ALERT: ATTENTION-BASED LONG-CONTEXT EARLY-WARNING VIA ROUTING TRAJECTORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Long conversations can retain relevant evidence while a language model answers incorrectly. We introduce ALERT, which predicts retrieval failure before the first answer token. A learned ranker selects one history message using attention at the final prompt position; a second probe predicts failure from that source's attention profile. The LLM remains frozen, and warnings trigger re-injection of selected evidence while preserving the full history. Paired evaluations through 32K separate localization, failure prediction, and recovery across NoLiMa-style tasks, BABILong, turn-based RULER, and Scatter-MCQ. In an initial confirmation on 576 new BABILong stories at 16K, frozen monitors achieve failure AUROC 0.710 for Gemma-4-E4B and 0.759 for Qwen3-4B, versus 0.518 and 0.517 for raw next-token margin. For these two models, respectively, frozen-monitor AUROCs are 0.941 and 0.802 on RULER, and 0.672 and 0.662 on Scatter-MCQ. Selective re-injection improves accuracy by 14.9 and 17.0 percentage points on BABILong, and 9.5 and 5.7 points on Scatter-MCQ, respectively. These results support pre-generation failure prediction and net recovery, although neither BABILong policy exceeds always-re-injection accuracy. Source-selection errors and frequent warnings remain constraints on effective selective intervention.

1 INTRODUCTION

We study *pre-generation detection of retrieval failures*: predicting answer correctness before the first answer token. A user states that Sam left the key in a drawer, then asks for its location after a long conversation (Figure 1). The statement remains available, yet the model may answer incorrectly. An early warning would enable the assistant to revisit relevant evidence before answering.

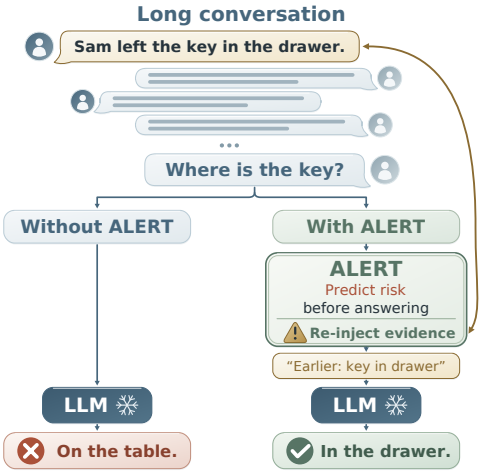


Figure 1: Evidence can remain in context while the answer is wrong. ALERT warns before decoding and repeats selected evidence, preserving the history.

Long-context benchmarks test evidence position (Liu et al., 2024), indirect retrieval cues (Modarressi et al., 2025), and facts in distracting text (Kuratov et al., 2024). Internal analyses identify retrieval heads (Wu et al., 2025). Lookback Lens uses attention ratios between input context and generated tokens to detect contextual hallucinations (Chuang et al., 2024), motivating attention as a signal for monitoring evidence use.

Before generation, however, a monitor cannot use answer tokens and must locate evidence without its annotation. Localization alone is insufficient: our diagnostic examples show attention to supporting evidence followed by incorrect answers. Attention describes information routing without establishing successful evidence use (Jain & Wallace, 2019). A pre-generation monitor must therefore link the model's treatment of a source to subsequent answer correctness. This distinction makes source localization and warning quality separate evaluation targets.

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.

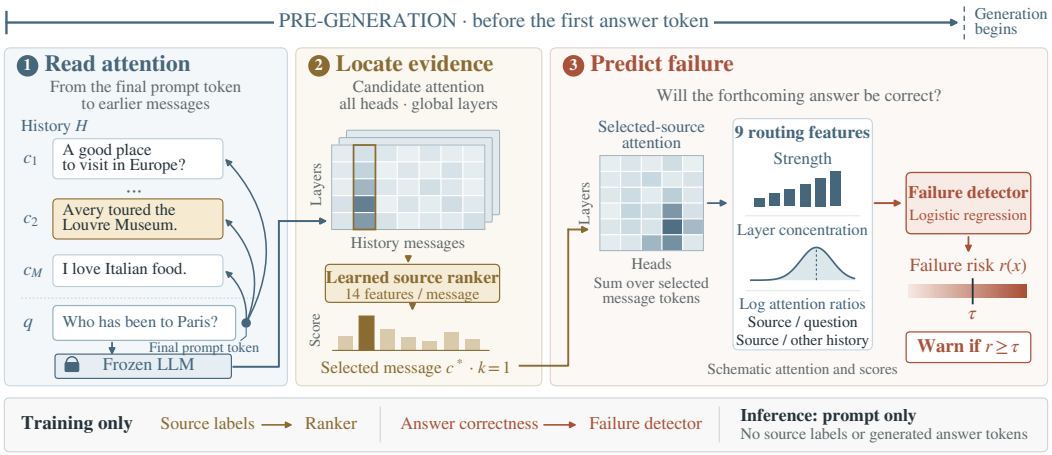


Figure 2: ALERT: predict answer failure before generation. 1 Read attention from the final prompt position to history messages. **2** Rank messages and select one candidate source ($K = 1$). **3** Predict failure from its routing features; warn when risk $r(x) \geq \tau$, with τ fixed on calibration data. Supervision is used only during fitting. Maps and scores are schematic; depicted layers have global attention.

ALERT predicts retrieval failure from attention to an automatically selected source (Figure 2). A learned ranker selects a history message; a separate probe estimates answer risk from its attention strength, distribution across layers, and competition with other context. The LLM remains frozen. Source annotations train the ranker, and subsequent answer outcomes train the failure probe; neither is supplied at inference.

> Attention to evidence does not guarantee a correct answer. ALERT separates source selection from failure prediction, then uses the warning to trigger evidence re-injection.

To test whether warnings improve answering, we use *selective re-injection*: repeat selected evidence before the question when the detector signals failure. The full history stays intact. For state-tracking queries, the excerpt identifies its original position so an earlier event is not presented as a new update. Recovery depends on both the selected evidence and the warning policy.

We evaluate four benchmarks adapted to multi-turn conversations: NoLiMa-style associative recall, BABILong state tracking, RULER needle-in-a-haystack retrieval, and our Scatter-MCQ benchmark for scattered multiple-choice recall. Contexts span 8K, 16K, and 32K within each model's validated limit. Matched short-context controls for BABILong and Scatter-MCQ assess underlying task competence. Component-wise analyses distinguish source localization, failure prediction, and intervention outcomes.

In an initial frozen-method confirmation on 576 new BABILong stories at 16K, Gemma-4-E4B achieves failure AUROC 0.710, versus 0.518 for raw next-token margin (one minus the top-two probability gap). Selective re-injection raises accuracy from 28.6% to 43.6%, but intervenes on 91.5% of prompts; always re-injecting reaches 44.4%. This supports pre-generation discrimination and net recovery on that cohort, while warning selectivity remains unresolved. We therefore assess whether risk scores distinguish failures separately from how often the calibrated rule warns on correct answers and triggers intervention.

We make four contributions:

- **Automatic evidence selection:** an attention-based ranker selects a candidate source without source annotations at inference.
- **Pre-generation failure detection:** a separate probe predicts answer failure from the selected source's attention profile before decoding.
- **Selective evidence re-injection:** a frozen warning rule triggers repetition of selected evidence while preserving the history and event order.
- **Retrieval benchmarking and analysis:** a paired four-benchmark protocol varies context length while retaining evidence, and separates localization, detection, rescues, harms, and intervention cost.

2 Related Work

Failures within long and multi-turn contexts. Longer input capacity does not ensure reliable use of retained information. Liu et al. (2024) show that answer accuracy depends on evidence position, with relevant information in the middle of a context often used less reliably. NoLiMa removes literal matching cues to test associative retrieval (Modarressi et al., 2025); BABILong embeds reasoning facts within long distractor text (Kuratov et al., 2024); and RULER extends needle retrieval with tasks requiring tracing and aggregation (Hsieh et al., 2024). Conversational failures introduce further sources of difficulty: Laban et al. (2026) find that models can commit to premature solutions and fail to recover as instructions arrive over multiple turns, while **?** measure sensitivity to dialogue histories prepended to a question. Du et al. (2025) additionally report degradation despite successful evidence retrieval. These settings motivate evaluating retained evidence, source localization, and answer correctness separately. Our conversational adaptations use paired contexts to study whether an individual failure can be anticipated before decoding, and whether acting on that warning improves the answer.

Internal representations and failure prediction. Hidden representations can encode information that is not expressed in an answer. Gao et al. (2024) recover the location of relevant information from hidden states even when the model fails to answer correctly, demonstrating a gap between internal accessibility and successful use. INSIDE estimates hallucination risk through the diversity of internal representations across sampled responses (Chen et al., 2024). ICR Probe instead tracks contributions to the residual stream across layers to detect hallucinations (Zhang et al., 2025c). Such approaches motivate examining internal dynamics, but their signals and supervision targets must be distinguished from source retrieval itself. ALERT reads one prompt pass and predicts the forthcoming answer's failure without observing generated answer tokens or sampling multiple responses. Its source ranker and failure probe have separate targets: selecting relevant evidence does not by itself certify that the model will use it correctly.

Attention-based monitoring. Lookback Lens uses per-head context-attention ratios during generation for hallucination detection and guided decoding (Chuang et al., 2024). AggTruth combines head selection with passage-attention statistics (Matys et al., 2025), whereas RAUQ recurrently combines selected-head attention with token probabilities without supervision (Vazhentsev et al., 2025). Other approaches use per-head divergence from uniform attention (van Dijk, 2026) or identify attention sinks as a source of detection signals (Binkowski et al., 2026). Together, these studies establish attention as a useful monitoring signal. Our focus is its allocation to candidate evidence before generation, rather than statistics computed over generated answer tokens.

Retrieval and context interventions. QRHead selects query-relevant heads using annotated passages and uses them for training-free reranking (Wu et al., 2025; Zhang et al., 2025b). It demonstrates transfer across evaluated datasets but does not predict subsequent answer failure. FACT iteratively extracts and removes candidate facts through repeated full-context calls (Wang et al., 2025), while Recursive Language Models decompose prompts through recursive calls over snippets (Zhang et al., 2025a). These methods improve access to evidence through retrieval or additional model calls. Our intervention instead repeats selected evidence when a failure warning is triggered, retaining the original conversation.

Why ALERT? **ALERT** reads attention at the final prompt token, requiring neither generated text nor sampled answers. It averages heads within globally attending layers and tracks attention to candidate source turns, the question, and the remaining history across depth. Sink tokens are excluded to focus these measurements on contextual attention. Source annotations supervise the ranker, and answer outcomes supervise the failure probe; neither is required at inference. The resulting warning controls selective re-injection of evidence immediately before the question, with original positions identified to preserve event order. Thus, the same attention signal supports source selection, pre-generation failure prediction, and targeted intervention without head selection or auxiliary generation calls for candidate selection. An extended discussion of related work and detailed comparisons are provided in §B.

3 WHY RETRIEVAL NEEDS A FAILURE MONITOR

An assistant answering a question about a long conversation must recover the relevant earlier statements amid unrelated history. For example, "Where is Mary?" requires her latest reported location, even when subsequent messages concern other topics. Document QA and indirect-cue retrieval pose related evidence-selection problems (Liu et al., 2024; Kuratov et al., 2024; Modarressi et al., 2025). This motivates our question: can signals collected while processing the prompt warn that the model will fail to use evidence already present? We first examine how adding unrelated context affects the model's answer while the relevant facts remain available. This provides a starting point for exploring how the model attends to those facts and how its internal representations evolve across layers, motivating the signals a failure monitor could use.

Measure distractibility while retaining the task facts. Our conversational BABILong adaptation pairs each long history with a short-story control containing the same task facts and question. We call the resulting performance loss *contextual distractibility*, following the operational logic of measuring interference from irrelevant input (Forster & Lavie, 2014). Figure 3 shows that retaining the evidence does not retain answer accuracy. The paired control distinguishes this loss from errors already present without distractors. Distance and distractor amount grow together, so the contrast measures their combined cost. It establishes the behavior that needs monitoring; we next trace a retained fact through prompt processing to examine the model's internal dynamics.

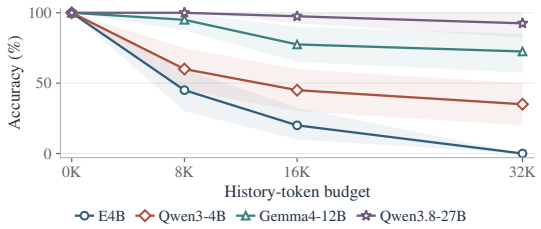


Figure 3: Retaining evidence does not retain accuracy. Generated-answer accuracy on the same 40 exploratory stories; four illustrative models. 0K: no book distractors. Bands: pointwise 95% paired-story bootstrap intervals.

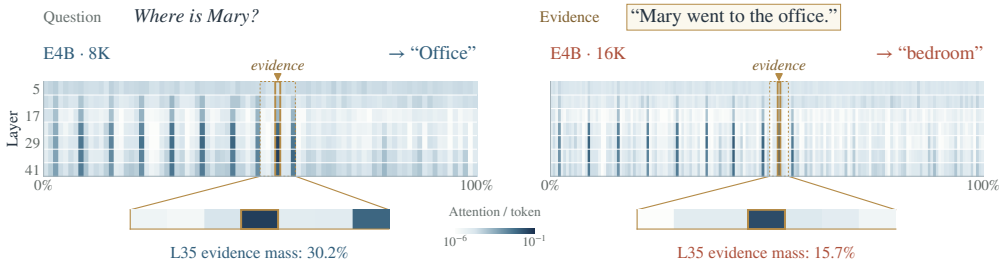


Figure 4: The source receives attention while the answer becomes wrong. Arrows name the later generated answers; "office" is correct. Gemma-4-E4B: head-mean attention per message token at seven global layers; final prompt position, shared log scale. Gold marks the evidence message and layer-35 zoom, not token attribution. Same story at 8,192 / 16,384 history tokens; retrospective example, with layer 35 chosen descriptively.

Attending to the evidence does not guarantee a correct answer. In Figure 4, the model assigns 15.7% of its head-averaged attention at layer 35 to the supporting message at 16K, yet answers "bedroom" when the correct location is "office". The error therefore coexists with attention to the relevant evidence. Attention weights describe how value vectors are mixed, $z = \sum_t a_t v_t$; answering also depends on the information those vectors carry and how subsequent computations transform it. Retrieval-head ablations show that routing can matter (Wu et al., 2025), while attention weights alone do not explain the output (Jain & Wallace, 2019). Although this example does not identify the failed computation, it shows why a monitor must evaluate more than whether attention reaches the correct source.

> Finding evidence and answering correctly are separate targets. A source ranker identifies a relevant earlier message. A failure monitor must distinguish correct from incorrect answers even when that message receives attention. Its predictions must therefore be validated against the later answer, including errors that occur despite attention to the correct source.

This distinction leads to the next question: does the representation approaching the output reveal what source attention leaves unresolved? We therefore follow the same story through every decoder layer before examining the wider cohort.

Inspecting the model before it answers. *Our signals are measured before the first answer token is generated.* We read each layer's hidden state at the final prompt token. Figure 5 projects it through the frozen output normalization and vocabulary head to score possible next tokens. The sketch marks the still-empty answer slot, y_1. These untuned, uncalibrated readouts (Belrose et al., 2023) follow layers within one prompt pass, without using any generated answer.



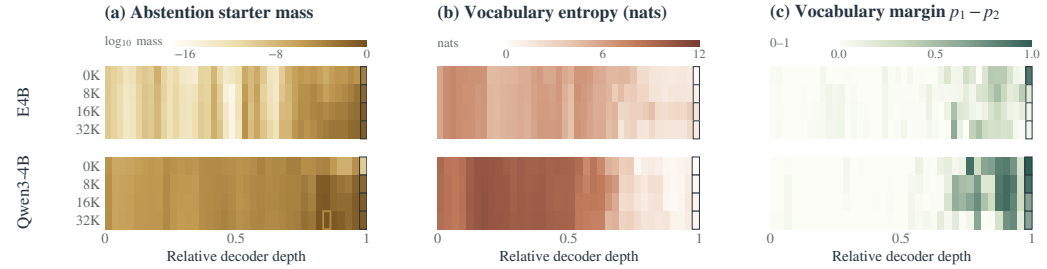


Figure 5: Next-token readouts across layers and context lengths. Same question; all 42 E4B / 36 Qwen layers, read at the final prompt token. Starter mass: first tokens of "Not stated" / "not stated"; entropy and margin: full vocabulary. Dark outlines mark final layers; gold marks the transient starter-mass peak.

A response tendency does not determine answer correctness. The 32K traces in Figure 5 end with similar abstention-starter mass, but markedly different margins and opposite answer outcomes. Their paths across layers also differ: a similar endpoint can conceal different trajectories. Figure 6 checks the response-type ambiguity across all 40 stories: abstention curves can separate while correct and wrong-location curves remain close. These group means motivate testing signals against later answer correctness; they do not establish individual-answer discrimination.

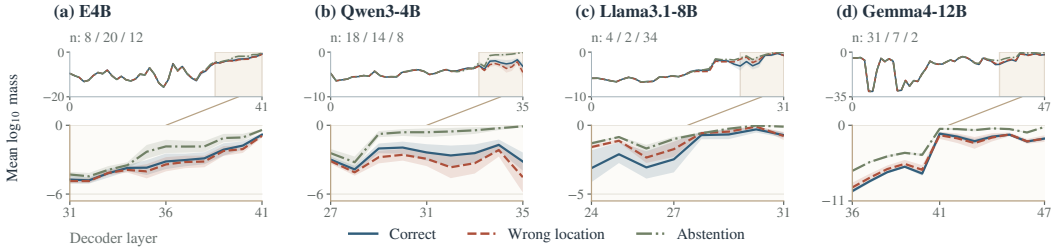


Figure 6: Layer trajectories grouped by answer outcome. All 40 stories at 16K; mean $\log_{10}$ starter mass. Full trajectories above; final quarter below. Bands: pointwise 95% story bootstrap (2,000 within-group resamples). Counts follow legend order; vertical ranges differ.

Recognizing abstention does not establish detection of wrong facts. Correct answers, incorrect locations and abstentions must remain distinguishable in evaluation; otherwise a change in their mixture can look like a more general improvement in failure detection.

Routing supplies an evidence-specific measurement. Vocabulary confidence summarizes competition between output tokens; routing describes attention to a particular source. Equal mean source mass can accompany attention spread across layers or concentrated in a few. Layer entropy and top-layer shares capture this distinction; source-to-history ratios measure competition with other messages. We test whether these patterns add predictive information beyond confidence after automatic source selection. Supervision refers to the eventual answer; logit-lens readouts are diagnostic and do not enter the current routing detector.

Hypothesis. At fixed context length, routing to automatically selected support adds information about later answer failure beyond next-token confidence. Test on unseen stories with frozen selection and warning thresholds, reporting wrong factual answers separately from abstentions.

Testing the warning by revisiting its source. The selected source links prediction to intervention: re-injection repeats its excerpt while preserving the question, history and chronology. Paired rescues and harms test whether renewed exposure helps. Section 4 formalizes this sequence: select a source, estimate failure risk, and re-inject only on warning.

4 SOURCE-RANKED DETECTION AND SELECTIVE RE-INJECTION

Section 3 motivates two components around a frozen LLM. Before generation, *(i) the detector* predicts whether the forthcoming answer will be correct or incorrect (Steps 1–3; Figure 7). *(ii) Selective re-injection* repeats automatically selected evidence on predicted failure (Step 4; Figure 8).

We begin with detection. Selective re-injection requires deciding which source to repeat and whether to intervene. Steps 1–3 provide the selected source and failure warning that guide re-injection in Step 4.

Step 1: Observe prompt attention. Prompt $x = (H, q)$ combines history messages $H = (c_1, \ldots, c_M)$ with question q. Each message is a candidate source. At the final prompt position, we record attention to its tokens across all heads of global-attention layers.

Step 2: Rank candidate sources. For each message, ϕ_j summarizes attention strength, distribution across layers and heads, and competition with the question and other messages. Appendix E defines all 14 standardized features. A linear probe selects one source: $\hat{j} = \arg\max_j (u^\top \phi_j + b_u)$, with learned weights u, intercept b_u and public source ID breaking ties.

Step 3: Score failure risk. The selected source may still be used incorrectly. A second linear probe predicts failure from its attention strength, layer distribution and competition with context: nine standardized routing features $\psi_{\hat{j}}$ (Appendix E). Weights v, intercept b_v and sigmoid σ give the uncalibrated risk

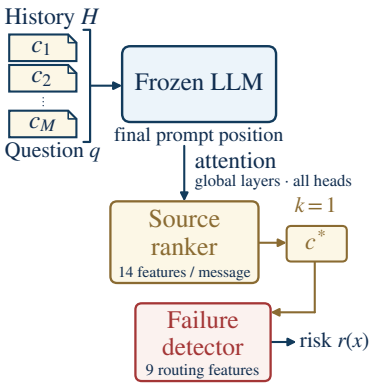


Figure 7: Steps 1–3: Detect before generation. Predict answer correctness before decoding from attention to one ranked message ($c^* = c_{\hat{j}}$). Gold support supervises fitting only.

$$r(x) = \sigma(v^\top \psi_{\hat{j}} + b_v), \qquad w(x) = \mathbf{1}\{r(x) \geq \tau\}. \tag{1}$$

(i) Pre-generation detector. Predict whether the forthcoming answer will be correct or incorrect. Warning $w(x) = 1$ flags predicted failure; abstention and truncation count as failures.

Offline fitting and freezing. Fit L_2 probes (inverse regularization strength $C = 1$) per LLM/dataset on all fitting items. Each prompt splits ranker weight equally between its annotated source and other candidates; balance failure classes. To include selection errors, train the detector on out-of-fold sources from three story-grouped BABILong folds or held-out NoLiMa fitting families; then refit the ranker on all fitting data. Disjoint calibration sets threshold τ to the highest value attaining $\geq 95\%$ empirical failure recall, keeping ties together. Freeze scalers, probes and τ; one-class fitting or calibration leaves the policy undefined.

(ii) Selective re-injection. On predicted failure, repeat the selected evidence before the question; otherwise, keep the original prompt. Identify the copy as an earlier message.

Step 4: Re-inject on warning. When $w(x) = 1$, we quote c^* before q in the final user turn, marking its original position as an earlier event. Otherwise, we decode from x unchanged.

We copy at most 64 native tokens verbatim, labeling clipped prefixes as excerpts. The wrapper adds at most 128 prompt tokens; 64 answer tokens remain reserved within the context limit. Re-injection follows prefill, saves no prefill computation and leaves model weights frozen.

Paired policy evaluation. For each evaluation item, we collect all interventions and replay frozen warnings offline, retaining initially correct items. Always-on and confidence-triggered re-injection compare trigger policies; lexical, length-matched random and privileged gold-source copies share flags to isolate source choice.

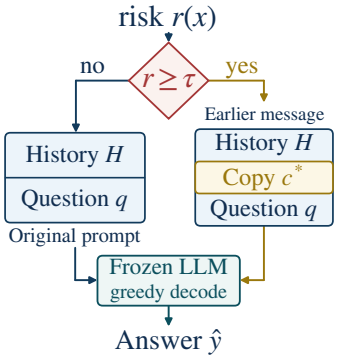


Figure 8: Step 4: Re-inject on warning. Copy the selected excerpt before the unchanged question when $r \geq \tau$;

Table 1: Failure detection on targeted retrieval tasks. Each context and Overall cell reports AUROC ↑/FPR ↓. R@1: source recall@1 pooled over contexts ↑. FPR uses calibrated thresholds.

	NoLiMa					RULER				
	AUROC/FPR				R@1	AUROC/FPR				R@1
Model	8K	16K	32K	Overall	Overall	8K	16K	32K	Overall	Overall
Gemma-4-E4B	0.943/0.081	0.984/0.192	1.000/0.375	0.970/0.177	0.986	0.943/0.256	0.967/0.262	0.950/0.252	0.955/0.257	0.890
Gemma-3-4B	0.980/0.030	0.967/0.069	0.891/0.350	0.945/0.122	0.924	0.942/0.194	0.947/0.287	0.976/0.453	0.954/0.302	0.583
Llama-3.1-8B	0.808/0.500	0.904/0.464	0.873/0.464	0.873/0.479	0.993	0.715/0.699	0.741/0.661	0.778/0.540	0.763/0.635	0.785
Qwen3-4B	0.857/0.688	0.899/0.467	0.895/0.714	0.880/0.605	0.882	0.780/0.384	0.898/0.370	0.801/0.366	0.824/0.373	0.791
Gemma-4-12B	0.985/0.044	0.952/0.088	0.873/0.200	0.939/0.096	0.993	0.918/0.232	0.943/0.178	0.957/0.187	0.941/0.200	0.777
Ministral-8B	–/0.021	0.936/0.064	0.913/0.125	0.958/0.067	0.986	0.885/0.225	0.874/0.213	–	0.881/0.219	0.654
Phi-4	0.948/0.244	0.910/0.265	–	0.931/0.253	0.990	0.604/0.723	–	–	0.604/0.723	0.558
Qwen3-14B	–/NP	–/NP	0.522/NP	0.704/NP	1.000	0.472/0.800	0.778/1.000	0.885/0.857	0.753/0.905	0.833
Qwen3.8-27B	0.532/NP	0.957/NP	0.874/NP	0.845/NP	1.000	0.676/0.832	0.673/0.844	0.756/0.740	0.706/0.804	0.581
Llama-3.2-11B	0.847/0.529	0.871/0.480	0.904/0.577	0.876/0.529	0.993	0.686/0.789	0.670/0.734	0.769/0.651	0.730/0.726	0.808

5 Benchmarks and Evaluation Setup

We evaluate pre-generation failure detection on four benchmarks adapted to multi-turn conversations at **8K, 16K, and 32K** tokens: NoLiMa (Modarressi et al., 2025) for associative retrieval, RULER (Hsieh et al., 2024) for needle-in-a-haystack retrieval, BABILong (Kuratov et al., 2024) for location tracking, and our proposed Scatter benchmark for scattered multiple-choice recall. We organise synthetic contexts into user turns and preserve natural exchange boundaries in conversational contexts, placing relevant evidence before each query amid unrelated content. NoLiMa embeds authored facts in MMLU exchanges, RULER places needles within synthetic filler, and BABILong interleaves movement facts with book passages. Scatter embeds three MMLU questions (Hendrycks et al., 2020) at approximately 25%, 75%, and 100% of WildChat conversations (Zhao et al., 2024), with each question's four options distributed across preceding user turns. The model answers each question before receiving further context.

We evaluate source localization and failure prediction separately, using benchmark-specific partitions that hold out facts, relation families, background contexts, or question identities. Matched short-context controls assess underlying task competence for BABILong and Scatter. We measure failure detection using AUROC and a calibrated operating point: the highest alert threshold achieving at least 95% failure recall on calibration predictions is fixed for evaluation, where we report the achieved true-positive and false-positive rates. Source recall@1 measures whether the top-ranked candidate turn matches the annotated evidence source. Full construction procedures, partitioning and calibration protocols, sample sizes, and controls are provided in §F.

6 Results

We first examine whether ALERT can locate relevant evidence and warn before an incorrect answer. We measure localization through source recall@1 (R@1), the fraction of examples in which the top-ranked history turn matches the annotated evidence source. We assess warning quality separately: AUROC measures how well scores distinguish failures from correct answers, while FPR measures how often the calibrated rule flags correct answers. We begin with NoLiMa and RULER, then turn to BABILong and Scatter-MCQ. These comparisons motivate our analysis of the detector's layer and feature use, before testing whether acting on its warnings improves answers through re-injection. More results, including bootstrap analyses and supplementary findings, can be found in D and C, respectively.

6.1 Targeted retrieval: NoLiMa and RULER

The results support the presence of a useful failure signal before answer generation (Table 1). Across NoLiMa and RULER, the three Gemma models achieve pooled AUROCs of 0.939–0.970, showing that attention to the selected source can distinguish future failures from successes before the first answer token. This signal extends beyond whether the ranker finds the annotated evidence: Gemma-3-4B reaches RULER AUROC of 0.954 despite source recall@1 of 0.583. Source-associated attention therefore provides information about answer reliability even when localization is imperfect. The signal

Table 2: Failure detection on state tracking and distributed recall. Each context and Overall cell reports AUROC ↑/FPR ↓. R@1: source recall@1 pooled over contexts ↑. FPR uses calibrated thresholds.

	BABILong					Scatter-MCQ				
	AUROC/FPR				R@1	AUROC/FPR				R@1
Model	8K	16K	32K	Overall	Overall	8K	16K	32K	Overall	Overall
Gemma-4-E4B	0.777/0.582	0.709/0.836	0.727/0.880	0.784/0.702	0.448	0.664/0.852	0.644/0.966	0.657/0.991	0.675/0.923	0.321
Gemma-3-4B	0.589/1.000	0.827/0.667	0.777/0.750	0.741/0.786	0.378	0.616/0.873	0.586/0.967	0.579/0.984	0.602/0.936	0.229
Llama-3.1-8B	0.612/0.540	0.718/0.900	–/–	0.872/0.573	0.343	0.677/0.921	0.587/0.980	0.612/0.986	0.651/0.955	0.251
Qwen3-4B	0.786/0.472	0.805/0.440	0.831/0.544	0.811/0.477	0.628	0.676/0.901	0.620/0.984	0.635/0.992	0.650/0.953	0.314
Gemma-4-12B	0.647/0.832	0.688/0.938	0.656/1.000	0.741/0.914	0.832	0.680/0.875	0.669/0.934	0.669/0.972	0.692/0.918	0.352
Ministral-8B	–/NP	–/NP	–	–/NP	0.107	0.634/0.936	0.603/0.968	–	0.624/0.950	0.256
Phi-4	0.680/0.667	0.622/0.989	–	0.710/0.791	0.407	0.726/0.868	–	–	0.726/0.868	0.332
Qwen3-14B	0.712/0.263	0.646/0.511	0.717/0.841	0.791/0.453	0.692	0.685/1.000	0.466/1.000	0.455/1.000	0.532/1.000	0.300
Qwen3.8-27B	0.947/NP	1.000/NP	0.894/NP	0.938/NP	0.983	0.512/0.958	0.557/0.967	0.404/0.931	0.501/0.952	0.133
Llama-3.2-11B	0.696/0.655	–/–	–/–	0.904/0.655	0.318	0.644/0.885	0.550/0.980	0.607/1.000	0.619/0.945	0.251

also remains informative at longer contexts; Gemma-3-4B's RULER AUROC rises from 0.942 at 8K to 0.976 at 32K. Its accompanying FPR increase shows that exploiting this predictive information requires careful threshold calibration, but the discrimination results establish an opportunity to detect likely failures before committing to an answer.

6.2 State tracking and distributed recall: BABILong and Scatter-MCQ

Pre-generation attention remains informative when answering requires tracking changing facts or retrieving options distributed across turns (Table 2). On BABILong, Qwen3-4B achieves AUROCs of 0.786, 0.805, and 0.831 at 8K, 16K, and 32K, respectively, indicating that the failure signal persists as the history grows. Its source recall@1 of 0.628 further suggests that useful warnings do not require perfect evidence localization. Scatter-MCQ presents a harder setting, yet Gemma-4-E4B retains similar discrimination at 8K and 32K (0.664 and 0.657), consistent with a weaker but persistent signal under distributed recall. The high false-positive rates limit how selectively these scores can currently trigger action; they do not erase the predictive information available before generation. Together, these findings motivate using attention as an early warning cue and testing whether acting on that cue through re-injection can convert predicted risk into improved answers.

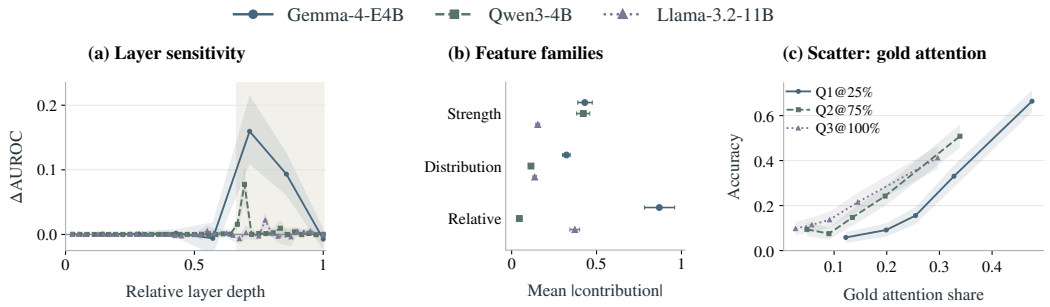


Figure 9: Attention patterns behind the warning. (a) BABILong layer-permutation sensitivity. **(b)** Feature-family contributions (mean absolute contribution per feature). **(c)** Scatter-MCQ accuracy by annotated gold-attention quintile. Details in Appendix E.1.

6.3 Attention patterns behind the warning

The detector depends most on late, pre-final attention. Sensitivity peaks at layers 30/42, 25/36, and 31/40 for Gemma, Qwen, and Llama, respectively (Figure 9a). Jointly shuffling the final depth third reduces AUROC by 0.289, 0.245, and 0.125, whereas shuffling the first third changes it by less than 0.002 in magnitude. Thus, the strongest measured dependence lies in late, pre-final routing, which is available before any answer is generated. This locates dependence within the attention-based detector; it does not establish whether earlier hidden states already encode the correct answer, or identify layers that causally determine answering success.

Models use different combinations of attention features. Relative-attention features have the largest mean contribution per feature for Gemma and Llama, while strength features dominate Qwen

(Figure 9b). The warning therefore depends on how the selected source competes with the question and other history, as well as the attention it receives across layers. Appendix Figure 10 shows all nine features separately. These magnitudes describe the fitted score, not unique predictive information or a universal relationship between attention and successful evidence use.

Scatter links attention to evidence use when location is supplied. In the complementary Scatter diagnostic, answer accuracy increases across gold-attention-share quintiles at each conversation probe (Figure 9c). This relationship persists even as attention spreads over more options at later probes: the gold option remains individually favoured, but receives a smaller share of total option attention. Its correctness AUROC declines from 0.82 to 0.79 to 0.74. These annotated-option associations motivate attention as a reliability cue; they do not establish the performance of automatic source selection or a pre-generation warning. The following re-injection results test whether acting on the warning improves answers.

6.4 Frozen-method confirmation on new stories

Table 3: Selective re-injection by context length: answer accuracy (%).

		BABILong				RULER				Scatter-MCQ			
Model	Ctx.	Base	Re-inj.	Δ (pp)	Int.	Base	Re-inj.	Δ (pp)	Int.	Base	Re-inj.	Δ (pp)	Int.
Gemma-4-E4B	8K	49.0	60.5	+11.5	77.0	75.5	75.5	+0.0	46.0	28.5	37.5	+9.0	95.0
	16K	27.5	47.0	+19.5	92.5	71.5	71.5	+0.0	47.0	20.5	28.5	+8.0	99.5
	32K	12.5	26.0	+13.5	96.0	75.5	76.0	+0.5	40.0	15.5	27.0	+11.5	100.0
	All	29.7	44.5	+14.8	88.5	74.2	74.3	+0.2	44.3	21.5	31.0	+9.5	98.2
Qwen3-4B	8K	54.0	67.0	+13.0	66.0	75.5	75.5	+0.0	57.0	26.5	31.0	+4.5	95.5
	16K	45.5	64.5	+19.0	67.5	71.5	72.0	+0.5	56.0	19.6	24.1	+4.5	98.0
	32K	28.5	43.0	+14.5	80.5	77.5	77.5	+0.0	49.5	18.6	26.6	+8.0	99.0
	All	42.7	58.2	+15.5	71.3	74.8	75.0	+0.2	54.2	21.6	27.3	+5.7	97.5
Llama-3.2-11B	8K	27.5	54.0	+26.5	75.5	93.0	97.0	+4.0	81.5	19.0	28.0	+9.0	94.5
	16K	0.0	25.5	+25.5	99.0	93.5	97.5	+4.0	73.0	14.6	18.6	+4.0	100.0
	32K	0.0	3.5	+3.5	100.0	88.5	95.0	+6.5	72.0	13.6	17.6	+4.0	100.0
	All	9.2	27.7	+18.5	91.5	91.7	96.5	+4.8	75.5	15.7	21.4	+5.7	98.2

Table 3 shows that intervention rates vary with benchmark difficulty. BABILong and Scatter-MCQ exhibit weaker baseline accuracy and higher firing rates, with re-injection improving aggregate accuracy by 14.8–18.5 and 5.7–9.5 percentage points, respectively. RULER is comparatively easier for the evaluated models, with stronger baseline accuracy (74.2–91.7%), lower intervention rates within each model, and smaller gains (0.2–4.8 percentage points).

Table 4: Top-K ablation on RULER. AUROC ↑; FPR at the 95%-sensitivity threshold ↓; source R@K ↑.

		AUROC			Overall	
Model	K	8K	16K	32K	AUROC/FPR	R@K
Gemma-4-E4B	1	0.935	0.969	0.957	0.956/0.269	0.885
	3	0.780	0.870	0.928	0.866/0.953	0.982
	5	0.753	0.816	0.867	0.812/0.991	0.993
Qwen3-4B	1	0.744	0.890	0.801	0.812/0.405	0.760
	3	0.742	0.661	0.715	0.702/0.851	0.887
	5	0.703	0.626	0.650	0.650/0.980	0.907
Llama-3.2-11B	1	0.633	0.680	0.791	0.722/0.740	0.813
	3	0.589	0.634	0.605	0.607/0.998	0.907
	5	0.522	0.667	0.591	0.588/1.000	0.933

Overall pools contexts. $K > 1$: frozen head, max-pooled scores (naive control).

We concatenate the top-K candidate turns into a single input to the detector. Table 7 shows that $K = 1$ achieves the highest detection AUROC across three models and context lengths. Increasing K improves source recall but reduces overall AUROC for all three models. False-positive rates also rise sharply, reaching 0.980–1.000 at $K = 5$. Thus, concatenating additional candidates improves evidence coverage but weakens failure detection, supporting our choice of $K = 1$. The detector also introduces negligible inference latency (§G).

7 Conclusion

We introduce ALERT, which ranks source turns in multi-turn conversations using attention and predicts answer failure from the selected evidence's attention profile before any answer token is generated. Paired with selective evidence re-injection immediately before the question, ALERT improves accuracy and recovers otherwise incorrect answers. We also introduce Scatter-MCQ, a benchmark that distributes answer options across authentic long conversations to evaluate retrieval and failure detection when relevant evidence is dispersed across turns.

AI USE STATEMENT

In this work, we used generative AI tools to draft portions of the manuscript and assist in producing result visualisations. We have not used generative AI tools to develop research ideas, design experiments, implement the proposed method, analyse experimental results, or formulate conclusions. Additionally, we used generative AI tools to improve readability through grammar, spelling, and word-choice edits. We have reviewed all AI-assisted work, and manuscript drafts produced with AI assistance were subsequently revised by the authors. All research ideas, experimental design, implementations, experiments, analyses, and conclusions are the authors' own. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Jakub Binkowski, Kamil Adamczewski, and Tomasz Kajdanowicz. Attention sinks as internal signals for hallucination detection in large language models. *arXiv preprint arXiv:2604.10697*, 2026.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms' internal states retain the power of hallucination detection. In *International Conference on Learning Representations*, volume 2024, pp. 3056–3076, 2024.

Jingxuan Chen, Mohammad Taher Pilehvar, and Jose Camacho-Collados. Understanding llm performance degradation in multi-instance processing: The roles of instance count and context length. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31851–31884, 2026.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James Glass. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1419–1436, 2024.

Hamed Damirchi, Ehsan Abbasnejad, Afshar Shamsi, Zhen Zhang, Javen Qinfeng Shi, et al. Truth as a trajectory: What internal representations reveal about large language model reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 44774–44790, 2026.

Yufeng Du, Minyang Tian, Srikanth Ronanki, Subendhu Rongali, Sravan Bodapati, Aram Galstyan, Azton Wells, Roy Schwartz, Eliu A Huerta, and Hao Peng. Context length alone hurts llm performance despite perfect retrieval. *arXiv preprint arXiv:2510.05381*, 2025.

Sophie Forster and Nilli Lavie. Distracted by your mind? individual differences in distractibility predict mind wandering. *Journal of experimental psychology: learning, memory, and cognition*, 40(1):251, 2014.

Muhan Gao, TaiMing Lu, Kuai Yu, Adam Byerly, and Daniel Khashabi. Insights into llm long-context failures: when transformers know but don't tell. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7611–7625, 2024.

Robert Hankache, Kingsley Nketia Acheampong, Liang Song, Marek Brynda, Raad Khraishi, and Greig A Cowan. Evaluating the sensitivity of llms to prior context. *arXiv preprint arXiv:2506.00069*, 2025.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

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? *arXiv preprint arXiv:2404.06654*, 2024.

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, Volume 1 (Long and Short Papers)*, pp. 3543–3556, 2019.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. *Advances in Neural Information Processing Systems*, 37:106519–106554, 2024.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation. In *International Conference on Learning Representations*, volume 2026, pp. 54738–54778, 2026.

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.

Piotr Matys, Jan Eliasz, Konrad Kiełczyński, Mikołaj Langner, Teddy Ferdinan, Jan Kocoń, and Przemysław Kazienko. Aggtruth: contextual hallucination detection using aggregated attention scores in llms. In *International Conference on Computational Science*, pp. 227–243. Springer, 2025.

Aman Mehta. When agents commit too soon: Diagnosing premature commitment in llm agents. *arXiv preprint arXiv:2606.22936*, 2026.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A Rossi, Seunghyun Yoon, and Hinrich Schütze. Nolima: Long-context evaluation beyond literal matching. *arXiv preprint arXiv:2502.05167*, 2025.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 1797–1807, 2018.

Gijs van Dijk. Detecting hallucinations in large language models via internal attention divergence signals. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026)*, pp. 60–75, 2026.

Artem Vazhentsev, Lyudmila Rvanova, Gleb Kuzmin, Ekaterina Fadeeva, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Timothy Baldwin, Mrinmaya Sachan, Preslav Nakov, et al. Uncertainty-aware attention heads: Efficient unsupervised uncertainty quantification for llms. *arXiv e-prints*, pp. arXiv–2505, 2025.

Jinlin Wang, Suyuchen Wang, Ziwen Xia, Sirui Hong, Yun Zhu, Bang Liu, and Chenglin Wu. Fact: examining the effectiveness of iterative context rewriting for multi-fact retrieval. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 3382–3392, 2025.

Weiwei Wang, Jiyong Min, and Weijie Zou. Intelligence degradation in long-context llms: critical threshold determination via natural length distribution analysis. *arXiv preprint arXiv:2601.15300*, 2026.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. In *International Conference on Learning Representations*, volume 2025, pp. 62143–62156, 2025.

Alex L Zhang, Tim Kraska, and Omar Khattab. Recursive language models. *arXiv preprint arXiv:2512.24601*, 2025a.

Wuwei Zhang, Fangcong Yin, Howard Yen, Danqi Chen, and Xi Ye. Query-focused retrieval heads improve long-context reasoning and re-ranking. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 23802–23816, 2025b.

Zhenliang Zhang, Xinyu Hu, Huixuan Zhang, Junzhe Zhang, and Xiaojun Wan. Icr probe: Tracking hidden state dynamics for reliable hallucination detection in llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17986–18002, 2025c.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatgpt interaction logs in the wild. In *International Conference on Learning Representations*, volume 2024, pp. 34590–34605, 2024.

A Appendix

Limitations. ALERT's classifier is fitted per model and per benchmark, so a new setting requires a labelled calibration set. It also reads attention directly, so it applies only to open-weight (white-box) models.

B Extended Related Work

This section expands the comparisons in §2, retaining quantitative findings, implementation distinctions, and qualifications omitted from the main text.

Degradation in long and multi-turn contexts. Laban et al. (2026) evaluate over 200K simulated multi-turn conversations with under-specified requests across 15 models, reporting an average performance drop of 39% relative to single-turn prompts. They attribute the decline primarily to unreliability: models commit prematurely, rely on earlier incorrect attempts, and rarely recover. Hankache et al. (2025) prepend 4K–64K tokens of dialogue history to GPQA questions and report relative performance drops of up to 73%. Cross-domain histories are more disruptive than same-domain histories, and larger models also degrade. Repeating the task description before the query recovers most of the loss, suggesting dilution of task-relevant information rather than a loss of capacity.

Du et al. (2025) report accuracy reductions of 13.9–85 percentage points in essay-filler experiments despite successful verbatim evidence retrieval. Separate experiments show that degradation persists with whitespace or fully masked filler, supporting an effect of length beyond retrieval failure or semantic distraction. NoLiMa removes lexical overlap between queries and needles; eleven of thirteen models with 128K-token context windows fall below half their short-context accuracy at 32K tokens (Modarressi et al., 2025). Using natural-length inputs, Wang et al. (2026) identify a performance cliff for Qwen2.5-7B at approximately 40–50% of its context window.

Chen et al. (2026) examine processing and aggregating multiple instances of an individually tractable task in one prompt. Performance declines noticeably above 200 instances and approaches collapse beyond 1,000, with average success below 20% at 2,000 instances. Success correlates more strongly with instance count (Spearman -0.61) than with total context length (-0.37). Wang et al. (2025) additionally use layer-wise probes to show that models progressively lose relevant facts during generation.

These studies characterise conditions and mechanisms of failure and motivate mitigations such as conversation recaps, task repetition, and evidence recitation. Our objective is a per-prompt warning that anticipates answer failure and guides an input-side intervention. Our conversational evaluation draws on these retrieval and prior-context settings through adaptations of NoLiMa, RULER (Hsieh et al., 2024), and BABILong (Kuratov et al., 2024), together with our Scatter-MCQ benchmark.

Hidden-state signals of failure. INSIDE detects hallucinations through the dispersion of mid-layer sentence embeddings across K sampled generations (Chen et al., 2024). Gao et al. (2024) show that linear probes recover the location of the gold document from hidden states even when the generated answer is incorrect. Relevant information can therefore remain internally accessible despite output failure, with targets in the middle of the context localised only in deeper layers. ICR Probe analyses module contributions to the residual stream (Zhang et al., 2025c), whereas Truth-as-a-Trajectory uses layer-wise displacement vectors (Damirchi et al., 2026). Both model evolution across depth and report improved generalisation over static probes. Mehta (2026) define representational commitment through cross-run hidden-state similarity at a fixed agent step. This predicts trajectory consistency rather than correctness and, like INSIDE, requires multiple runs.

Attention-based monitoring. Lookback Lens computes, for each head and generated token, the ratio of attention to the input context to attention over the combined context and generated tokens (Chuang et al., 2024). A linear classifier on these ratios matches hidden-state and entailment-based detectors, transfers from a 7B to a 13B model without retraining, and supports guided decoding that reduces hallucination on XSum (Narayan et al., 2018) by 9.6%. AggTruth combines head selection with generated-token attention statistics over the passage, including total mass, entropy, inter-head

similarity, and Jensen–Shannon distance, reporting stronger cross-task transfer than hidden-state probes (Matys et al., 2025). RAUQ is unsupervised: it selects one uncertainty-aware head per layer and recurrently combines attention to previous tokens with token probabilities (Vazhentsev et al., 2025).

van Dijk (2026) use each head's KL divergence from uniform attention, finding signals concentrated in middle layers and on entity tokens. SinkProbe shows that attention-based detectors can implicitly depend on attention sinks, whose dominance indicates a shift from input-grounded to prior-driven computation (Binkowski et al., 2026).

ALERT differs in its observation point, feature construction, and treatment of sinks. It reads attention at the final prompt token, before any answer is generated, and therefore requires neither generated text nor sampled answers. Head-averaged attention to candidate source turns, the question, and the remaining history is tracked across globally attending layers, capturing allocation across depth without head selection or per-head classifiers. Motivated by SinkProbe, sink tokens are excluded so that region-level measurements focus on contextual attention rather than default routing. As in Lookback Lens and AggTruth, answer outcomes supervise the failure classifier. Source annotations additionally supervise the learned ranker, but neither source annotations nor answer outcomes are supplied at inference, when source turns are ranked using final-prompt-token attention features.

Retrieval heads, context selection, and mitigation. QRHead extends the copy-paste retrieval heads identified by Wu et al. (2025) by scoring heads according to attention from query tokens to annotated gold passages (Zhang et al., 2025b). The selected heads support a training-free reranker that outperforms dense retrievers on long-context reasoning tasks. Head identification uses labelled data, and transfer is demonstrated across the evaluated datasets; universal task independence is not established. The method retrieves evidence without predicting subsequent generation failure.

FACT repeatedly extracts candidate facts and removes them from the context to improve multi-fact retrieval (Wang et al., 2025). Each iteration requires an additional full-context call, extracted candidates remain unverified, and context rewriting degrades reasoning-intensive QA performance for Llama-3.1 and Qwen-2.5. Recursive Language Models avoid directly processing the entire prompt by decomposing it and recursively invoking the model on snippets, requiring multiple sub-calls (Zhang et al., 2025a).

ALERT instead uses a frozen warning rule to control selective re-injection. Short excerpts from the top-K turns selected using the same head-averaged attention signal are repeated immediately before the question, with their original positions identified to preserve event order. The model answers using the augmented full history. Candidate selection requires no source annotations at inference, head detection, iterative rewriting, or auxiliary generation calls; the original conversation remains intact.

C Additional Results

Table 5: Failure-detection TPR per context (at the calibrated threshold)

	RULER				Scatter-MCQ			
Model	8K	16K	32K	Overall	8K	16K	32K	Overall
Gemma-4-E4B	0.956	0.979	0.907	0.946	0.948	0.976	0.996	0.975
Gemma-3-4B	0.963	0.995	1.000	0.987	0.967	0.985	0.998	0.984
Llama-3.1-8B	0.857	0.929	0.961	0.930	0.963	0.994	0.996	0.985
Qwen3-4B	0.944	0.990	0.873	0.935	0.956	0.983	0.994	0.978
Gemma-4-12B	0.972	0.992	1.000	0.989	0.950	0.983	0.993	0.977
Ministral-8B	0.929	0.924	–	0.926	0.966	0.986	–	0.976
Phi-4	0.939	–	–	0.939	0.983	–	–	0.983
Qwen3-14B	0.976	0.953	0.938	0.956	0.990	0.964	0.913	0.956
Qwen3.8-27B	0.956	0.950	0.930	0.945	0.988	0.961	0.900	0.950
Llama-3.2-11B	0.909	0.885	0.980	0.939	0.959	0.994	0.998	0.984

–: no records at that context (model not run there) or metric undefined (single outcome class; TPR with no failures; FPR with no successes). Overall pools the 8K/16K/32K records for that cell.

Table 6: Source recall@1 per context (↑; automatic source selection)

	RULER				Scatter-MCQ			
Model	8K	16K	32K	Overall	8K	16K	32K	Overall
Gemma-4-E4B	0.933	0.905	0.834	0.890	0.422	0.298	0.242	0.321
Gemma-3-4B	0.615	0.601	0.534	0.583	0.307	0.206	0.175	0.229
Llama-3.1-8B	0.850	0.811	0.695	0.785	0.350	0.251	0.152	0.251
Qwen3-4B	0.801	0.818	0.753	0.791	0.415	0.265	0.263	0.314
Gemma-4-12B	0.824	0.765	0.743	0.777	0.463	0.345	0.247	0.352
Ministral-8B	0.749	0.560	–	0.654	0.311	0.201	–	0.256
Phi-4	0.558	–	–	0.558	0.332	–	–	0.332
Qwen3-14B	0.948	0.847	0.708	0.833	0.395	0.285	0.219	0.300
Qwen3.8-27B	0.571	0.573	0.599	0.581	0.193	0.131	0.075	0.133
Llama-3.2-11B	0.863	0.839	0.723	0.808	0.350	0.256	0.147	0.251

–: no records at that context (model not run there) or metric undefined (single outcome class; TPR with no failures; FPR with no successes). Overall pools the 8K/16K/32K records for that cell.

Table 7: Top-K ablation on RULER: failure detection and source recall. AUROC ↑ (chance = 0.5); FPR at the 95%-sensitivity threshold ↓; R@K: source recall at K ↑.

	AUROC			AUROC / FPR	R@K
Model	8K	16K	32K	Overall	Overall
$K = 1$					
Gemma-4-E4B	0.935	0.969	0.957	0.956/0.269	0.885
Qwen3-4B	0.744	0.890	0.801	0.812/0.405	0.760
Llama-3.2-11B	0.633	0.680	0.791	0.722/0.740	0.813
$K = 3$					
Gemma-4-E4B	0.780	0.870	0.928	0.866/0.953	0.982
Qwen3-4B	0.742	0.661	0.715	0.702/0.851	0.887
Llama-3.2-11B	0.589	0.634	0.605	0.607/0.998	0.907
$K = 5$					
Gemma-4-E4B	0.753	0.816	0.867	0.812/0.991	0.993
Qwen3-4B	0.703	0.626	0.650	0.650/0.980	0.907
Llama-3.2-11B	0.522	0.667	0.591	0.588/1.000	0.933

Overall pools the 8K/16K/32K records. FPR is reported pooled over contexts only. R@K: fraction of records whose evidence-bearing turn appears in the top K ranked turns. $K>1$ rows use the frozen head with max-pooled scores (naive control). Overall AUROC falls and R@K rises monotonically in K for every model. FPR = 1.000 means the threshold flags every record (degenerate).

D Bootstrap Confidence Intervals

We quantify evaluation-item uncertainty for the NoLiMa and BABILong results using 10,000 paired cluster bootstrap resamples (random seed 7). The resampling units are the 48 independent NoLiMa fact IDs and the 200 independent BABILong story IDs per model. Each resample draws the same number of IDs with replacement and retains all observed context lengths, detector scores, warning decisions, and paired intervention outcomes for each selected ID. Repeated contexts therefore do not increase the independent sample size.

We recompute each metric within every resample and report its empirical 2.5th and 97.5th percentiles. Overall metrics pool the admitted context records while preserving their grouping by ID. Re-injection gains are calculated from the paired difference between selective-policy and baseline correctness. Detector fits, calibration thresholds, and evaluation folds remain fixed throughout resampling.

Tables 15–13 report failure-detection AUROC, calibrated false-positive rate, and automatic source recall@1 for all ten models, wherever the metric is defined. Table 18 reports baseline accuracy, selective re-injection accuracy, paired gains, and intervention rates for the three models in the main re-injection table.

Table 8: Top-K ablation: failure-detection AUROC per context (↑; chance $= 0.5$)

	RULER				Scatter-MCQ			
Model	8K	16K	32K	Overall	8K	16K	32K	Overall
$K = 1$								
Gemma-4-E4B	0.935	0.969	0.957	0.956	0.645	0.605	0.596	0.643
Qwen3-4B	0.744	0.890	0.801	0.812	0.668	0.659	0.649	0.663
Llama-3.2-11B	0.633	0.680	0.791	0.722	0.634	0.541	0.600	0.602
$K = 3$								
Gemma-4-E4B	0.780	0.870	0.928	0.866	0.590	0.595	0.546	0.616
Qwen3-4B	0.742	0.661	0.715	0.702	0.623	0.577	0.635	0.620
Llama-3.2-11B	0.589	0.634	0.605	0.607	0.520	0.535	0.521	0.546
$K = 5$								
Gemma-4-E4B	0.753	0.816	0.867	0.812	0.541	0.518	0.515	0.557
Qwen3-4B	0.703	0.626	0.650	0.650	0.537	0.558	0.629	0.576
Llama-3.2-11B	0.522	0.667	0.591	0.588	0.483	0.532	0.477	0.525

Overall pools the 8K/16K/32K records for that cell. Overall AUROC falls monotonically in K for every model on both RULER and Scatter-MCQ. $K>1$ rows are the frozen head with max-pooled scores (naive control), evaluated on a 200/context subsample.

Table 9: Operating point: false-positive rate at the 95%-sensitivity threshold (↓), pooled over contexts

	RULER			Scatter-MCQ		
Model	K=1	K=3	K=5	K=1	K=3	K=5
Gemma-4-E4B	0.269	0.953	0.991	0.938	1.000	1.000
Qwen3-4B	0.405	0.851	0.980	0.961	1.000	1.000
Llama-3.2-11B	0.740	0.998	1.000	0.926	1.000	1.000

$K>1$ columns are the frozen head with max-pooled scores (naive control). FPR $= 1.000$ with TPR $= 1.000$ means the threshold flags every record (degenerate).

Table 10: Top-K ablation: source recall at K (↑)

	RULER				Scatter-MCQ			
Model	8K	16K	32K	Overall	8K	16K	32K	Overall
$K = 1$								
Gemma-4-E4B	0.930	0.900	0.825	0.885	0.335	0.190	0.190	0.238
Qwen3-4B	0.770	0.785	0.725	0.760	0.360	0.226	0.216	0.268
Llama-3.2-11B	0.865	0.815	0.760	0.813	0.280	0.186	0.070	0.179
$K = 3$								
Gemma-4-E4B	1.000	0.980	0.965	0.982	0.655	0.480	0.455	0.530
Qwen3-4B	0.945	0.895	0.820	0.887	0.690	0.523	0.487	0.567
Llama-3.2-11B	0.945	0.900	0.875	0.907	0.540	0.417	0.327	0.428
$K = 5$								
Gemma-4-E4B	1.000	0.990	0.990	0.993	0.805	0.650	0.595	0.683
Qwen3-4B	0.960	0.925	0.835	0.907	0.850	0.719	0.633	0.734
Llama-3.2-11B	0.960	0.935	0.905	0.933	0.770	0.613	0.533	0.639

Fraction of records whose evidence-bearing turn appears in the top K ranked turns. Non-decreasing in K in every cell, and non-increasing in context length in every cell except Qwen3-4B on RULER at K=1. $K>1$ rows are the frozen head with max-pooled scores (naive control), evaluated on the full RULER test split ($n \approx 2235$).

These intervals are pointwise, not simultaneous, and describe uncertainty conditional on the saved models, detector fits, folds, and background pools. They do not estimate variation across training seeds or repeated detector fitting. In particular, resampling fixed NoLiMa out-of-fold predictions does not capture dependence induced by shared training data. The BABILong 200-story amendment followed inspection of earlier outcomes and overlaps the original 576-story cohort; it is not an

independent replication. Undefined single-class or zero-denominator resamples are excluded and counted in the accompanying analysis. Rare classes can yield unstable intervals, and zero-width empirical intervals do not establish certainty.

Table 11: Failure-detection AUROC per context (↑)

	NoLiMa				BABILong			
Model	8K	16K	32K	Overall	8K	16K	32K	Overall
Gemma-4-E4B	0.943 [0.850, 1.000]	0.984 [0.951, 1.000]	1.000† [1.000, 1.000]	0.970 [0.946, 0.990]	0.777 [0.711, 0.839]	0.709 [0.628, 0.783]	0.727 [0.624, 0.816]	0.784 [0.741, 0.824]
Gemma-3-4B	0.980 [0.937, 1.000]	0.967 [0.911, 1.000]	0.891 [0.760, 0.989]	0.945 [0.899, 0.982]	0.589† [0.412, 0.813]	0.827† [0.720, 0.929]	0.777† [0.412, 0.966]	0.741 [0.632, 0.849]
Llama-3.1-8B	0.808 [0.630, 0.955]	0.904 [0.805, 0.976]	0.873 [0.764, 0.958]	0.873 [0.794, 0.938]	0.612 [0.534, 0.690]	0.718 [0.563, 0.858]	–	0.872 [0.844, 0.899]
Qwen3-4B	0.857 [0.719, 0.964]	0.899 [0.776, 0.986]	0.895† [0.784, 0.977]	0.880 [0.819, 0.932]	0.786 [0.721, 0.846]	0.805 [0.741, 0.864]	0.831 [0.768, 0.885]	0.811 [0.772, 0.848]
Gemma-4-12B	0.985† [0.936, 1.000]	0.952 [0.886, 0.993]	0.873 [0.762, 0.958]	0.939 [0.893, 0.974]	0.647† [0.422, 0.869]	0.688 [0.605, 0.768]	0.656 [0.575, 0.732]	0.741 [0.686, 0.793]
Ministral-8B	–	0.936† [0.851, 1.000]	0.912† [0.793, 0.993]	0.958† [0.904, 0.994]	–	–	–	–
Phi-4	0.948† [0.872, 0.997]	0.910 [0.810, 0.987]	–	0.931 [0.874, 0.977]	0.680 [0.599, 0.759]	0.622 [0.542, 0.698]	–	0.710 [0.659, 0.761]
Qwen3-14B	–	–	0.522† [0.064, 0.957]	0.704† [0.371, 0.986]	0.712 [0.627, 0.791]	0.646 [0.562, 0.726]	0.717 [0.639, 0.792]	0.791 [0.758, 0.824]
Qwen3.8-27B	0.532† [0.386, 0.674]	0.957† [0.891, 1.000]	0.874† [0.575, 1.000]	0.845† [0.638, 1.000]	0.947† [0.869, 1.000]	1.000† [1.000, 1.000]	0.894† [0.720, 0.993]	0.938† [0.833, 0.993]
Llama-3.2-11B	0.847 [0.707, 0.962]	0.871 [0.758, 0.962]	0.904 [0.811, 0.971]	0.876 [0.797, 0.941]	0.696 [0.616, 0.774]	–	–	0.904 [0.873, 0.931]

Each cell gives the estimate above its pointwise 95% paired cluster bootstrap interval (10,000 resamples). NoLiMa: 48 facts; BABILong: 200 stories. Overall pools contexts while retaining repeated observations within each ID. Intervals condition on saved fits, folds and backgrounds, not training seeds. †: rare supporting classes, undefined resamples, or a degenerate empirical interval; a zero-width interval does not establish certainty. –: context unavailable, missing detector score, or only one outcome class.

Table 12: False-positive rate per context (↓; calibrated thresholds)

	NoLiMa				BABILong			
Model	8K	16K	32K	Overall	8K	16K	32K	Overall
Gemma-4-E4B	0.081 [0.000, 0.182]	0.192 [0.045, 0.357]	0.375 [0.143, 0.625]	0.177 [0.100, 0.264]	0.582 [0.484, 0.680]	0.836 [0.736, 0.929]	0.880 [0.737, 1.000]	0.702 [0.629, 0.777]
Gemma-3-4B	0.030 [0.000, 0.100]	0.069 [0.000, 0.179]	0.350 [0.150, 0.571]	0.122 [0.051, 0.202]	1.000† [1.000, 1.000]	0.667† [0.200, 1.000]	0.750† [0.000, 1.000]	0.786 [0.533, 1.000]
Llama-3.1-8B	0.500 [0.342, 0.658]	0.464 [0.280, 0.654]	0.464 [0.276, 0.652]	0.479 [0.341, 0.619]	0.540 [0.441, 0.638]	0.900 [0.667, 1.000]	–	0.573 [0.480, 0.664]
Qwen3-4B	0.688 [0.444, 0.917]	0.467 [0.214, 0.727]	0.714† [0.333, 1.000]	0.605 [0.425, 0.769]	0.472 [0.379, 0.567]	0.440 [0.338, 0.543]	0.544 [0.414, 0.673]	0.477 [0.408, 0.547]
Gemma-4-12B	0.044 [0.000, 0.111]	0.088 [0.000, 0.194]	0.200 [0.045, 0.368]	0.096 [0.030, 0.176]	0.832 [0.777, 0.883]	0.938 [0.899, 0.974]	1.000† [1.000, 1.000]	0.914 [0.888, 0.938]
Ministral-8B	0.021 [0.000, 0.062]	0.064 [0.000, 0.146]	0.125 [0.026, 0.238]	0.067 [0.015, 0.131]	–	–	–	–
Phi-4	0.244 [0.119, 0.381]	0.265 [0.125, 0.419]	–	0.253 [0.141, 0.377]	0.667 [0.588, 0.743]	0.989 [0.963, 1.000]	–	0.791 [0.741, 0.840]
Qwen3-14B	–	–	–	–	0.263 [0.199, 0.329]	0.511 [0.429, 0.596]	0.841 [0.745, 0.926]	0.453 [0.401, 0.508]
Qwen3.8-27B	–	–	–	–	–	–	–	–
Llama-3.2-11B	0.529 [0.359, 0.694]	0.480 [0.280, 0.679]	0.577 [0.379, 0.767]	0.529 [0.378, 0.681]	0.655 [0.526, 0.780]	–	–	0.655 [0.526, 0.780]

Each cell gives the estimate above its pointwise 95% paired cluster bootstrap interval (10,000 resamples). NoLiMa: 48 facts; BABILong: 200 stories. Overall pools contexts while retaining repeated observations within each ID. Intervals condition on saved fits, folds and backgrounds, not training seeds. †: rare supporting classes, undefined resamples, or a degenerate empirical interval; a zero-width interval does not establish certainty. –: context unavailable, no calibrated policy, or no correct answers. NoLiMa FPR uses the separately archived calibration audit, not the AUROC fits.

Table 13: Source recall@1 per context (↑; automatic source selection)

	NoLiMa				BABILong			
Model	8K	16K	32K	Overall	8K	16K	32K	Overall
Gemma-4-E4B	1.000† [1.000, 1.000]	1.000† [1.000, 1.000]	0.958 [0.896, 1.000]	0.986 [0.965, 1.000]	0.600 [0.530, 0.670]	0.480 [0.410, 0.550]	0.265 [0.205, 0.325]	0.448 [0.398, 0.498]
Gemma-3-4B	1.000† [1.000, 1.000]	1.000† [1.000, 1.000]	0.771 [0.646, 0.875]	0.924 [0.882, 0.958]	0.410 [0.345, 0.480]	0.410 [0.340, 0.480]	0.315 [0.250, 0.380]	0.378 [0.328, 0.428]
Llama-3.1-8B	1.000† [1.000, 1.000]	1.000† [1.000, 1.000]	0.979 [0.938, 1.000]	0.993 [0.979, 1.000]	0.545 [0.475, 0.610]	0.385 [0.315, 0.455]	0.100 [0.060, 0.145]	0.343 [0.302, 0.385]
Qwen3-4B	0.938 [0.854, 1.000]	0.938 [0.854, 1.000]	0.771 [0.646, 0.875]	0.882 [0.826, 0.931]	0.695 [0.630, 0.760]	0.685 [0.620, 0.750]	0.505 [0.435, 0.575]	0.628 [0.575, 0.682]
Gemma-4-12B	1.000† [1.000, 1.000]	1.000† [1.000, 1.000]	0.979 [0.938, 1.000]	0.993 [0.979, 1.000]	0.940 [0.905, 0.970]	0.835 [0.785, 0.885]	0.720 [0.655, 0.780]	0.832 [0.798, 0.865]
Ministral-8B	1.000† [1.000, 1.000]	1.000† [1.000, 1.000]	0.958 [0.896, 1.000]	0.986 [0.965, 1.000]	0.095 [0.055, 0.135]	0.120 [0.075, 0.165]	–	0.107 [0.075, 0.142]
Phi-4	1.000† [1.000, 1.000]	0.979 [0.938, 1.000]	–	0.990 [0.969, 1.000]	0.550 [0.480, 0.620]	0.265 [0.205, 0.325]	–	0.407 [0.357, 0.458]
Qwen3-14B	1.000† [1.000, 1.000]	1.000† [1.000, 1.000]	1.000† [1.000, 1.000]	1.000† [1.000, 1.000]	0.775 [0.715, 0.830]	0.730 [0.670, 0.790]	0.570 [0.505, 0.640]	0.692 [0.643, 0.738]
Qwen3.8-27B	1.000† [1.000, 1.000]	1.000† [1.000, 1.000]	1.000† [1.000, 1.000]	1.000† [1.000, 1.000]	0.995 [0.985, 1.000]	0.990 [0.975, 1.000]	0.965 [0.940, 0.990]	0.983 [0.968, 0.995]
Llama-3.2-11B	1.000† [1.000, 1.000]	1.000† [1.000, 1.000]	0.979 [0.938, 1.000]	0.993 [0.979, 1.000]	0.535 [0.465, 0.600]	0.345 [0.280, 0.415]	0.075 [0.040, 0.115]	0.318 [0.278, 0.358]

Each cell gives the estimate above its pointwise 95% paired cluster bootstrap interval (10,000 resamples). NoLiMa: 48 facts; BABILong: 200 stories. Overall pools contexts while retaining repeated observations within each ID. Intervals condition on saved fits, folds and backgrounds, not training seeds. †: rare supporting classes, undefined resamples, or a degenerate empirical interval; a zero-width interval does not establish certainty. –: no admitted records at that context. Source recall does not require both answer-outcome classes.

Table 14: Selective re-injection on BABILong by context length

		BABILong			
Model	Context	Base (%)	Re-inj. (%)	Δ (pp)	Int. (%)
Gemma-4-E4B	8K	49.0 [42.5, 56.0]	60.5 [53.5, 67.5]	11.5 [5.5, 18.0]	77.0 [71.0, 82.5]
	16K	27.5 [21.5, 34.0]	47.0 [40.0, 54.0]	19.5 [13.5, 26.0]	92.5 [88.5, 96.0]
	32K	12.5 [8.0, 17.5]	26.0 [20.0, 32.0]	13.5 [8.5, 19.0]	96.0 [93.0, 98.5]
	Overall	29.7 [25.5, 34.2]	44.5 [39.5, 49.5]	14.8 [11.0, 18.7]	88.5 [85.3, 91.5]
Qwen3-4B	8K	54.0 [47.0, 60.5]	67.0 [60.5, 73.5]	13.0 [7.0, 19.0]	66.0 [59.5, 72.5]
	16K	45.5 [38.5, 52.5]	64.5 [58.0, 71.0]	19.0 [13.0, 25.5]	67.5 [61.0, 74.0]
	32K	28.5 [22.5, 35.0]	43.0 [36.0, 50.0]	14.5 [8.5, 20.5]	80.5 [75.0, 85.5]
	Overall	42.7 [37.5, 48.2]	58.2 [53.0, 63.3]	15.5 [11.5, 19.8]	71.3 [66.5, 76.0]
Llama-3.2-11B	8K	27.5 [21.5, 34.0]	54.0 [47.0, 61.0]	26.5 [20.0, 33.0]	75.5 [69.5, 81.5]
	16K	0.0† [0.0, 0.0]	25.5 [19.5, 31.5]	25.5 [19.5, 31.5]	99.0 [97.5, 100.0]
	32K	0.0† [0.0, 0.0]	3.5 [1.0, 6.0]	3.5 [1.0, 6.0]	100.0† [100.0, 100.0]
	Overall	9.2 [7.2, 11.3]	27.7 [24.2, 31.2]	18.5 [15.3, 21.7]	91.5 [89.3, 93.5]

Estimate above pointwise 95% paired story-bootstrap interval (10,000 resamples); 200 independent stories at each context. Overall retains all contexts within story. †: degenerate empirical interval, not certainty. The cohort amendment followed inspection of earlier outcomes. RULER and Scatter-MCQ intervals require their individual paired records, which are unavailable here.

Table 15: Failure-detection AUROC per context (↑; chance $= 0.5$) Each cell gives the estimate above its pointwise 95% paired cluster bootstrap interval (10,000 resamples, root seed 7

	RULER				Scatter-MCQ			
Model	8K	16K	32K	Overall	8K	16K	32K	Overall
Gemma-4-E4B	0.943 [0.912, 0.968]	0.967 [0.948, 0.983]	0.950 [0.925, 0.970]	0.955 [0.941, 0.967]	0.664 [0.615, 0.712]	0.644 [0.592, 0.695]	0.657 [0.596, 0.716]	0.675 [0.645, 0.704]
Gemma-3-4B	0.942 [0.913, 0.965]	0.947 [0.919, 0.970]	0.976 [0.964, 0.986]	0.954 [0.941, 0.965]	0.616 [0.542, 0.688]	0.586 [0.505, 0.667]	0.579 [0.495, 0.662]	0.602 [0.557, 0.646]
Llama-3.1-8B	0.715 [0.585, 0.825]	0.741 [0.645, 0.827]	0.778 [0.725, 0.828]	0.763 [0.719, 0.804]	0.677 [0.622, 0.728]	0.587 [0.523, 0.650]	0.612 [0.544, 0.680]	0.651 [0.614, 0.686]
Qwen3-4B	0.780 [0.735, 0.822]	0.898 [0.867, 0.928]	0.801 [0.759, 0.842]	0.824 [0.802, 0.846]	0.676 [0.624, 0.726]	0.620 [0.558, 0.680]	0.635 [0.580, 0.688]	0.650 [0.617, 0.682]
Gemma-4-12B	0.918 [0.890, 0.943]	0.943 [0.919, 0.964]	0.957 [0.936, 0.974]	0.941 [0.927, 0.953]	0.680 [0.631, 0.727]	0.669 [0.620, 0.718]	0.669 [0.613, 0.720]	0.692 [0.664, 0.719]
Ministral-8B	0.885 [0.851, 0.916]	0.874 [0.839, 0.908]	–	0.881 [0.857, 0.904]	0.634 [0.572, 0.693]	0.603 [0.541, 0.663]	–	0.624 [0.580, 0.667]
Phi-4	0.604 [0.518, 0.692]	–	–	0.604 [0.518, 0.692]	0.726 [0.676, 0.775]	–	–	0.726 [0.676, 0.775]
Qwen3-14B	0.472 [0.074, 0.946]	0.778 [0.606, 0.915]	0.885 [0.827, 0.935]	0.753 [0.624, 0.865]	0.685 [0.550, 0.823]	0.466 [0.326, 0.639]	0.455 [0.299, 0.632]	0.532 [0.450, 0.622]
Qwen3.8-27B	0.676 [0.618, 0.733]	0.673 [0.612, 0.733]	0.756 [0.702, 0.805]	0.706 [0.673, 0.738]	0.512 [0.399, 0.626]	0.557 [0.469, 0.651]	0.404 [0.303, 0.506]	0.501 [0.445, 0.557]
Llama-3.2-11B	0.686 [0.554, 0.801]	0.670 [0.559, 0.772]	0.769 [0.709, 0.824]	0.730 [0.681, 0.777]	0.644 [0.587, 0.698]	0.550 [0.486, 0.615]	0.607 [0.541, 0.673]	0.619 [0.582, 0.654]

Table 16: Failure-detection recall (TPR) per context at the calibrated threshold (↑) Each cell gives the estimate above its pointwise 95% paired cluster bootstrap interval (10,000 resamples, root seed 7

	RULER				Scatter-MCQ			
Model	8K	16K	32K	Overall	8K	16K	32K	Overall
Gemma-4-E4B	0.956 [0.908, 0.990]	0.979 [0.948, 1.000]	0.907 [0.846, 0.958]	0.946 [0.918, 0.970]	0.948 [0.922, 0.970]	0.976 [0.960, 0.989]	0.996 [0.990, 1.000]	0.975 [0.965, 0.983]
Gemma-3-4B	0.963 [0.934, 0.988]	0.995 [0.984, 1.000]	1.000 [1.000, 1.000]	0.987 [0.978, 0.995]	0.967 [0.950, 0.983]	0.985 [0.974, 0.994]	0.998 [0.994, 1.000]	0.984 [0.977, 0.990]
Llama-3.1-8B	0.857 [0.684, 1.000]	0.929 [0.816, 1.000]	0.961 [0.900, 1.000]	0.930 [0.876, 0.975]	0.963 [0.943, 0.980]	0.994 [0.986, 1.000]	0.996 [0.990, 1.000]	0.985 [0.978, 0.991]
Qwen3-4B	0.944 [0.894, 0.988]	0.990 [0.966, 1.000]	0.873 [0.806, 0.936]	0.935 [0.905, 0.962]	0.956 [0.934, 0.974]	0.983 [0.970, 0.994]	0.994 [0.985, 1.000]	0.978 [0.970, 0.985]
Gemma-4-12B	0.972 [0.937, 1.000]	0.992 [0.973, 1.000]	1.000 [1.000, 1.000]	0.989 [0.978, 0.997]	0.950 [0.926, 0.972]	0.983 [0.970, 0.995]	0.993 [0.985, 1.000]	0.977 [0.969, 0.985]
Ministral-8B	0.929 [0.874, 0.976]	0.924 [0.874, 0.968]	–	0.926 [0.890, 0.959]	0.966 [0.947, 0.983]	0.986 [0.975, 0.996]	–	0.976 [0.966, 0.986]
Phi-4	0.939 [0.846, 1.000]	–	–	0.939 [0.846, 1.000]	0.983 [0.970, 0.994]	–	–	0.983 [0.970, 0.994]
Qwen3-14B	0.976 [0.961, 0.990]	0.953 [0.931, 0.974]	0.938 [0.913, 0.962]	0.956 [0.944, 0.967]	0.990 [0.981, 0.997]	0.964 [0.947, 0.979]	0.913 [0.891, 0.934]	0.956 [0.946, 0.965]
Qwen3.8-27B	0.956 [0.929, 0.980]	0.950 [0.924, 0.975]	0.930 [0.898, 0.959]	0.945 [0.929, 0.961]	0.988 [0.979, 0.996]	0.961 [0.945, 0.975]	0.900 [0.874, 0.924]	0.950 [0.939, 0.960]
Llama-3.2-11B	0.909 [0.769, 1.000]	0.885 [0.750, 1.000]	0.980 [0.932, 1.000]	0.939 [0.887, 0.980]	0.959 [0.938, 0.977]	0.994 [0.986, 1.000]	0.998 [0.994, 1.000]	0.984 [0.977, 0.991]

Table 17: Failure-detection FPR per context at the calibrated threshold (↓) Each cell gives the estimate above its pointwise 95% paired cluster bootstrap interval (10,000 resamples, root seed 7

	RULER				Scatter-MCQ			
Model	8K	16K	32K	Overall	8K	16K	32K	Overall
Gemma-4-E4B	0.256 [0.207, 0.306]	0.262 [0.213, 0.312]	0.252 [0.202, 0.303]	0.257 [0.228, 0.286]	0.852 [0.802, 0.899]	0.966 [0.933, 0.993]	0.991 [0.971, 1.000]	0.923 [0.898, 0.946]
Gemma-3-4B	0.194 [0.141, 0.251]	0.287 [0.225, 0.352]	0.453 [0.375, 0.529]	0.302 [0.264, 0.340]	0.873 [0.798, 0.941]	0.967 [0.921, 1.000]	0.984 [0.945, 1.000]	0.936 [0.902, 0.967]
Llama-3.1-8B	0.699 [0.651, 0.747]	0.661 [0.612, 0.709]	0.540 [0.488, 0.593]	0.635 [0.606, 0.664]	0.921 [0.869, 0.969]	0.980 [0.947, 1.000]	0.986 [0.957, 1.000]	0.955 [0.928, 0.979]
Qwen3-4B	0.384 [0.329, 0.439]	0.370 [0.315, 0.426]	0.366 [0.311, 0.420]	0.373 [0.342, 0.405]	0.901 [0.854, 0.944]	0.984 [0.959, 1.000]	0.992 [0.974, 1.000]	0.953 [0.931, 0.972]
Gemma-4-12B	0.232 [0.185, 0.283]	0.178 [0.133, 0.225]	0.187 [0.140, 0.236]	0.200 [0.173, 0.228]	0.875 [0.829, 0.919]	0.934 [0.893, 0.970]	0.972 [0.943, 0.993]	0.918 [0.894, 0.941]
Ministral-8B	0.225 [0.177, 0.274]	0.213 [0.167, 0.263]	–	0.219 [0.185, 0.254]	0.936 [0.876, 0.984]	0.968 [0.929, 1.000]	–	0.950 [0.912, 0.981]
Phi-4	0.723 [0.676, 0.769]	–	–	0.723 [0.676, 0.769]	0.868 [0.808, 0.921]	–	–	0.868 [0.808, 0.921]
Qwen3-14B	0.800 [0.333, 1.000]	1.000 [1.000, 1.000]	0.857 [0.500, 1.000]	0.905 [0.762, 1.000]	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]
Qwen3.8-27B	0.832 [0.767, 0.892]	0.844 [0.771, 0.909]	0.740 [0.663, 0.815]	0.804 [0.763, 0.845]	0.958 [0.857, 1.000]	0.967 [0.885, 1.000]	0.931 [0.829, 1.000]	0.952 [0.900, 0.989]
Llama-3.2-11B	0.789 [0.746, 0.829]	0.734 [0.689, 0.779]	0.651 [0.601, 0.700]	0.726 [0.699, 0.752]	0.885 [0.825, 0.939]	0.980 [0.948, 1.000]	1.000 [1.000, 1.000]	0.945 [0.916, 0.970]

Table 18: Selective re-injection by context length Each cell gives the estimate above its pointwise 95% paired cluster bootstrap interval (10,000 resamples, root seed 7

		RULER				Scatter-MCQ			
Model	Context	Base (%)	Re-inj. (%)	Δ (pp)	Int. (%)	Base (%)	Re-inj. (%)	Δ (pp)	Int. (%)
Gemma-4-E4B	8K	75.5 [69.5, 81.5]	75.5 [69.5, 81.5]	0.0 [0.0, 0.0]	46.0 [39.0, 53.0]	28.5 [22.5, 34.5]	37.5 [31.0, 44.5]	9.0 [2.5, 15.5]	95.0 [92.0, 98.0]
	16K	71.5 [65.0, 77.5]	71.5 [65.0, 77.5]	0.0 [0.0, 0.0]	47.0 [40.0, 54.0]	20.5 [15.0, 26.0]	28.5 [22.5, 35.0]	8.0 [2.0, 14.0]	99.5 [98.5, 100.0]
	32K	75.5 [69.5, 81.0]	76.0 [70.5, 81.5]	0.5 [0.0, 1.5]	40.0 [33.5, 47.0]	15.5 [10.5, 21.0]	27.0 [21.0, 33.5]	11.5 [5.0, 18.0]	100.0 [100.0, 100.0]
	Overall	74.2 [70.7, 77.7]	74.3 [70.8, 77.7]	0.2 [0.0, 0.5]	44.3 [40.3, 48.3]	21.5 [18.3, 24.8]	31.0 [27.3, 34.7]	9.5 [5.8, 13.2]	98.2 [97.0, 99.2]
Qwen3-4B	8K	75.5 [69.5, 81.0]	75.5 [69.5, 81.0]	0.0 [0.0, 0.0]	57.0 [50.0, 64.0]	26.5 [20.5, 33.0]	31.0 [24.5, 37.5]	4.5 [-2.0, 11.0]	95.5 [92.5, 98.0]
	16K	71.5 [65.0, 77.5]	72.0 [65.5, 78.0]	0.5 [0.0, 1.5]	56.0 [49.0, 63.0]	19.6 [14.6, 25.1]	24.1 [18.1, 30.2]	4.5 [-1.0, 10.1]	98.0 [96.0, 99.5]
	32K	77.5 [71.5, 83.5]	77.5 [71.5, 83.5]	0.0 [0.0, 0.0]	49.5 [42.5, 56.5]	18.6 [13.6, 24.1]	26.6 [20.6, 32.7]	8.0 [1.5, 15.1]	99.0 [97.5, 100.0]
	Overall	74.8 [71.3, 78.3]	75.0 [71.5, 78.5]	0.2 [0.0, 0.5]	54.2 [50.2, 58.2]	21.6 [18.4, 24.9]	27.3 [23.6, 30.8]	5.7 [2.2, 9.2]	97.5 [96.2, 98.7]
Llama-3.2-11B	8K	93.0 [89.0, 96.5]	97.0 [94.5, 99.0]	4.0 [0.0, 8.0]	81.5 [76.0, 86.5]	19.0 [13.5, 24.5]	28.0 [22.0, 34.5]	9.0 [3.0, 15.0]	94.5 [91.0, 97.5]
	16K	93.5 [90.0, 96.5]	97.5 [95.0, 99.5]	4.0 [1.0, 7.0]	73.0 [66.5, 79.0]	14.6 [10.1, 19.6]	18.6 [13.6, 24.1]	4.0 [-2.5, 10.6]	100.0 [100.0, 100.0]
	32K	88.5 [84.0, 93.0]	95.0 [92.0, 97.5]	6.5 [2.0, 11.0]	72.0 [66.0, 78.0]	13.6 [9.0, 18.6]	17.6 [12.6, 23.1]	4.0 [-2.0, 10.1]	100.0 [100.0, 100.0]
	Overall	91.7 [89.3, 93.8]	96.5 [95.0, 97.8]	4.8 [2.7, 7.2]	75.5 [72.0, 78.8]	15.7 [12.9, 18.7]	21.4 [18.2, 24.7]	5.7 [2.0, 9.4]	98.2 [97.0, 99.2]

E ATTENTION FEATURES FOR SOURCE RANKING AND FAILURE DETECTION

Both probes in Section 4 use attention measured at the final encoded prompt position, before the first answer token. The ranker describes every candidate message with 14 features. The failure probe uses nine of these features for the selected message. Source annotations and later answer outcomes supervise fitting; neither enters feature extraction at inference.

Attention measurements and notation. Let $A_{\ell hj}$ be the attention mass from the final prompt position to all tokens in candidate message c_j, at global-attention layer ℓ and head h. We sum over message tokens, then average over heads to obtain the layer profile $s_{\ell j}$. Let q_ℓ and b_ℓ be head-mean mass on the question and full history, respectively, and n_j the message's token count. Only global-attention layers enter these features. With L such layers, the late set T contains the last $\lceil L/4 \rceil$ layers in decoder order.

For numerical stability, floor $s_{\ell j}$ and q_ℓ at $\epsilon = 10^{-12}$ and define $o_{\ell j} = \max(b_\ell - s_{\ell j}, \epsilon)$, the remaining-history mass. Below, bars denote means over all L layers, $p_{\ell j} = s_{\ell j} / \sum_m s_{mj}$, and $p_{(i)j}$ is the ith largest layer share. All logarithms are natural.

Nine routing features shared by both probes. The raw vector ρ_j contains the following features, in order. The first three measure strength, the next four describe its distribution over layers, and the last two compare the source with other prompt regions.

Feature	Definition
Log mean source mass	$\log \bar{s}_j$
Log peak-layer source mass	$\log \max_\ell s_{\ell j}$
Log late-layer mean mass	$\log\big(\lvert T\rvert^{-1} \sum_{\ell \in T} s_{\ell j}\big)$
Late-layer fraction	$\sum_{\ell \in T} p_{\ell j}$
Normalized layer entropy	$-\sum_\ell p_{\ell j} \log p_{\ell j} / \log L$
Top-one layer share	$p_{(1)j}$
Top-three layer share	$\sum_{i=1}^{\min(3,L)} p_{(i)j}$
Log source-to-question ratio	$\log(\bar{s}_j / \bar{q})$
Log source-to-other-history ratio	$\log(\bar{s}_j / \bar{o}_j)$

The ratios compare attention mass, not token counts. If $L = 1$, the implementation sets entropy and both top-layer shares to zero. The failure probe receives $\psi_{\hat{j}}$, the standardized form of $\rho_{\hat{j}}$ for the ranked source.

Five additional source-ranking features. The ranker augments ρ_j with density, head specialization and competition among candidates. Let $P_j = \max_{\ell,h} A_{\ell hj}$ and $P_j^T = \max_{\ell \in T, h} A_{\ell hj}$, each floored at ϵ. Let D be the sum of unfloored mean attention masses across all candidates, floored at ϵ. The additional features, in order, are

$$\log(\bar{s}_j / n_j), \quad \log P_j, \quad \log P_j^T, \quad \bar{s}_j / D, \quad \log(P_j / \bar{s}_j).$$

They represent attention per message token, the strongest individual head over all layers and over late layers, the candidate's share of candidate attention, and peak-to-mean head selectivity. Together with ρ_j, they form the 14 raw features standardized into ϕ_j.

Standardization and probe fitting. Each probe has its own feature-wise standardizer fitted on its training inputs. Means and scales are frozen for calibration and evaluation. Both linear probes are fitted with logistic loss and L_2 regularization ($C = 1$); source ranking uses the linear score, while failure prediction applies a sigmoid. The warning threshold is fixed on disjoint calibration data as specified in Section 4.

E.1 INDIVIDUAL DETECTOR FEATURE CONTRIBUTIONS

In Figure 9a, sensitivity is the AUROC decrease after permuting one captured global-attention layer across stories, averaged over 64 permutations. Shading marks the final third of model depth. Bands and bars in panels (a,b) are pointwise 95% story-bootstrap intervals from 1,000 resamples. Panel (c) reuses the annotated-option Scatter diagnostic, pooled across Gemma-4-E4B, Qwen3-4B, and

Llama-3.2-11B; its curves correspond to Q1, Q2, and Q3 at 25%, 75%, and 100% of the conversation budget.

The main figure summarizes feature families using their mean absolute contribution per feature. Here we retain all nine individual contributions for the same 200 BABILong stories at 8K, with automatic source selection and frozen probes. For standardized feature z_j, its contribution to the failure logit is $\beta_j z_j$; the figure reports the mean of its absolute value over stories. Intervals use 1,000 paired story-bootstrap resamples. These magnitudes describe score use; correlated contributions can cancel and do not identify unique predictive information or causal effects.

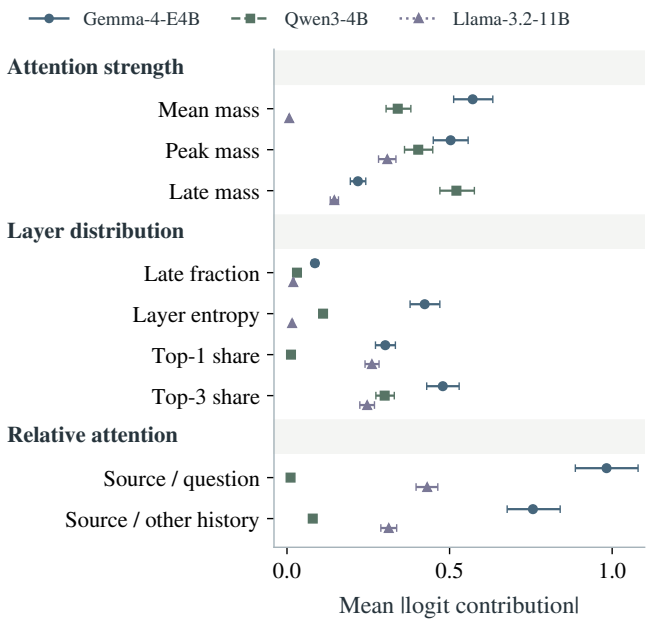


Figure 10: Individual routing-feature contributions on BABILong at 8K. Strength measures mean, peak, and late-layer source mass; distribution measures late fraction, normalized layer entropy, and top-1/top-3 layer shares; relative attention compares source mass with question and other-history mass. Late features use the final quarter of captured global layers. Mass and ratio features are log-transformed before standardization. Bars: pointwise 95% story-bootstrap intervals.

F Benchmarks and Evaluation Setup

We evaluate our pre-generation failure detector on four benchmarks adapted to a long-context, multi-turn conversational setting: *NoLiMa* Modarressi et al. (2025), *RULER* Hsieh et al. (2024), *BABILong*Kuratov et al. (2024), and *Scatter*, our proposed evaluation setup constructed from MMLU Hendrycks et al. (2020) questions and WildChat Zhao et al. (2024) conversations. These benchmarks assess associative retrieval, needle-style fact retrieval, location tracking, and scattered multiple-choice recall, respectively. We apply a common procedure for converting benchmark instances into conversational turns and detail the benchmark-specific experimental configurations below.

Motivation. Long-context retrieval in multi-turn dialogue requires models to recover relevant evidence from earlier turns amid intervening content. Representing the context as a single monolithic prompt obscures these conversational boundaries. We therefore formulate four retrieval settings, associative fact retrieval, needle retrieval, location tracking, and scattered multiple-choice recall, as multi-turn conversations at matched context lengths. This design enables us to assess whether an attention-based detector can localize evidence-bearing turns and predict retrieval failures across tasks and models, and whether its predictive performance generalizes beyond individual benchmarks.

Constructing multi-turn contexts. Across all four benchmarks, we organize each long context as a multi-turn conversation, using fixed-size background chunks as user turns for synthetic benchmarks

and preserving natural exchanges for conversational benchmarks. Depending on the benchmark, task-relevant evidence consists of an authored fact, a needle, a movement statement, or MMLU answer options, as detailed below. We place this evidence in designated turns interspersed with filler turns, all preceding the final query. The model then generates a response to this query. This construction provides discrete candidate turns for source localization and a common setting for pre-generation retrieval-failure detection. All comparisons use three context lengths: 8K, 16K, and 32K tokens (8,192, 16,384, and 32,768, respectively).

NoLiMa (associative retrieval). We embed 48 authored facts within unrelated MMLU exchanges and pose indirect questions that require retrieving a name through one of four association types, such as landmark-to-city associations (Modarressi et al., 2025). Our initial evaluation uses fact-grouped cross-validation. A subsequent evaluation introduces matched distractor facts and holds out both relation families and background contexts: the test relation family and its associated backgrounds are excluded from training, while a separate relation family is used to calibrate the alert threshold. Detection performance is estimated using nested, fact-grouped out-of-fold predictions.

RULER (needle-in-a-haystack). We construct this benchmark from the NVIDIA RULER synthetic suite, embedding target facts, or needles, within synthetic filler and querying them directly. We use eight needle-in-a-haystack tasks (`niah_single_{1,2,3}`, `niah_multikey_{1,2,3}`, `niah_multivalue`, and `niah_multiquery`) with disjoint training and test questions. These tasks provide an isolatable source turn for localization, whereas the remaining tasks (`vt`, `cwe`, `fwe`, and `qa`) serve only as filler. We convert each context into user turns of approximately 500 tokens, respecting sentence and line boundaries and preserving each needle as a separate, complete turn. Detector fitting follows the nested, source-grouped out-of-fold calibration procedure used throughout. We calibrate the alert threshold to a fixed target true-positive rate on the training split and hold it fixed when evaluating the disjoint test split.

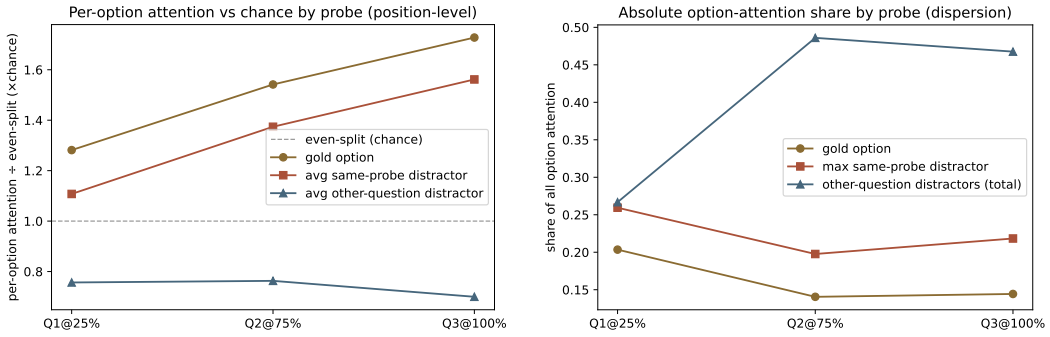
BABILong (location tracking). We interleave bAbI QA1 movement facts with book passages and query the latest location (Kuratov et al., 2024): 60 fitting, 20 calibration and 40 evaluation stories, with disjoint books across partitions. A further 576 stories test the frozen method using only the evaluation books. Stories are paired across lengths and models, and short-story controls assess competence without distractors.

Scatter (scattered MCQ). We construct Scatter conversations from real WildChat dialogues, preserving natural exchange boundaries and turn lengths within each token budget. We use these dialogues as filler and insert three MMLU questions at approximately 25%, 75%, and 100% of the context length. Each question's four answer options are scattered as bracketed notes across preceding user turns, strictly before the question itself. Evaluation proceeds incrementally, with the model answering each question with a single letter before receiving further context. We record correctness for each question separately. For each context length, we use 500 training and 200 test conversations with globally disjoint MMLU items. To control for underlying task competence, we also evaluate a test-only baseline that removes filler while retaining the same question triples.

Metrics. We evaluate failure detection using threshold-independent ranking and a calibrated operating point. Our primary metric is the area under the receiver operating characteristic curve (AUROC), computed from predicted failure-risk scores and binary failure labels, where 0.5 denotes chance-level ranking. We calibrate the alert threshold using nested out-of-fold training predictions, selecting the highest threshold that achieves a true-positive rate (TPR, or failure recall) of at least 95%. Holding this threshold fixed, we report the achieved TPR (TPR@95) and false-positive rate (FPR) on the disjoint test split. We also assess source localization using source recall@1, defined as the proportion of queries for which the top-ranked candidate turn corresponds to the true evidence source.

F.1 Motivation: Where Attention Goes on Scattered Options

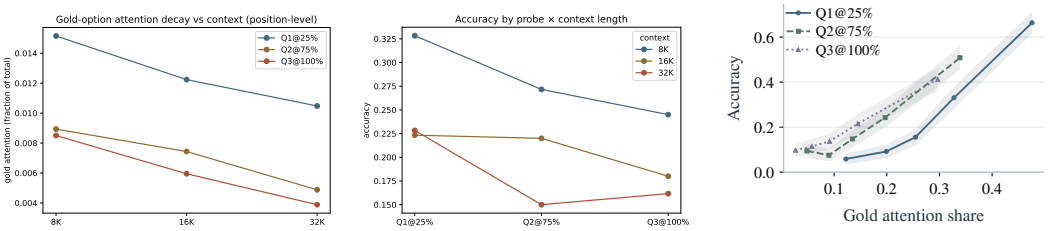
We examine how attention at the answer token of each scatter probe is distributed across scattered multiple-choice options. We measure attention separately for each `[(letter) text]` option span, distinguishing the gold option, same-question distractors, and options from other questions. We



(a) Per-option attention ÷ even-split baseline (chance = 1.0). Gold and same-question distractors are above chance and rise with probe position, while other-question options stay below chance.

(b) Absolute share of option attention. The gold option's slice shrinks as attention disperses over more options at deeper probes

Figure 11: Answer-time attention over scattered options, by probe. *(a)* Per-option *preference* for the gold option rises with probe depth, while *(b)* its *absolute* share falls as attention disperses over more options.



(a) Gold attention (fraction of all attention) vs. context length.

(b) Accuracy by probe and context length.

(c) Accuracy vs. gold-attention share (quintiles).

Figure 12: Long-context behaviour and the attention–correctness link. *(a)* the gold option's absolute attention decays with context length, and *(b)* accuracy tracks that decay across probes and lengths, while *(c)* accuracy rises with how much attention the gold option draws.

report results by probe position (Q1, Q2, and Q3 at 25%, 75%, and 100% of the budget, respectively), pooled across Gemma-4-E4B, Qwen3-4B, and Llama-3.2-11B.

Gold-option enrichment increases across probes, while its declining attention share predicts answer correctness. As the number of contextual options increases ($\sim$!4, $\sim$!8, and $\sim$!12 at Q1–Q3), we normalise per-option attention shares by the uniform baseline ($1/n$), with values above 1 indicating enrichment (Figs. 11a and 11b). Gold-option enrichment rises from $1.28\times$ to $1.54\times$ and $1.73\times$, alongside increasing enrichment of same-question distractors ($1.11 \rightarrow 1.37 \rightarrow 1.56\times$), while other-question options remain below baseline (0.70–$0.76\times$). Within each probe's four options, the gold share remains nearly constant ($0.276/0.270/0.270$; uniform baseline 0.25), indicating stable, modest gold selectivity. However, its share of total option attention falls from 0.20 at Q1 to 0.14 at Q2–Q3, while the combined share allocated to other-question options rises to $\sim$!0.47. Thus, the gold option remains individually favoured but receives a smaller share as attention disperses across more competing options. This decline accompanies reduced accuracy at later probes (Figs. 12a and 12b). Gold attention is predictive of answer correctness, with pooled AUROCs of $0.82/0.79/0.74$ at Q1/Q2/Q3 and point-biserial correlations declining from 0.53 to 0.39. Accuracy also increases monotonically across gold-share bins (Fig. 12c). The association is strongest at Q1 and weakens at later probes, although lower gold attention remains associated with a greater likelihood of error. Together, these results suggest that the attention share reaching the gold option carries a failure signal even when its relative enrichment remains high. Table 19 reports AUROC and recall for Scatter-MCQ questions at different conversation depths.

G LATENCY

Table 19: Scatter-MCQ failure detection (AUROC / FPR) and source recall@1 per question slot. Each conversation carries three probe questions (Q1–Q3) asked in order. AUROC ↑ (chance = 0.5); FPR at the calibrated threshold ↓; R@1: source recall@1 with automatic source selection ↑.

	AUROC / FPR				Source R@1			
Model	Q1	Q2	Q3	Overall	Q1	Q2	Q3	Overall
Gemma-4-E4B	0.651/0.883	0.708/0.954	0.643/0.938	0.675/0.923	0.408	0.315	0.238	0.321
Gemma-3-4B	0.682/0.878	0.572/0.969	0.544/0.969	0.602/0.936	0.326	0.196	0.166	0.229
Llama-3.1-8B	0.649/0.977	0.654/0.945	0.614/0.935	0.651/0.955	0.360	0.222	0.172	0.251
Qwen3-4B	0.643/0.933	0.642/0.969	0.663/0.961	0.650/0.953	0.395	0.281	0.268	0.314
Gemma-4-12B	0.648/0.896	0.690/0.914	0.703/0.959	0.692/0.918	0.453	0.328	0.273	0.352
Ministral-8B	0.697/0.924	0.589/0.962	0.572/0.968	0.624/0.950	0.385	0.228	0.153	0.256
Phi-4	0.693/0.900	0.744/0.762	0.701/0.941	0.726/0.868	0.480	0.305	0.210	0.332
Qwen3-14B	0.534/1.000	0.466/1.000	0.505/1.000	0.532/1.000	0.391	0.273	0.236	0.300
Qwen3.8-27B	0.521/1.000	0.448/1.000	0.407/0.886	0.501/0.952	0.237	0.100	0.062	0.133
Llama-3.2-11B	0.605/0.962	0.612/0.940	0.602/0.926	0.619/0.945	0.358	0.217	0.179	0.251

Table 20: Latency overhead of the detector on RULER (Δ = with detector $-$ without, mean seconds per query)

Model	8K	16K	32K	Overall
Gemma-4-E4B	$+0.05$	$+0.05$	$+0.03$	$+0.04$
Gemma-3-4B	-0.17	-0.23	-0.60	-0.33
Llama-3.1-8B	$+0.08$	$+0.17$	$+0.25$	$+0.17$
Qwen3-4B	$+0.06$	$+0.10$	$+0.16$	$+0.11$
Gemma-4-12B	$+0.01$	$+0.03$	-0.52	-0.16
Ministral-8B	$+0.06$	$+0.11$	–	$+0.09$
Phi-4	$+0.20$	–	–	$+0.20$
Qwen3-14B	$+0.16$	$+0.29$	$+0.58$	$+0.34$
Llama-3.2-11B	$+0.09$	$+0.14$	$+0.20$	$+0.15$