Taking the PULSE of Medical Reasoning: Step-Level Retrieval Timing for Knowledge-Grounded Inference

Anonymous ACL submission

Abstract

Reliable clinical reasoning requires LLMs to dynamically access external biomedical evidence to improve factual trustworthiness and reasoning accuracy. However, indiscriminate retrieval can introduce noise and latency, raising a critical retrieval-timing problem: at each reasoning step, the system must decide whether the model's internal knowledge is reliable or external evidence is needed. We propose **PULSE**, a lightweight probing method that uses the model's internal representations as intrinsic signals for step-level retrieval control. Our posterior trajectory analysis reveals that erroneous steps exhibit substantially higher inter-layer volatility than correct ones, suggesting unstable internal knowledge. Based on this finding, PULSE trains a compact probe on inter-layer geometric features to detect factual errors upon step completion, and uses it as an adaptive trigger for retrieval and step regeneration. Across four medical benchmarks, PULSE-timed retrieval matches exhaustive per-step retrieval in accuracy while reducing retrieval calls by 72%, and can transfer to an unseen legal domain, indicating potential domain-agnostic generality. Overall, our study provides an initial step toward shifting retrieval-trigger design from surface-level confidence estimates to internal activation dynamics for adaptive access.

1 Introduction

Reliable clinical reasoning requires Large Language Models (LLMs) to dynamically access external biomedical evidence to improve factual trustworthiness and reasoning accuracy (Tu et al., 2025; Singhal et al., 2023; Bayat et al., 2023; Berger et al., 2025). This need is particularly salient in medical question answering, where multi-step reasoning often depends on specialized and rapidly evolving knowledge. Importantly, factual errors may arise only at specific reasoning steps rather than throughout the entire reasoning chain. Thus, effective medical retrieval-augmented generation (RAG) should not retrieve evidence indiscriminately. Instead, it should determine *when* external evidence is needed during the reasoning process.

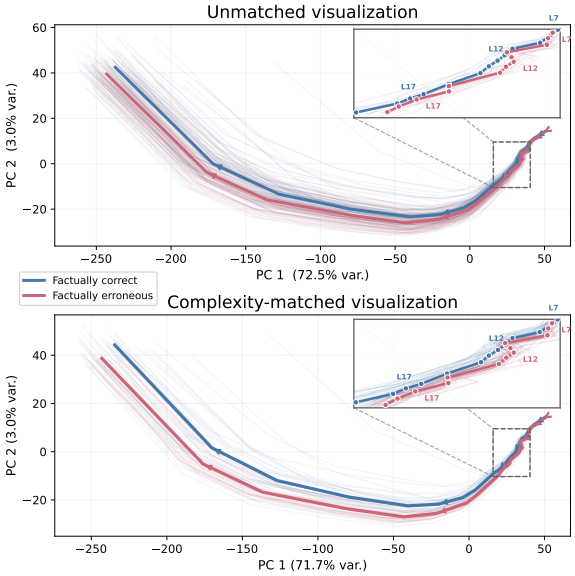

Figure 1: Layer-wise PCA trajectories of hidden states. Bold lines: group means; faint lines: individual samples. (a) Unmatched. (b) Complexity-matched.

Existing retrieval-augmented methods differ mainly in the granularity at which retrieval is triggered. Question-level retrieval retrieves evidence before reasoning begins (Lewis et al., 2020; Lu et al., 2025), but is too coarse to address factual errors that emerge only in intermediate reasoning steps. Exhaustive step-level retrieval can provide targeted support, but it incurs substantial retrieval cost and may introduce unnecessary or distracting evidence (Huang and Huang, 2024). Token-level retrieval further refines the trigger granularity (Jiang et al., 2023; Su et al., 2024), yet it fails to preserve the semantic integrity of a complete reasoning step. Step-level adaptive retrieval therefore offers a natural granularity (Yao et al., 2025), and its effectiveness hinges on a critical retrieval-timing problem: after each reasoning step, the system must decide

whether the model's internal knowledge is reliable or external evidence is required.

A central challenge is how to obtain reliable trigger signals. Most existing approaches rely on output-level uncertainty signals, such as perplexity (Si et al., 2022), entropy (Kuhn et al., 2023), or maximum softmax probability, to estimate whether the model is likely to be wrong. While useful, these signals are derived from the final output distribution and may provide only a surface-level view of the model's reasoning state. Another line of work has shown that factual knowledge and reasoning-related information are encoded in the model's hidden representations (Meng et al., 2022; Geva et al., 2021; Azaria and Mitchell, 2023; Lu et al., 2026), and that probes over multi-layer representations can help assess reasoning quality (Ni et al., 2025). These findings suggest that layer-wise internal dynamics may contain richer trigger signals.

Motivated by this perspective, we investigate whether the layer-wise evolution of hidden states can serve as an intrinsic signal for step-level factual error detection. Our posterior trajectory analysis reveals that factually erroneous reasoning steps exhibit substantially higher inter-layer volatility than correct ones, even after controlling for reasoning complexity, as shown in Figure 1. This pattern suggests that when a model fails to ground a step in its internal knowledge, its representations undergo unstable transitions across layers. We term this phenomenon *internal knowledge instability* and build upon it to propose PULSE (**P**robing **U**ngrounded reasoning via **L**ayer-wise **S**tate **E**volution), a lightweight probing method for adaptive retrieval control. PULSE extracts compact geometric features from hidden-state trajectories and detects factual errors upon step completion. Rather than retrieve at every step, it selectively triggers biomedical retrieval and step regeneration only when a potential factual error is detected, thereby reducing unnecessary retrieval while introducing negligible generation overhead.

Our study evaluates PULSE on four medical QA benchmarks. PULSE-timed retrieval matches exhaustive per-step retrieval in downstream accuracy while reducing retrieval calls by 72%, achieving a favorable balance between accuracy and retrieval cost. Moreover, the same trigger transfers zero-shot to an unseen legal domain, suggesting that layer-wise trajectory volatility may capture a domain-agnostic signal of factual grounding failure rather than dataset-specific artifacts. Overall, our contributions can be summarized as follows:

- **Geometric Signal.** We identify elevated inter-layer volatility in hidden states as a geometric signature of step-level factual errors, providing an intrinsic basis for retrieval timing.
- **Retrieval Trigger.** We propose PULSE, a lightweight probe that determines retrieval timing from the geometric evolution of hidden states across layers.
- **Broad Efficiency.** We show that PULSE-timed retrieval matches exhaustive retrieval with 72% fewer calls and generalizes zero-shot across domains.

2 Related Work

Retrieval-Augmented Generation (RAG) mitigates LLM factual deficiencies by incorporating external knowledge (Lewis et al., 2020). A central challenge in multi-step reasoning is deciding *when* to retrieve and *how* to retrieve—exhaustive per-step retrieval introduces latency and noise, while omitting retrieval at knowledge-deficient steps allows errors to propagate.

Retrieval Strategy. To handle multi-step reasoning, methods such as IRCoT (Trivedi et al., 2023) and ReAct (Yao et al., 2023) interleave retrieval with generation, while recent work improves query formulation and chunk selection (Su et al., 2024). These advances address the retrieval pipeline itself and are orthogonal to our contribution; PULSE can serve as a front-end trigger for any such retriever.

Retrieval Timing. The more fundamental question is *when* retrieval is needed (Tang et al., 2025). Existing triggering signals fall into two categories by their source.

At the output level, methods rely on generation-time statistics: token probability thresholds (Jiang et al., 2023), semantic entropy over sampled outputs (Kuhn et al., 2023), logit-based uncertainty regressors (Liu et al., 2024), and composite heuristics combining token probability with attention-based importance scoring (Su et al., 2024). Trained reflection tokens (Asai et al., 2024) also operate in the output space. These signals are cheap to compute but remain surface-level proxies that conflate linguistic uncertainty with genuine knowledge deficiency.

At the internal level, recent work examines hidden-state geometry for factual assessment.

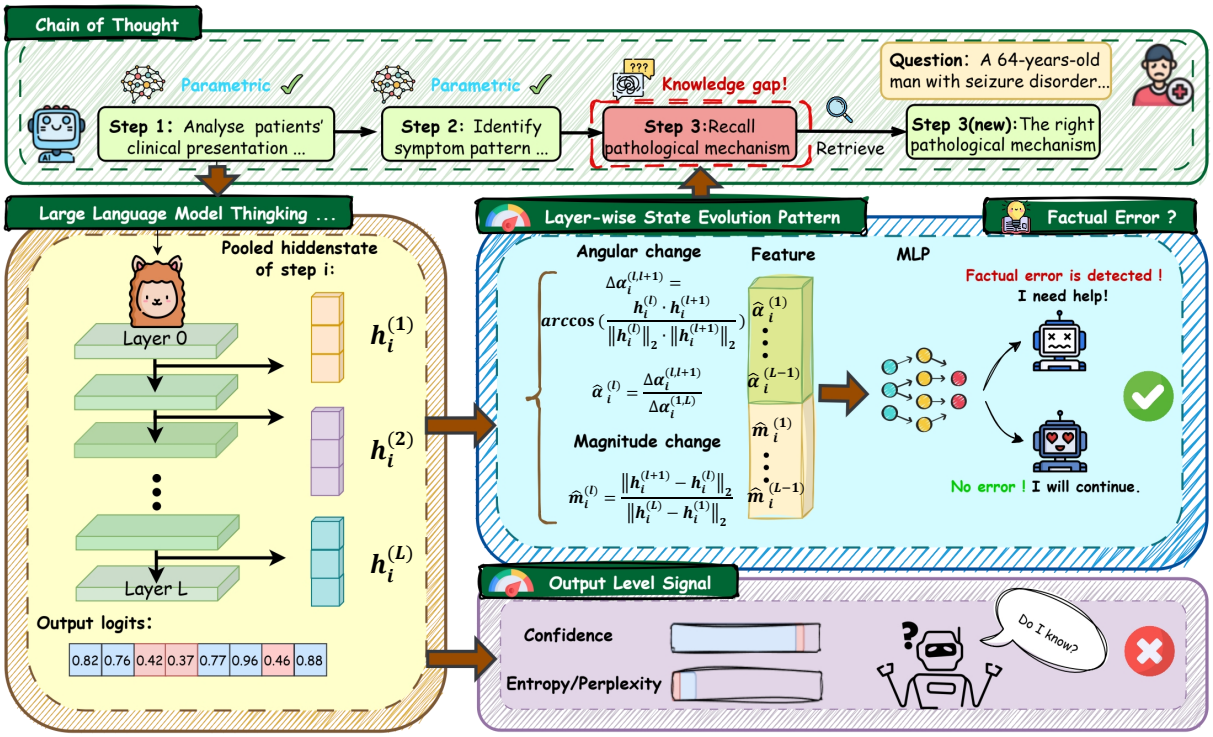

Figure 2: Overview of PULSE. Layer-wise geometric trajectory features of each reasoning step are fed into an MLP probe to detect factual errors. Flagged steps trigger retrieval, whose accuracy gain interventionally validates the detection. Output-level signals (purple) serve as baselines.

SeaKR (Yao et al., 2025) uses hidden-state uncertainty to trigger retrieval adaptively. CoE (Wang et al., 2025) aggregates cross-layer geometric feature changes into scalar scores, while Cheang et al. (2025) and Lee et al. (2025) analyze internal state geometry during knowledge recall and self-verification. However, these efforts operate at the granularity of a complete query or response—none localizes knowledge deficiency to individual reasoning steps.

PULSE bridges this gap by bringing layer-wise geometric trajectory analysis to the step level. It learns how trajectories co-evolve across layers, capturing discriminative evolution patterns that signal factual errors within each reasoning step and providing fine-grained, domain-transferable retrieval triggering signals.

3 Checking the PULSE

We propose PULSE (Probing Ungrounded via Layer-wise State Evolution), a method that detects factual errors at individual reasoning steps by exploiting the geometric structure of layer-wise hidden-state trajectories. Figure 2 illustrates the overall workflow.

3.1 Posterior Trajectory Analysis

To understand how a model's internal representations differ between factually correct and erroneous reasoning steps, we examine the layer-wise evolution of hidden states during chain-of-thought generation.

Step-level trajectory construction. Consider a reasoning step s_i spanning tokens $(t_i^{(1)}, \ldots, t_i^{(K_i)})$. For a Transformer with L layers, let $\mathbf{z}_{i,k}^{(l)} \in \mathbb{R}^d$ denote the hidden state of token $t_i^{(k)}$ at layer l. We obtain a single vector summarizing the step at each layer via mean pooling:

$$h_i^{(l)} = \frac{1}{K_i} \sum_{k=1}^{K_i} \mathbf{z}_{i,k}^{(l)} \tag{1}$$

The ordered sequence $(h_i^{(1)}, \ldots, h_i^{(L)})$ then constitutes the *layer-wise trajectory* of step s_i, encoding how the model's representation of that step transforms as it propagates through successive layers.

Trajectory visualization. Each step is annotated as factually correct or erroneous (annotation protocol detailed in Appendix A.3). We project these layer-wise representations into two dimensions via

PCA and trace each step's trajectory across layers (Figure 1, top). Two patterns emerge clearly: (1) trajectories of factually erroneous steps exhibit markedly higher volatility—particularly in mid-to-late layers—compared to correct steps, and (2) the two classes show separation in the projected space. Crucially, this separation is not attributable to any single layer's activation; rather, it arises from the *sequential pattern* of how representations evolve across the layer stack.

Controlling for step complexity. A natural concern is that the observed volatility merely reflects differences in reasoning complexity rather than factual correctness. To control for this confound, we stratify steps by token count and internal line-break structure as coarse complexity proxies, then repeat the PCA visualization within matched strata (Figure 1, bottom). The mid-to-late-layer volatility gap between correct and erroneous steps persists across all strata, confirming that the geometric divergence is not reducible to step complexity.

Implication. These observations indicate that layer-wise evolution carries an intrinsic, complexity-independent signal for distinguishing grounded from ungrounded reasoning, motivating the detection method we formalize next.

3.2 Prior Error Detection

The posterior analysis reveals that factual errors leave a distinctive geometric footprint in layer-wise trajectories. To quantify trajectory volatility in a form amenable to learned classification, we decompose each step's trajectory into two complementary feature sequences that capture how the representation evolves across layers, then train a lightweight probe on these features.

Angular change sequence. We measure directional reorientation between adjacent layers via the normalized inter-layer angle:

$$\hat{\alpha}_i^{(l)} = \frac{\alpha_i^{(l,\,l+1)}}{\alpha_i^{(l_s,\,l_e)}}, \quad l = l_s, \ldots, l_e{-}1 \tag{2}$$

where

$$\alpha_i^{(a,b)} = \arccos\left(\frac{h_i^{(a)} \cdot h_i^{(b)}}{\|h_i^{(a)}\|_2 \, \|h_i^{(b)}\|_2}\right) \tag{3}$$

This yields $A_i = (\hat{\alpha}_i^{(l_s)}, \ldots, \hat{\alpha}_i^{(l_e-1)}) \in \mathbb{R}^{l_e-l_s}$; larger values indicate more pronounced directional shifts at that layer transition. Normalizing by the endpoint angle $\alpha_i^{(l_s,l_e)}$ removes the influence of a step's overall angular span.

Magnitude change sequence. Complementarily, we quantify how the representation scale evolves across successive layers by measuring the normalized inter-layer displacement:

$$\Delta m_i^{(l)} = \frac{\|h_i^{(l+1)} - h_i^{(l)}\|_2}{\|h_i^{(l_e)} - h_i^{(l_s)}\|_2}, \quad l = l_s, \ldots, l_e{-}1 \tag{4}$$

This yields the sequence $M_i = (\Delta m_i^{(l_s)}, \ldots, \Delta m_i^{(l_e-1)}) \in \mathbb{R}^{l_e-l_s}$, where each element captures the relative expansion or contraction at a single layer transition. Normalizing by the total displacement $\|h_i^{(l_e)} - h_i^{(l_s)}\|_2$ ensures comparability across steps of different overall magnitude.

Feature concatenation and probe. Together, angular and magnitude sequences provide a structured geometric characterization of trajectory volatility—the former tracking "in what direction" it turns, the latter tracking "how far" the representation moves. We concatenate them into a unified feature vector:

$$\mathbf{f}_i = [A_i;\ M_i] \in \mathbb{R}^{2(l_e-l_s)} \tag{5}$$

Here $[l_s, l_e]$ denotes a contiguous layer range restricted to mid-to-late layers, motivated by Section 3.1; its specific boundaries are validated in the ablation study (Section 5.1). An MLP probe $g\colon \mathbb{R}^{2(l_e-l_s)} \to [0,1]$ with two hidden layers (dimensions 128 and 64, ReLU activations) maps the trajectory feature vector to an error probability, trained with binary cross-entropy loss for up to 1,000 iterations using Adam. At inference time, given a new step s_i, we extract its hidden states across $[l_s, l_e]$, compute $\mathbf{f}_i$, and obtain $g(\mathbf{f}_i)$; steps exceeding a threshold are flagged as likely factual errors.

3.3 Detection Experiments

Datasets. We train the probe on 1,000 MedMCQA (Pal et al., 2022) and 173 AfriMedQA (Nimo et al., 2025) questions (70/30 train/val split; labeling protocol and step-level statistics in Appendix A). We evaluate on an in-distribution set of 300 held-out MedMCQA questions and an out-of-distribution set from MMLU-Pro (Biology) (Wang et al., 2024), testing whether the geometric signal

Table 1: Factual error detection performance on Qwen2.5 and Llama3.1. Threshold-independent metrics (AUROC, PR AUC) are reported as primary metrics. Baseline thresholds are selected via Youden's J statistic; probe-based methods use a default threshold of 0.5. Best results per dataset are in **bold**.

	MedMCQA (ID)					MMLU-Pro (Biology) (OOD)				
Method	Prec.	Rec.	F1	AUROC	PR AUC	Prec.	Rec.	F1	AUROC	PR AUC
Backbone: Qwen2.5-7B-Instruct										
PPL	.4433	.6160	.5155	.6112	.4768	.4643	.0138	.0268	.4776	.2200
MaxProb	.4480	.5436	.4912	.6043	.4727	.5417	.0138	.0269	.4838	.2224
Entropy	.4538	.6289	.5272	.6264	.4932	.2727	.0477	.0812	.4695	.2177
CoE-R	.4419	.5436	.4875	.5843	.4204	.2774	.6140	.3822	.5981	.2808
CoE-C	.4571	.7607	.5710	.6560	.4948	.3032	.6394	.4113	.6442	.3630
ReProbe	.4420	**.9325**	.5998	.7350	**.5847**	.3287	.6300	.4320	.6789	.3866
PULSE	**.5027**	.8571	**.6337**	**.7385**	.5666	**.3608**	**.6829**	**.4721**	**.7273**	**.4316**
Backbone: Llama3.1-8B-Instruct										
PPL	.3306	.3015	.3154	.5310	.3185	**.6352**	.2741	.3856	.6090	.5156
MaxProb	.3822	.1854	.2497	.5299	.3141	**.6352**	.2793	.3830	.6093	.5149
Entropy	.3719	.2772	.3176	.5446	.3428	.5634	.2993	.3909	.5921	.4836
CoE-R	.2964	**.9757**	.4546	.5323	.2704	.5683	.5471	.5575	.7204	.6341
CoE-C	.3047	.7697	.4365	.6018	.3552	.5274	.6575	.5853	.7257	.6144
ReProbe	**.4474**	.6211	.5201	.7431	.5103	.6223	.5830	.6021	.7406	.6451
PULSE	.4450	.6816	**.5385**	**.7528**	**.5150**	.6060	**.6844**	**.6428**	**.7933**	**.6982**

transfers without re-training. The resulting label distribution is imbalanced, motivating the use of PR AUC alongside AUROC.

Baselines. We compare against training-free output-level indicators (PPL, MaxProb, Entropy) and hidden-state-based methods (CoE-R/C (Wang et al., 2025), ReProbe (Ni et al., 2025)). Full descriptions are provided in Appendix B.

Implementation details. We conduct experiments on two backbone models: Qwen2.5-7B-Instruct (Yang et al., 2024) and Llama3.1-8B-Instruct (Grattafiori et al., 2024). For each reasoning step, we extract hidden states from layers 7–25 (Qwen) and 6–31 (Llama), and compute the magnitude and angular change sequences defined in Section 3.2. Layer range selection is validated in the ablation study (Section 5.1). We report AUROC, PR AUC, F1, Precision, and Recall as detection metrics.

Results and analysis. Table 1 reports detection performance on both backbones.

On the in-distribution MedMCQA set, PULSE achieves the highest AUROC on both backbones (0.7385 on Qwen, 0.7528 on Llama) and consistently leads in F1. Compared with ReProbe, which concatenates all layer states without modeling their order, the gain confirms that preserving layer-by-layer geometric structure provides additional discriminative signal.

The advantage widens on the out-of-distribution MMLU-Biology set: on Llama, AUROC reaches 0.7933 versus 0.7406 for ReProbe (+5.3 pp) without any re-training, indicating that the learned patterns reflect a general internal signature of factual errors rather than dataset-specific artifacts. Output-level methods (PPL, MaxProb, Entropy) degrade substantially under distribution shift, suggesting that surface-level confidence cannot reliably reveal the model's epistemic state regarding factual grounding.

4 Retrieving on PULSE

Having established that PULSE reliably detects step-level factual errors from trajectory geometry, we now validate its practical utility through an adaptive retrieval framework: if the probe correctly pinpoints where factual errors occur, supplying external evidence at those steps should yield greater accuracy gains with fewer retrievals.

4.1 Error-Triggered Retrieval Framework

Using the probe output as a retrieval trigger, we construct a step-level adaptive retrieval framework. At inference time, the model is prompted to generate chain-of-thought reasoning in which each step begins with a "Step" prefix, yielding well-delimited semantic units. After each step s_i is generated, we

extract its layer-wise trajectory, compute the feature vector $\mathbf{f}_i$, and pass it to the probe g. If $g(\mathbf{f}_i) = 1$ (factual error is detected), the system extracts important tokens from the factually erroneous step and retrieves relevant passages from an external corpus, and then injects them into the context for subsequent reasoning; otherwise, the model continues with its parametric knowledge alone, avoiding unnecessary external evidence.

We explore two knowledge injection strategies: **(i) PULSE$_{\text{front}}$**: retrieved passages are prepended before the question following IRCoT (Trivedi et al., 2023), and the model regenerates the full reasoning chain; **(ii) PULSE$_{\text{step}}$**: retrieved passages are inserted immediately before the flagged step, which is then regenerated while all prior steps are kept intact.

4.2 Retrieval Experiments

Datasets. We employ benchmarks entirely *unseen* during probe training: (1) MedMCQA (Pal et al., 2022), a large-scale medical entrance exam dataset spanning diverse clinical subjects; (2) JAMA Clinical Challenge (Chen et al., 2025), a set of expert-level clinical reasoning questions from the *Journal of the American Medical Association*; and (3) the health and biology subsets of MMLU-Pro (Wang et al., 2024), two challenging multiple-choice benchmarks covering diverse biomedical topics.

Baselines. We compare against methods spanning the retrieval-frequency spectrum: **CoT** (Wei et al., 2022) (no retrieval), **IRCoT** (Trivedi et al., 2023) (retrieve at every step), **DRAGIN** (Su et al., 2024) and **SeaKR** (Yao et al., 2025) (output-signal-triggered adaptive retrieval), and **ReProbe** (Ni et al., 2025) (hidden-state probe as a drop-in trigger replacement within our framework). All methods share the same backbone LLM, retriever, and knowledge corpus. Full descriptions are in Appendix B.

Knowledge corpus. Our evaluation targets scenarios in which institutions maintain curated, high-coverage knowledge bases—such as departmental case repositories or diagnostic guideline collections—where retrieval timing rather than corpus quality is the primary bottleneck. To isolate the effect of triggering precision from retrieval noise, we construct a high-coverage corpus for each benchmark (details in Appendix C.2). A pilot study with

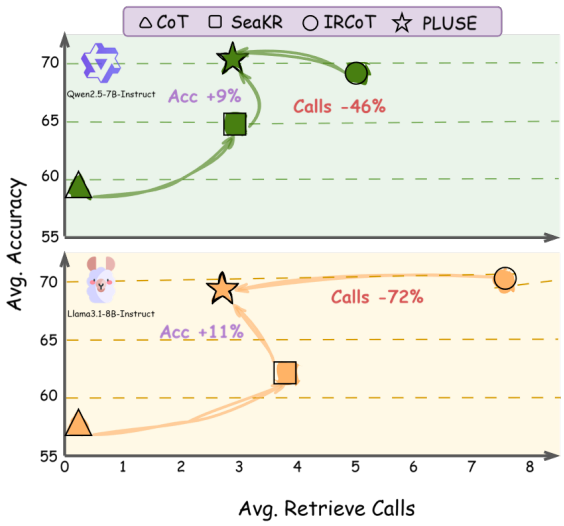

Figure 3: Accuracy–efficiency Pareto front on the four medical QA benchmarks (averaged). **Top**: Qwen2.5-7B-Instruct; **Bottom**: Llama3.1-8B-Instruct. The x-axis reports the mean number of retrieval calls per question and the y-axis reports average accuracy.

the open-source MedRAG Textbook corpus (Appendix C.1) confirms that under insufficient coverage all methods degrade below unassisted CoT, showing that corpus quality and retrieval timing are orthogonal problems; this work addresses the latter.

Evaluation metrics. We report answer accuracy (%) as the primary metric and the average number of retrieval calls per instance (#Ret) to quantify computational efficiency.

Results and analysis. Table 2 presents end-to-end results on four medical QA benchmarks; Figure 3 visualizes the accuracy–efficiency trade-off. We report two deployment options of PULSE: PULSE$_{\text{step}}$ injects evidence at the flagged step and continues reasoning (minimal cost), while PULSE$_{\text{front}}$ prepends evidence to the query and re-reasons from scratch (higher ceiling).

Supplying external evidence at detected steps yields large accuracy gains. Compared with unassisted CoT, PULSE produces substantial improvements across all benchmarks and backbones (e.g., +24.67% on MedMCQA and +8.53% on JAMA with Qwen; +21.67% and +10.42% on MedMCQA and Biology with Llama). Since the only intervention is injecting evidence at probe-flagged steps, these improvements suggest that the flagged steps correspond to genuine factual errors. This interventional evidence provides support for the detection

Table 2: Main results on four medical QA benchmarks using BM25 retrieval. Acc: accuracy (%). #Ret: average number of retrieval calls per test instance. **Bold**: best; underline: second best within each backbone group (excluding CoT for #Ret).

	MedMCQA		JAMA		MMLU-Pro (Health)		MMLU-Pro (Biology)	
Method	Acc (%)	#Ret ↓	Acc (%)	#Ret ↓	Acc (%)	#Ret ↓	Acc (%)	#Ret ↓
Backbone: Qwen2.5-7B-Instruct								
CoT	57.00	0	50.30	0	57.21	0	72.52	0
IRCoT	**85.33**	5.06	53.87	4.38	**65.28**	5.34	72.80	5.51
DRAGIN	58.33	3.52	53.74	5.70	58.19	4.16	73.78	2.76
SeaKR	79.00	2.43	54.91	4.06	58.07	2.83	63.88	2.56
ReProbe	77.33	2.41	56.19	3.13	62.59	2.90	72.66	2.35
PULSE$_{step}$	77.00	**2.05**	58.57	**2.72**	65.16	**2.52**	74.90	**1.93**
PULSE$_{front}$	81.67	2.75	**58.83**	3.00	63.08	2.61	**76.43**	2.62
Backbone: Llama 3.1 8B Instruct								
CoT	62.00	0	53.47	0	53.79	0	60.15	0
IRCoT	**87.67**	6.96	**58.31**	6.76	62.22	8.60	**71.69**	8.29
DRAGIN	63.67	3.89	51.49	2.80	54.77	3.00	62.76	2.42
SeaKR	71.67	3.57	55.72	4.94	58.01	4.04	62.34	2.66
ReProbe	80.33	2.62	55.66	2.52	56.23	2.46	67.09	2.98
PULSE$_{step}$	82.33	2.28	57.84	2.01	60.15	2.39	68.62	**1.78**
PULSE$_{front}$	83.67	**2.05**	57.78	**1.86**	**63.50**	**2.36**	70.57	2.34

signal beyond correlation-based metrics in Table 1.

More precise targeting achieves greater gains with fewer retrievals. DRAGIN and SeaKR, despite moderate retrieval frequency, often fail to improve—and sometimes degrade—accuracy relative to CoT (e.g., DRAGIN drops to 51.49% on JAMA with Llama versus 53.47% for CoT), suggesting that imprecise triggering introduces noise rather than useful knowledge. Under the same step-level injection protocol, PULSE$_{step}$ consistently outperforms ReProbe with fewer calls: on Qwen, +2.38% on JAMA with 13% fewer calls, +2.57% on Health with 13% fewer calls, and +2.24% on Biology with 18% fewer calls; similar patterns hold on Llama (+2.18% on JAMA with 20% fewer calls, +1.53% on Biology with 40% fewer calls). This simultaneous improvement on both axes confirms that trajectory geometry provides a more precise error signal than concatenated hidden states.

Targeted retrieval matches exhaustive retrieval at a fraction of the cost. IRCoT retrieves at every step (5–9 calls per instance). PULSE$_{front}$ surpasses it on JAMA (+4.96%) and MMLU-Pro (Biology) (+3.63%) with Qwen using only 2.6–3.0 calls, and on MMLU-Pro (Health) (+1.28%) with Llama at 2.36 versus 8.60 calls. On the remaining benchmarks it stays within 2–4% at roughly one-third the budget. Retrieving for the right steps matters more than retrieving for all steps.

All trends hold across both backbones, providing strong empirical evidence that PULSE identifies genuine factual errors with precision that translates into a favorable accuracy–efficiency trade-off.

5 Analysis

This section addresses two questions: whether the probe's design choices (feature type and layer interval) are well-justified (Section 5.1), and whether the detected signal generalizes beyond the training domain (Section 5.2).

5.1 Ablation

Feature ablation. Table 3 compares detection and downstream performance under different feature configurations on MMLU-Pro (Biology), with the probe trained on MedMCQA (out-of-distribution). Using magnitude alone yields the lowest AUROC (0.6565) and under-triggers retrieval (2.51 calls), leaving many factual errors undetected. Using angle alone improves detection (AUROC 0.6866) but over-triggers (2.94 calls), introducing unnecessary retrievals that may distract the reasoning process. Combining both features achieves the best AUROC (0.7273) and downstream accuracy (76.43%) at moderate retrieval cost (2.62 calls), confirming that the two

Table 3: Feature ablation on MMLU-Pro (Biology) with Qwen2.5-7B-Instruct. Probe is trained on MedMCQA and evaluated out-of-distribution on MMLU-Pro (Biology).

	Detection		End-to-End	
Feature	AUROC	PR AUC	Acc(%)	Avg #Ret
Mag	.6565	.3547	75.31	2.51
Angle	.6866	.3903	75.59	2.94
Mag + Ang	**.7273**	**.4316**	**76.43**	2.62

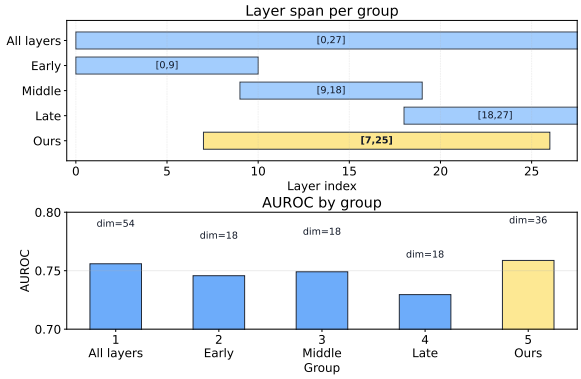

Figure 4: Effect of layer interval on detection AUROC (Qwen2.5-7B-Instruct, MedMCQA).

dimensions capture complementary aspects of the model's internal state and are both necessary for reliable factual error detection.

Layer interval selection. Figure 4 examines how the choice of layer interval affects detection. The late-only interval ([18,27]) performs worst (AUROC 0.7295): when restricted to final layers, the representations increasingly commit to the output distribution and lose discriminative signal about factual errors. The early interval ([0,9]) also underperforms (AUROC 0.7457), as shallow layers primarily encode token-level syntax before higher-level reasoning representations have formed. The middle interval ([9,18]) reaches 0.7491, confirming that intermediate layers carry the strongest per-segment signal.

Our selected interval [7,25] achieves the highest AUROC (0.7588) and F1 (0.5474), outperforming even the all-layer configuration (AUROC 0.7559, F1 0.5378) despite using only 36 of 54 features. This indicates that early and late layers contribute domain-specific token-distribution noise that harms both discrimination and generalization; excluding them yields a more robust and compact representation. We therefore adopt [7,25] as the default interval. Llama 3.1 exhibits the same pattern; we adopt layers [6,31] following the same selection logic. Since retrieval is triggered by the probe's binary output, this choice directly determines when evidence is supplied during reasoning.

5.2 Cross-Domain Generalization

To test whether the geometric signal is domain-agnostic, we directly apply the probe trained exclusively on medical data (MedMCQA + AfriMedQA) to the MMLU-Pro Law subset (1,101 instances) without any adaptation. As shown in Table 4, PULSE achieves 42.23% accuracy with 2.36 retrieval calls per instance, outperforming both the CoT baseline (28.52%) and IRCoT (39.96% at 4.48 calls). The probe attains higher accuracy than exhaustive retrieval while requiring roughly half the cost, indicating that triggering decisions remain precise in an entirely unseen domain. Combined with the layer-selection analysis above, these results support the interpretation that middle-layer trajectory fluctuations reflect a general computational signature of failed internal knowledge retrieval, independent of the specific domain of the missing knowledge.

Table 4: Cross-domain transfer to MMLU-Pro (Law) with Qwen2.5-7B-Instruct. The probe is trained solely on medical data and applied without adaptation.

Method	Acc (%)	Avg #Ret
CoT	28.52	0
IRCoT	39.96	4.48
PULSE	**42.23**	**2.36**

6 Conclusion

We have shown that factually erroneous reasoning steps exhibit a distinctive geometric signature in middle-layer hidden states—a latent *pulse* encoded in its forward pass. By monitoring this signal at the step level, PULSE grounds retrieval timing in the model's computational trajectory, matching or surpassing exhaustive per-step retrieval with only 28–54% of the calls across four medical benchmarks and transferring zero-shot to legal reasoning. Just as a clinician takes a patient's pulse to decide when to intervene, a reasoning system can take the pulse of its hidden dynamics to decide when external knowledge is needed. Future work may extend trajectory-based error signals as process-level rewards in agentic settings and verify whether the signature persists at larger model scales.

Limitations

Model access. PULSE requires access to intermediate hidden states, limiting its applicability to open-weight models. It cannot be directly deployed on black-box APIs.

Model scale. Our experiments use 7–8B parameter models. Whether the geometric signature persists—and whether a single probe transfers—at 70B+ scales remains unverified.

Step structure. The method assumes the model produces explicit step-by-step reasoning with clear boundaries (e.g., special token-delimited steps). For free-form or single-paragraph generations, more flexible logical segmentation strategies—such as semantic chunking or clause-level boundary detection—may be needed to define the units over which trajectory features are computed.

Corpus Dependency. All retrieval-augmented methods in medical reasoning scenarios are contingent on the underlying corpus providing sufficient step-level coverage for the target domain. PULSE is therefore best suited to deployment settings where curated, high-coverage knowledge bases are available (e.g., institutional case repositories in clinical domains), rather than open-ended web-scale retrieval scenarios.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, volume 2024, pages 9112–9141.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Farima Fatahi Bayat, Kun Qian, Benjamin Han, Yisi Sang, Anton Belyy, Samira Khorshidi, Fei Wu, Ihab Ilyas, and Yunyao Li. 2023. Fleek: Factual error detection and correction with evidence retrieved from external knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 124–130.

Armin Berger, Sarthak Khanna, David Berghaus, and Rafet Sifa. 2025. Reasoning llms in the medical domain: A literature survey. *2025 IEEE 12th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10.

Chi Seng Cheang, Hou Pong Chan, Wenxuan Zhang, and Yang Deng. 2025. Large language models do not really know what they don't know. *arXiv preprint arXiv:2510.09033*.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2025. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3563–3599, Albuquerque, New Mexico. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yizheng Huang and Jimmy X. Huang. 2024. A survey on retrieval-augmented text generation for large language models. *ACM Computing Surveys*.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 7969–7992.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Andrew Lee, Lihao Sun, Chris Wendler, Fernanda Viégas, and Martin Wattenberg. 2025. The geometry of self-verification in a task-specific reasoning model. In *Mechanistic Interpretability Workshop at NeurIPS 2025*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Linyu Liu, Yu Pan, Xiaocheng Li, and Guanting Chen. 2024. Uncertainty estimation and quantification for llms: A simple supervised approach. *arXiv preprint arXiv:2404.15993*.

Ming Lu, Chaozhuo Li, Pengbo Wang, Chenxu Wang, Litian Zhang, Zheng Liu, Qiwei Ye, Yi Hua, Yushan

Cai, Yuanbo Xu, Hao Zhao, Feiran Huang, Xi Zhang, and Philip S. Yu. 2026. Loki's dance of illusions: A comprehensive survey of hallucination in large language models. *IEEE Transactions on Computational Social Systems*, pages 1–32.

Yuxing Lu, Gecheng Fu, Wei Wu, Xukai Zhao, Sin Yee Goi, and Jinzhuo Wang. 2025. Doctorrag: Medical rag fusing knowledge with patient analogy through textual gradients. *ArXiv*, abs/2505.19538.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems 35*.

Jingwei Ni, Ekaterina Fadeeva, Tianyi Wu, Mubashara Akhtar, Jiaheng Zhang, Elliott Ash, Markus Leippold, Timothy Baldwin, See Kiong Ng, Artem Shelmanov, and Mrinmaya Sachan. 2025. Reprobe: Efficient test-time scaling of multi-step reasoning by probing internal states of large language models.

Charles Nimo, Tobi Olatunji, Abraham Toluwase Owodunni, Tassallah Abdullahi, Emmanuel Ayodele, Mardhiyah Sanni, Ezinwanne C. Aka, Folafunmi Omofoye, Foutse Yuehgoh, Timothy Faniran, Bonaventure F. P. Dossou, Moshood O. Yekini, Jonas Kemp, Katherine A Heller, Jude Chidubem Omeke, Chidi Asuzu Md, Naome A Etori, Aïmérou Ndiaye, Ifeoma Okoh, and 7 others. 2025. AfriMed-QA: A pan-African, multi-specialty, medical question-answering benchmark dataset. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1948–1973, Vienna, Austria. Association for Computational Linguistics.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. 2022. Prompting gpt-3 to be reliable. *arXiv preprint arXiv:2210.09150*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024. DRAGIN: Dynamic retrieval augmented generation based on the real-time information needs of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12991–13013, Bangkok, Thailand. Association for Computational Linguistics.

Xiangru Tang, Wanghan Xu, Yujie Wang, Zijie Guo, Yanjun Shao, Jiapeng Chen, Cixuan Zhang, Ziyi Wang, Lixin Zhang, Guancheng Wan, Wenlong Zhang, Lei Bai, Zhenfei Yin, Philip Torr, Hanrui Wang, and Di Jin. 2025. Eigen-1: Adaptive multi-agent refinement with monitor-based rag for scientific reasoning. *ArXiv*, abs/2509.21193.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, Elahe Vedadi, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Le Hou, Albert Webson, Kavita Kulkarni, S. Mahdavi, Christopher Semturs, and Vivek Natarajan. 2025. Towards conversational diagnostic artificial intelligence. *Nature*, 642:442–450.

Yiming Wang, Pei Zhang, Baosong Yang, Derek Wong, and Rui Wang. 2025. Latent space chain-of-embedding enables output-free llm self-evaluation. In *International Conference on Learning Representations*, volume 2025, pages 70938–70970.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251, Bangkok, Thailand. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. Publisher Copyright: © 2023 11th International Conference on Learning Representations, ICLR 2023. All rights reserved.; 11th International Conference on Learning Representations, ICLR 2023 ; Conference date: 01-05-2023 Through 05-05-2023.

Zijun Yao, Weijian Qi, Liangming Pan, Shulin Cao, Linmei Hu, Liu Weichuan, Lei Hou, and Juanzi Li. 2025. SeaKR: Self-aware knowledge retrieval for adaptive retrieval augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27022–27043, Vienna, Austria. Association for Computational Linguistics.

A Dataset and Annotation Details

A.1 Data Sources and Splits

Our experimental pipeline involves three categories of data: (1) probe training data, used to train the factual error detection probe; (2) probe evaluation data, used to assess detection performance under both in-distribution (ID) and out-of-distribution (OOD) settings; and (3) end-to-end evaluation data, used to measure downstream QA accuracy when the probe is integrated with retrieval-augmented generation.

For probe training, we combine MedMCQA (Pal et al., 2022) and AfriMedQA (Nimo et al., 2025) to ensure diversity in question difficulty and cultural context. For probe evaluation, we hold out a separate MedMCQA subset as the ID test and use MMLU-Pro-Biology (Wang et al., 2024) as the OOD test, which differs in domain granularity and question style. Table 5 summarizes question-level statistics, and Table 6 provides step-level statistics for the probe training and evaluation splits.

Table 5: Question-level statistics across all datasets.

Role	Questions	Avg Steps	Domain
Train/Valid	1173	3.7	Medicine
Test (ID)	300	4.9	Medicine
Test (OOD)	717	5.9	Biology

Table 6: Step-level statistics for probe training and evaluation. Error Ratio denotes the proportion of steps labeled as containing a factual error.

Split	Total	No-Error	Error	Error Ratio
Train/Valid	4348	3216	1132	26.0%
Test (ID)	1483	944	539	36.4%
Test (OOD)	4257	3314	943	22.1%

A.2 Step Segmentation Protocol

To obtain discrete reasoning steps from chain-of-thought outputs, we instruct the backbone model to structure its response using explicit step prefixes. The generation prompt includes the following instruction:

```
Think step by step and output the reasoning process before answering. Each step should be started with "Step".
```

We then segment the generated output using rule-based matching: each segment beginning with the pattern `"Step [N]:"` is extracted as an individual reasoning step. Steps that do not conform to this format (e.g., malformed outputs or steps lacking the required prefix) are discarded. In practice, the vast majority of generated outputs conform to the prescribed format across all datasets.

Figure 5 provides a complete example of a segmented reasoning chain along with its step-level annotations.

A.3 Factual Error Annotation Protocol

Definition. A reasoning step is labeled as a *factual error* if it contains a factual claim that is incorrect or unsupported by established domain knowledge, indicating that the model lacks the requisite factual knowledge to produce a correct statement. Each step is evaluated **independently**: a step is labeled as factually erroneous if it contains an error regardless of whether the error originates in that step or is inherited from a preceding step. We adopt this step-independent criterion because (1) it is more reproducible than causal error attribution, and (2) the probe operates at inference time without access to ground-truth labels of prior steps, so it must detect knowledge deficiency from local representations alone.

Automated Annotation with GPT. We employ GPT to generate step-level annotations at scale. The annotator model receives the question, the correct answer, and the model's full reasoning chain, and is instructed to evaluate each step independently. The complete prompt template is shown below:

Using this protocol, we annotated 4,348 steps in the training set and 5,740 steps across the two probe evaluation sets.

Human Verification. To assess the reliability of the automated annotations, we conducted a manual quality review of the GPT-generated step labels. For all questions where the backbone model produced an incorrect final answer, the step labels were reviewed against established medical and biological references, and any identified labeling errors

Annotated Example

Question: Characteristic of venous blood flow of lower limb in duplex Doppler is?
Options: (A) Monophasic (B) Biphasic (C) Triphasic (D) Non phasic
Ground Truth: A (Monophasic) **Model Answer:** C (Triphasic)

Step 1: Understanding the context of the question, we need to consider the nature of venous blood flow in the lower limbs and how it is visualized using duplex Doppler ultrasound.
`Label: No Error`

Step 2: Duplex Doppler ultrasound is a technique that combines the Doppler effect with B-mode imaging to visualize blood flow and its direction.
`Label: No Error`

Step 3: In the lower limbs, venous blood flow typically shows a characteristic pattern due to the presence of valves that prevent backflow.
`Label: No Error`

Step 4: When blood flows through the veins, it can be influenced by factors such as muscle contraction, gravity, and the presence of valves. This results in a flow pattern that can be categorized into phases.
`Label: No Error`

Step 5: The most common pattern observed in the lower limb veins using duplex Doppler is a triphasic flow, which includes a systolic phase, a diastolic phase, and a post-contraction phase.
`Label: Factual Error`
Reason: Incorrectly identifies lower limb venous flow as triphasic. Venous flow is monophasic, modulated by respiration and muscular compression. Triphasic flow is characteristic of peripheral arterial waveforms.

Step 6: This triphasic pattern is due to the sequential opening and closing of the venous valves, which allows blood to flow in one direction despite changes in pressure.
`Label: Factual Error`
Reason: Fabricates a nonexistent physiological mechanism; sequential valve opening/closing does not produce triphasic flow in veins.

Step 7: Considering the options provided, the correct answer is the one that matches the triphasic pattern observed in the lower limb veins.
`Label: Factual Error`
Reason: Asserts that triphasic flow is observed in lower limb veins, which is factually incorrect.

Figure 5: A complete annotated example showing the question, segmented reasoning steps, and step-level factual error labels with justifications.

were corrected. For questions where the model produced a correct final answer, we additionally reviewed a random sample of 100 questions, comprising approximately 400 reasoning steps. This audit found a step-level labeling accuracy above 95%. Based on this verification, we consider the automated annotations sufficiently reliable for probe training in our benchmark setting.

The review focused on reference-supported factual consistency rather than clinical decision-making. The annotation prompt also explicitly instructs the annotator model to evaluate each step independently and to avoid marking a step as incorrect solely due to error propagation. This matches our annotation definition and contributes to the high observed accuracy after manual verification.

A.4 Class Distribution and Training Considerations

As shown in Table 6, the training set exhibits a class imbalance with an error ratio of 26.0% (approximately 1:2.8 positive-to-negative ratio). To mitigate the effect of this imbalance on probe training, we apply a positive class weight of 3 in the binary cross-entropy loss, which approximately matches the inverse class frequency ratio. This weighting ensures that the probe does not trivially learn to predict the majority class. We found results to be robust across weight values in the range [2, 5].

The error ratio varies across evaluation sets: the

Annotation Prompt Template

```
You are a teacher specializing in identifying errors in students' problem-solving processes.
You will be provided with a question, the standard answer, and a student's response containing
reasoning steps numbered by "Step". Please provide your evaluation entirely in English and
strictly follow the structure below:
1. Step Evaluation
Evaluate each "Step" in the student's response independently. For each step, use the following
fixed format:
Step [number]: [Correct / Incorrect]
Reason: [For an incorrect step, briefly explain the reason. The evaluation should be based only
on whether the step itself contains a knowledge error; mark a step as Incorrect only if the step,
when judged independently, contains a knowledge error. Do not mark a step as Incorrect merely
because of an error in a previous step, an error in a later step, an incorrect final answer, or
inconsistency with the standard answer.]
2. Final Label
[Output a Python dictionary-formatted string, where the keys are step numbers and the values
are correctness labels.]
{'Step 1': 'Correct', 'Step 2': 'Incorrect', ...}
```

Figure 6: Annotation Prompt Template

ID test shows a higher error ratio (36.4%) because it is drawn from questions the model frequently answers incorrectly, while the OOD test has a lower error ratio (22.1%) reflecting a different difficulty distribution. This variation allows us to assess probe robustness under different prevalence conditions, which is why we report both AUROC (threshold-invariant) and PR AUC (sensitive to class prevalence) as evaluation metrics.

B Baseline Details

B.1 Baselines for Factual Error Detection (Section 3.3)

Perplexity (PPL) (Si et al., 2022). For each reasoning step, we compute the average token-level negative log-likelihood (i.e., surprisal) under the backbone model. A higher perplexity indicates that the model assigns lower probability to its own generation, which may signal factual uncertainty. We treat perplexity as a step-level anomaly score and select the detection threshold via Youden's J statistic on the validation set.

MaxProb. For each reasoning step, we record the maximum softmax probability across all generated tokens and average these values as the step-level confidence score. Lower average maximum probability suggests the model is less committed to its token choices, potentially indicating erroneous content. The decision threshold is again selected via Youden's J statistic.

Entropy (Kuhn et al., 2023). We compute the Shannon entropy of the next-token probability distribution at each decoding position, then average over all tokens in the step and normalize by sequence length. Higher entropy reflects greater distributional uncertainty. The threshold is selected identically to the other output-level methods.

CoE-R / CoE-C (Wang et al., 2025). Both methods compute the magnitude change and angular change between adjacent hidden states across all layers for a reasoning step. CoE-R averages the difference between local magnitude and angular ratios into a single scalar; CoE-C treats each layer transition as a complex number and takes the magnitude of their average resultant vector. Both are training-free and compress the full layer-wise progression into one score per step.

ReProbe (Ni et al., 2025). ReProbe concatenates the hidden-state vectors from all layers at the final token position of a reasoning step, forming a high-dimensional representation of the step's full internal state. This concatenated vector is then projected and classified by a lightweight Transformer encoder trained with binary cross-entropy. Unlike our method, ReProbe does not explicitly model the sequential, layer-by-layer structure of the representation trajectory.

B.2 Baselines for Retrieval Experiments (Section 4.2)

CoT (Wei et al., 2022). Standard chain-of-thought prompting without any retrieval, representing the backbone model's intrinsic reasoning capability.

IRCoT (Trivedi et al., 2023). At every reasoning step, the system issues a retrieval call, prepends the top-k passages to the query, and continues generation. This represents an indiscriminate full-retrieval strategy and serves as an upper bound on retrieval frequency.

DRAGIN (Su et al., 2024). Retrieval is triggered when the maximum token-level entropy within a reasoning step exceeds a calibrated threshold. The retrieval query is constructed from the tokens with the highest attention-weighted information gain in the flagged step.

SeaKR (Yao et al., 2025). The system samples multiple continuations for the current step and estimates the model's internal uncertainty via hidden-state covariance across samples. Retrieval is invoked when this covariance exceeds a threshold, indicating that the model lacks sufficient parametric support. A self-refinement module then selects between the original and the retrieval-augmented continuation.

ReProbe (Ni et al., 2025). A lightweight Transformer probe is trained on the concatenation of hidden states across all layers at the final token of each step. We integrate its binary classification output into our retrieval framework as a drop-in trigger replacement: retrieval is called whenever ReProbe flags a step as factually erroneous, using the same injection strategy and corpus as PULSE.

C Retrieval Corpus Design

This appendix describes our retrieval corpus design decisions. We first present a pilot study using a realistic open-domain medical corpus (Section C.1), which motivates the construction of a high-coverage corpus that isolates detection quality from retrieval noise (Section C.2). We then detail the retrieval configuration used in our main experiments (Section C.3).

C.1 Realistic Corpus Pilot Study

To assess the feasibility of factual-error-triggered retrieval under realistic conditions, we conducted a pilot experiment using the open-source Textbook corpus from MedRAG (Xiong et al., 2024), which contains approximately 160K passages drawn from medical textbooks. We evaluated on a random sample of 500 questions from JAMA using BM25 retrieval with top-3 candidates, selecting the first non-duplicate passage at each retrieval step.

Table 7 reports the results. All retrieval-augmented methods—including IRCoT, SeaKR, and PULSE—performed below the CoT baseline without retrieval. Notably, the degree of degradation correlates with retrieval frequency: SeaKR and IRCoT, which retrieve more aggressively (3.43 and 3.85 retrievals per question on average), suffer the largest accuracy drops, while PULSE retrieves more conservatively (2.25 per question) and degrades less. This pattern suggests that under a noisy corpus, each retrieval event has a net-negative expected value—the injected passage is more likely to distract than to help—and methods that retrieve more frequently accumulate more harm.

Table 7: Accuracy (%) on JAMA using the realistic MedRAG Textbook corpus (n=500). All retrieval-augmented methods degrade below the no-retrieval CoT baseline. Retrieval frequency correlates with performance loss.

Method	Acc (%)	Avg. Ret
CoT (no retrieval)	51.40	0
PULSE (Ours)	47.60	2.25
IRCoT	45.40	3.85
SeaKR	42.40	3.43

We attribute this failure to two compounding factors: (1) *insufficient coverage*—the general-purpose textbook corpus does not reliably contain the fine-grained factual knowledge matching each specific reasoning step, and (2) *retrieval noise*—passages that are topically related but factually irrelevant introduce distractors that mislead the model. These observations motivate the high-coverage corpus design described below, which enables controlled evaluation of factual error detection independent of retrieval quality, as discussed in Section 4.2 of the main text.

C.2 Corpus Construction

To isolate the contribution of error detection from the confound of retrieval quality, we construct a per-benchmark corpus of factual knowledge statements using GPT-5.1. The design goal is to ensure that when retrieval is triggered at a step with a genuine factual error, the corpus contains a relevant supporting fact.

Generation procedure. For each question in the dataset, we provide GPT-5.1 with the question, the gold answer, and the model-generated chain-of-thought reasoning. The model is prompted to eval-

uate each reasoning step and produce the factual knowledge required to support that step. The full prompt template is shown below:

Each step yields 1–4 knowledge statements depending on the complexity of the reasoning involved.

Filtering and quality control. We apply a two-stage filtering process. First, rule-based filtering removes statements that contain calculation processes, problem-specific reasoning chains, or direct answers (identified via keyword patterns such as "therefore the answer is", "substituting into the equation", and similar phrases). Second, a single human annotator reviews the filtered statements to verify that each entry constitutes an objective factual claim (e.g., a definition, theorem, biological fact, or clinical criterion) rather than a reasoning derivation. Duplicate statements within each benchmark corpus are removed.

Corpus statistics. Table 8 summarizes the resulting corpora. Each knowledge statement is 1–3 sentences in length.

Table 8: Corpus statistics (number of unique knowledge statements per benchmark).

Benchmark	# Statements
JAMA	20,225
MMLUPro_Health	4,522
MMLUPro_Biology	4,262
MedMCQA (test, 300 questions)	1,503

C.3 Retrieval Configuration

We use BM25 for retrieval over the high-coverage corpus. At each reasoning step where retrieval is triggered, we construct the query by extracting content words from the current step text, retaining only tokens with part-of-speech tags in {NOUN, PROPN, NUM, ADJ, VERB, ADV}. This filtering removes function words and improves retrieval precision for factual content.

We retrieve the top-3 candidate passages and select the first passage that has not already been used in a previous retrieval step for the same question, ensuring that each triggered retrieval injects novel information. The format for integrating retrieved knowledge into the prompt is illustrated in Appendix D.

Design rationale. By pairing a high-coverage corpus with a simple retrieval mechanism, we ensure that performance differences across detection methods (reported in Section 4.2 of the main text) reflect the quality of factual error detection rather than artifacts of retrieval failure. The pilot study in Section C.1 demonstrates that this controlled setup is necessary: under realistic retrieval conditions, all methods degrade equally, making it impossible to distinguish detection quality.

C.4 Computational Resources

PULSE introduces only a small trainable component. The step-level error probe is implemented as a two-layer MLP with hidden dimensions 128 and 64, and takes as input low-dimensional magnitude and angular features rather than full hidden states. As a result, the probe contains only on the order of tens of thousands of trainable parameters, which is negligible compared with the frozen backbone LLM. The backbone model is not fine-tuned.

The main computational and memory overhead comes from inference-time feature extraction. To compute the magnitude and angular trajectory features, PULSE retains step-level hidden states during generation. This memory cost scales with the number of generated reasoning steps, the number of recorded layers, and the hidden dimension of the backbone model. Adaptive retrieval introduces additional latency only for steps selected by the probe. Experiments were conducted on NVIDIA A100 GPUs (80 GB).

D Case Study

Analysis

Both methods detect a factual error early in the reasoning process and enter the iterative retrieval loop. PULSE detects the error at Step 2, where the model recognizes that the OCT morphology is atypical but fails to connect the spiculated epiretinal membrane pattern to NF2. A single targeted retrieval provides this missing diagnostic knowledge: epiretinal membranes in NF2 often have a spiculated appearance and project anteriorly into the vitreous. With this diagnostic-level evidence, the regenerated step correctly identifies NF2 as the likely underlying condition, and the model proceeds to select family history and genetic testing as the appropriate next step before treatment decisions are made.

ReProbe, in contrast, triggers retrieval earlier at Step 1 and retrieves management-oriented evidence about observation versus early vitrectomy

Corpus Construction Prompt Template

```
You are a teacher specializing in identifying student errors in problem-solving. You are provided
with a problem, a standard answer, and a student's response (including thought processes numbered
"Step"). Please strictly adhere to the following structure when using the output:
Step Evaluation
Please evaluate each "Step" in the student's response independently. For each step, please use
the following fixed format:
Step [Number]: [Correct / Incorrect]
Reason: [Briefly explain the reason. Important Principle: Evaluation should be based on whether
the logical derivation of this step conforms to subject-specific standards. Even if a previous
step is incorrect, as long as the current step's logical derivation based on the previous result
is correct, it should still be judged as Correct (unless the step itself introduces new incorrect
knowledge).]
Knowledge: [Key Fields: Please provide the multiple original factual knowledge points required to
support this step of the reasoning. Content Requirements: Only definitions, theorems, formulas,
constants, subject-specific knowledge, or objective facts are allowed. Prohibited Content:
Specific calculation processes, reasoning results, or final conclusions related to this problem
are strictly prohibited. Do not mention the problem or reveal the answer.]
```

Figure 7: Prompt template for corpus construction.

Case-study question from JAMA Clinical Challenge.

Question

An 8-year-old girl was referred to the vitreoretinal service for evaluation of bilateral macular lesions noted by her pediatric ophthalmologist. Bilateral lateral rectus muscle recession for exotropia was performed 4 years before this presentation. Medical history was otherwise unremarkable. [...] A single, 1-cm, oval, lightly pigmented macule with irregular borders was noted on her right upper arm. Dilated fundus examination revealed eccentric, yellow-clear, preretinal lesions in the maculas of both eyes. OCT revealed hyperreflective areas on the surface of the macula with a spiculated appearance, projecting anteriorly into the vitreous.

Options: A: Obtain a family medical history and genetic testing B: Recommend pars plana vitrectomy with epiretinal membrane peel C: Observe the lesions D: Give an intravitreal ocriplasmin injection

Figure 8: Case-study question from JAMA Clinical Challenge.

for NF2-related epiretinal membranes. Although this evidence is factually relevant, it is premature for the question, whose correct next step is diagnostic confirmation rather than immediate treatment. ReProbe later triggers a second retrieval at Step 4, which injects additional procedural knowledge about pars plana vitrectomy and epiretinal membrane peeling. This second retrieval further primes the model toward a treatment-oriented conclusion. As a result, even though the model mentions genetic testing in Step 5, it ultimately prioritizes surgical intervention in Step 6 and selects the incorrect answer.

This case highlights that retrieval timing can affect downstream reasoning, even when the retrieved evidence is factually relevant. Evidence about management or surgical procedures may be helpful in other contexts but in this question it is introduced before the diagnosis has been established, shifting the reasoning trajectory from diagnostic confirmation toward treatment selection. In this example, PULSE triggers retrieval at a step more directly tied to the missing diagnostic fact, and therefore resolves the key factual gap with fewer retrieval calls.

E Ethical Considerations and Responsible Use

E.1 Risks and Safety

PULSE is intended as a research framework for detecting potential step-level factual errors in LLM-generated medical reasoning and for studying adaptive retrieval. It is not a clinical decision-support system, has not been evaluated in real clinical workflows, and should not be used as a substitute for professional medical judgment.

Potential risks include both false negatives and false positives. A false negative may fail to detect a

factual error and thereby allow an incorrect medical claim to influence downstream reasoning. A false positive may trigger unnecessary retrieval, introduce distracting evidence, or shift the model toward an incorrect conclusion. In addition, retrieved passages may be incomplete, outdated, irrelevant, or inconsistent with current clinical guidelines, especially when retrieval is performed over open-world biomedical or clinical corpora.

There is also a risk of over-reliance. Users may incorrectly interpret a low PULSE error score or a successful retrieval intervention as a guarantee of medical correctness. However, PULSE detects signals associated with potential factual errors in generated reasoning steps; it does not verify clinical appropriateness, patient-specific applicability, treatment safety, or guideline compliance. Therefore, PULSE should be used only as an analysis and research tool unless further validated in clinically realistic settings.

To mitigate these risks, we evaluate PULSE only in benchmark medical QA settings and do not make claims about clinical deployment. We explicitly state that PULSE outputs should not be interpreted as medical advice. Any future use in clinical or patient-facing environments would require rigorous prospective validation, careful monitoring, high-quality and up-to-date retrieval sources, and oversight by qualified medical professionals.

E.2 Data Privacy

Our experiments use publicly available medical QA benchmarks and biomedical text corpora. We do not collect new patient data, access electronic health records, or process identifiable protected health information. The study is conducted in a benchmark evaluation setting and does not involve interaction with patients or human subjects.

E.3 Use of AI Assistants

We use large language models to support step-level factuality annotation and/or data quality filtering. The prompts, label schema, and filtering procedures are described in Appendix A.3. Automatically produced labels may contain errors and should not be interpreted as definitive medical judgments. We therefore use them as noisy supervision and evaluation aids for research purposes rather than as clinical ground truth.

PULSE – Case Study (JAMA Clinical Challenge)

Step 1 (No Error Detected)

Analyze the patient's clinical presentation and findings. The patient is an 8-year-old girl with bilateral macular lesions. She has a history of exotropia surgery. Visual acuity is stable at 20/40 in both eyes. There is a residual intermittent exotropia. A single, 1-cm, oval, lightly pigmented macule with irregular borders is noted on her right upper arm. Dilated fundus examination reveals eccentric, yellow-clear, preretinal lesions in the maculas of both eyes. OCT shows hyperreflective areas on the surface of the macula with a spiculated appearance, projecting anteriorly into the vitreous, and increased retinal thickness with loss of the foveal contour.

Step 2 Draft (Factual Error Detected by Probe)

Consider the differential diagnosis. The bilateral epiretinal membranes with a spiculated appearance and anterior vitreous projection on OCT are atypical for standard idiopathic epiretinal membranes. The pigmented macule on the right upper arm may represent a café-au-lait spot, which can be associated with phakomatoses such as Neurofibromatosis Type 1 (NF1). However, the specific significance of the spiculated OCT morphology with anterior projection in a pediatric patient is *unclear*. Other possibilities include combined hamartoma of the retina and retinal pigment epithelium, or vitreomacular traction syndrome. Further clinical evaluation and imaging may be needed to narrow down the differential diagnosis.

Retrieved Knowledge (1 retrieval triggered)

Epiretinal membranes in NF2 often have a spiculated appearance with curled edges, projecting anteriorly into the vitreous and are best visualized using optical coherence tomography.

Step 2 Regenerated (After Retrieval)

Consider the differential diagnosis. The clinical presentation and OCT findings are consistent with Neurofibromatosis Type 2 (NF2). NF2 is characterized by bilateral vestibular schwannomas and can present with retinal abnormalities, including epiretinal membranes. The cutaneous pigmented macule may raise consideration of an underlying phakomatosis, while the spiculated epiretinal membrane morphology is more directly suggestive of NF2-related ocular findings.

Step 3 (No Error Detected)

Determine the next steps. Given the clinical suspicion of NF2, the appropriate next step is to obtain a detailed family medical history and genetic testing to confirm the diagnosis before considering treatment options.

Final Answer

A: Obtain a family medical history and genetic testing Correct

Figure 9: Case study of PULSE on a JAMA Clinical Challenge question.

ReProbe – Case Study (Same Question)

Step 1 Draft (Factual Error Detected by ReProbe Trigger)

The patient is an 8-year-old girl with bilateral macular lesions and a history of bilateral lateral rectus muscle recession for exotropia. On examination, OCT reveals atypical epiretinal membranes with a spiculated appearance projecting anteriorly into the vitreous. A single pigmented macule with irregular borders is noted on the right upper arm. The combination of ocular findings and a cutaneous lesion may suggest an underlying systemic or genetic condition, but the specific clinical significance of the spiculated epiretinal membrane morphology is *not immediately clear* from the presentation alone.

Retrieved Knowledge

Observation is an option for NF2-related epiretinal membranes if vision loss is mild, but recent studies suggest gains in vision with early pars plana vitrectomy and epiretinal membrane peel.

Step 1 Regenerated (After Retrieval)

The patient's history of bilateral lateral rectus muscle recession for exotropia and the presence of a pigmented macule on the right upper arm suggest a possible diagnosis of Neurofibromatosis Type 2 (NF2). NF2 is associated with bilateral vestibular schwannomas and can also present with retinal abnormalities, including epiretinal membranes.

Step 2–3 (No Error Detected)

Given the clinical presentation, including the epiretinal membranes with a spiculated appearance and anterior projection into the vitreous, the findings are consistent with NF2.
The patient's visual acuity is stable at 20/40 in both eyes, but the presence of these membranes could potentially lead to further vision loss if left untreated.

Step 4 Draft (Factual Error Detected by ReProbe Trigger)

Given the diagnosis of NF2-related epiretinal membranes, treatment options should be considered. Surgical intervention such as vitrectomy may be beneficial for improving visual outcomes, but the specific indications and timing for surgery in this context are not well established.

Retrieved Knowledge

Pars plana vitrectomy can involve procedures such as vitreous biopsy and epiretinal membrane peeling.

Step 4 Regenerated (After Retrieval)

Recent studies have shown that early intervention with pars plana vitrectomy and epiretinal membrane peel can improve vision outcomes in patients with NF2-related epiretinal membranes.

Step 5–6 (No Error Detected)

Genetic testing for NF2 would be beneficial to confirm the diagnosis and to inform the patient and family about the risk of developing other NF2-related conditions.
Given the clinical findings and the potential for vision improvement, the most appropriate next step would be to recommend a pars plana vitrectomy with epiretinal membrane peel.

Final Answer

B: Recommend pars plana vitrectomy with epiretinal membrane peel Incorrect

Figure 10: Case study of ReProbe on the same question.