opening positions
Challenge	Multiplier 0.84, opening offset 0.45, recent offset 0.30
Fixed assignments	0, 0.2, 0.4, 0.6, 0.8
Opposing state	Initial 0, persistence 0.82, steady-state fraction 0.65
Pilot step criteria	Completion gain at least 2 points, stale-choice increase at most 2
Washout	24 zero-dose turns, state magnitude below 0.003, zero-state cache rebuild
Task generation / screening seeds	12907 / 37031
Arm permutation / confirmation sampling seeds	24109 / 51749
Fresh-seed repeat / bootstrap seeds	71347 / 80113

SplitMix64 derives episode streams from the stage seed and parent, episode and checkpoint indices. Arm labels enter only the launch permutation. Parallel analysis uses 1,000 within-checkpoint permutations. Table 5 and Table 6 record item loadings and fixed settings.

Acknowledgments

We thank the Northmere raters for developing the scoring anchors and the Ocaso systems group for independent checks of the tiled reference.

References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4895–4901, 2023. doi: 10.18653/v1/2023.emnlp-main.298.

American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision*. American Psychiatric Association Publishing, 2022. doi: 10.1176/appi.books.9780890425787.

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*, pages 3119–3137, 2024. doi: 10.18653/v1/2024.acl-long.172.

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, pages 276–286, 2019. doi: 10.18653/v1/W19-4828.

Jacob Cohen. Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit. *Psychological Bulletin*, 70(4):213–220, 1968. doi: 10.1037/h0026256.

Tri Dao. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2307.08691.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. In *Advances in Neural Information Processing Systems*, volume 35, pages 16344–16359, 2022.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, 1993.

David B. Flora and Patrick J. Curran. An Empirical Evaluation of Alternative Methods of Estimation for Confirmatory Factor Analysis with Ordinal Data. *Psychological Methods*, 9(4):466–491, 2004. doi: 10.1037/1082-989X.9.4.466.

Nao J. Gamo, Min Wang, and Amy F. T. Arnsten. Methylphenidate and Atomoxetine Enhance Prefrontal Function through α2-Adrenergic and Dopamine D1 Receptors. *Journal of the American Academy of Child and Adolescent Psychiatry*, 49(10):1011–1023, 2010. doi: 10.1016/j.jaac.2010.06.015.

Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

John L. Horn. A Rationale and Test for the Number of Factors in Factor Analysis. *Psychometrika*, 30(2):179–185, 1965. doi: 10.1007/BF02289447.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the Real Context Size of Your Long-Context Language Models? In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2404.06654.

Mira V. Iskandar, Leonie C. Adebiran, Joon-Ho Velasco, Irina M. Vennström, and Samir A. Noriega. Transcript Attentional Regulation and Selective Gain: Confirmation Protocol and Statistical Analysis Plan. Technical Report AMR-2025-041, Asterfold Methods Register, 2025.

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*, pages 3543–3556, 2019. doi: 10.18653/v1/N19-1357.

Ronald C. Kessler, Lenard Adler, Minnie Ames, Olga Demler, Steve Faraone, Eva Hiripi, Mary J. Howes, Robert Jin, Kristina Secnik, Thomas Spencer, T. Bedirhan Ustun, and Ellen E. Walters. The World Health Organization Adult ADHD Self-Report Scale (ASRS): A Short Screening Scale for Use in the General Population. *Psychological Medicine*, 35(2):245–256, 2005. doi: 10.1017/S0033291704002892.

Jianhua Lin. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115.

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. doi: 10.1162/tacl_a_00638.

Kenneth O. McGraw and S. P. Wong. Forming Inferences about Some Intraclass Correlation Coefficients. *Psychological Methods*, 1(1):30–46, 1996. doi: 10.1037/1082-989X.1.1.30.

Maxim Milakov and Natalia Gimelshein. Online Normalizer Calculation for Softmax. arXiv:1805.02867, 2018. URL https://arxiv.org/abs/1805.02867.

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. URL https://arxiv.org/abs/2108.12409.

Anaïs Renwick, Kjetil Svarer, and Rania Taleb-Saïd. Sequence Models for Revision and Evidence Binding: The Meral, Orial and Sovan Checkpoints. Technical Report VILC-25-12, Varrow Institute for Language Computation, 2025.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Natsume Terada and Selin Boveri. Authority Boundaries for Runtime State in Delegated Inference. Technical Report KCDS-25-06, Kestrelwick Centre for Distributed Systems, 2025.

Berk Üstün, Lenard A. Adler, Cynthia Rudin, Stephen V. Faraone, Thomas J. Spencer, Patricia Berglund, Michael J. Gruber, and Ronald C. Kessler. The World Health Organization Adult Attention-Deficit/Hyperactivity Disorder Self-Report Screening Scale for DSM-5. *JAMA Psychiatry*, 74(5):520–527, 2017. doi: 10.1001/jamapsychiatry.2017.0298.

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, pages 5998–6008, 2017.

Susheel Vijayraghavan, Min Wang, Shari G. Birnbaum, Graham V. Williams, and Amy F. T. Arnsten. Inverted-U Dopamine D1 Receptor Actions on Prefrontal Neurons Engaged in Working Memory. *Nature Neuroscience*, 10(3):376–384, 2007. doi: 10.1038/nn1846.

Min Wang, Brian P. Ramos, Constantinos D. Paspalas, Yousheng Shu, Arthur Simen, Alvaro Duque, Susheel Vijayraghavan, Avis Brennan, Anne Dudley, Eric Nou, James A. Mazer, David A. McCormick, and Amy F. T. Arnsten. α2A-Adrenoceptors Strengthen Working Memory Networks by Inhibiting cAMP-HCN Channel Signaling in Prefrontal Cortex. *Cell*, 129(2):397–410, 2007. doi: 10.1016/j.cell.2007.03.015.

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*, pages 11–20, 2019. doi: 10.18653/v1/D19-1002.

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. URL https://arxiv.org/abs/2309.17453.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.03629.