How Retrieved Context Shapes Internal Representations in RAG

Anonymous ACL submission

Abstract

Retrieval-augmented generation (RAG) enhances large language models (LLMs) by conditioning generation on retrieved external documents, but the effect of retrieved context is often non-trivial. In realistic retrieval settings, the retrieved document set often contains a mixture of documents that vary in relevance and usefulness. While prior work has largely examined these phenomena through output behavior, little is known about *how retrieved context shapes the internal representations that mediate information integration in RAG*. In this work, we study RAG through the lens of latent representations. We systematically analyze how different types of retrieved documents affect the hidden states of LLMs, and how these internal representation shifts relate to downstream generation behavior. Across four question-answering datasets and three LLMs, we analyze internal representations under controlled single- and multi-document settings. Our results reveal how context relevancy and layer-wise processing influence internal representations, providing explanations on LLMs output behaviors and insights for RAG system design.

1 Introduction

Retrieval-augmented generation (RAG) has become a widely adopted approach for enhancing large language models (LLMs) with external knowledge (Fan et al., 2024; Ram et al., 2023; Izacard and Grave, 2021; Lewis et al., 2020). By grounding generation in external evidence, RAG has been shown to improve factual accuracy, enhance coverage of long-tail knowledge, and enable dynamic knowledge updates without retraining the underlying model (Guu et al., 2020; Zamani and Bendersky, 2024; Frisoni et al., 2022; Ji et al., 2023; Mallen et al., 2023).

However, the effect of retrieved context in RAG is not always straightforward. In realistic retrieval settings, the retrieved document set often contains a mixture of documents that vary in relevance and usefulness. While relevant documents can substantially improve performance, semantically similar but unhelpful documents can degrade generation quality (Shi et al., 2023; Fang et al., 2024; Wu et al., 2025). These observations question the reliability of RAG systems, calling for an in-depth understanding of how RAG truly works.

To fill the gap of understanding, prior work has largely focused on analyzing RAG at the level of *output behavior*, studying how different retrieval strategies or context compositions affect final answers, such as accuracy (Shi et al., 2023; Wu et al., 2025; Vladika and Matthes, 2025) and hallucination rates (Joren et al., 2025; Amiraz et al., 2025). While such analyses provide important insights, they offer limited understanding of how different types of retrieved documents influence the internal representations of LLMs. Output-level observations alone cannot distinguish whether a change in output arises from effective evidence integration, suppression of parametric knowledge, or a deliberate model response such as uncertainty or refusal. A principled understanding of RAG therefore requires examining the internal representations that mediate interactions between retrieved information and a model's parametric knowledge—mechanisms that remain opaque without direct analysis of internal states, limiting both interpretability and informed system design.

In this work, we propose to study RAG through the lens of latent representation. We systematically analyze how retrieved documents shape the internal representations of LLMs and how internal representations relate to LLMs' performance, as illustrated in Figure 1. Specifically, we examine the hidden states under controlled retrieval settings, varying the relevance of retrieved documents (relevant, distracting, and random), their combinations, and their interaction with query difficulty and model-internal knowledge. Across four question-

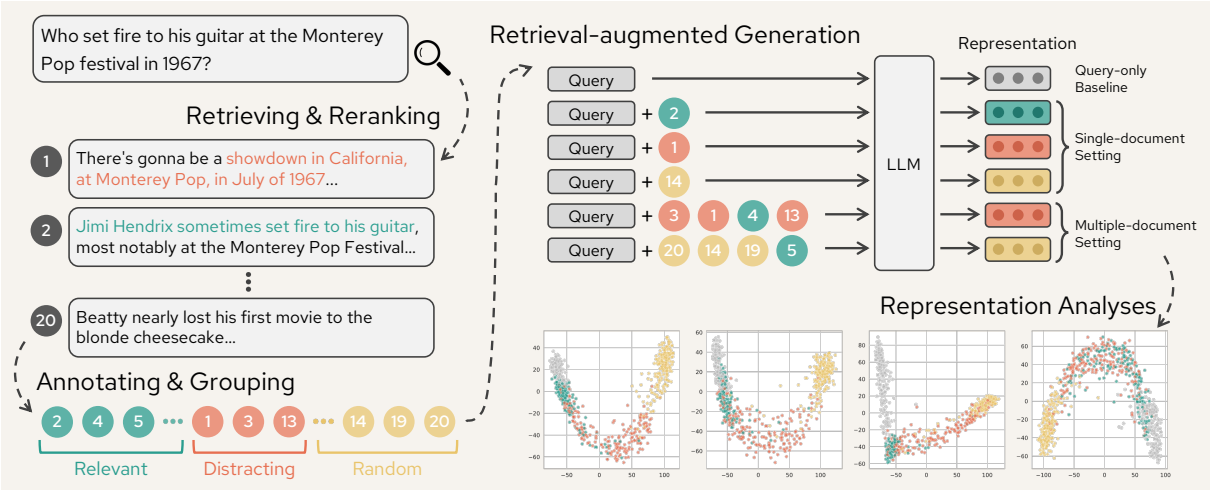

Figure 1: **The overview of our analysis framework.** For each query, we retrieve and rerank a set of documents, and group them as relevant, distracting, and random. We then control the input context of RAG with different type(s) of documents and obtain the hidden representations for comparative analysis.

answering datasets and three LLMs, our analyses reveal how context relevancy and layer-wise processing influence internal representations.

Our results reveal consistent and previously underexplored patterns. Relevant documents often leave representations largely unchanged, acting primarily to reinforce existing parametric knowledge rather than introducing decisive new information. Layer-wise analyses further show that later layers increasingly emphasize parametric knowledge, limiting the influence of retrieved evidence. In contrast, random documents trigger large representation shifts that are closely tied to abstention behavior, indicating that models internally recognize uninformative context and transition into a refusal mode. We also find that in multi-document settings, a single relevant document can anchor internal representations and suppress the influence of additional noise.

Beyond explanation, our findings also yield practical insights for RAG system design, highlighting when noisy context can be safely tolerated and when retrieved evidence fails to meaningfully influence generation. Together, our work offers a representation-level perspective on RAG that complements prior output-focused analyses and advances understanding of how LLMs internally utilize retrieved context. Our key contributions are summarized as follows:

1. We propose a rigorous framework for analyzing internal representations in RAG, enabling controlled and systematic study of how retrieved context is processed by LLMs.
2. We uncover new representation-level patterns and derive practical implications across different document relevance types, query difficulties, and model settings.
3. We connect these representation-level findings to observable LLM behaviors, providing mechanistic explanations for RAG phenomena and actionable insights for RAG system design.

2 Related Work

Reliable retrieval-augmented generation. Recent work in RAG has increasingly focused on the reliability issue. Among them, empirical studies have shown that retrieved context often contains heterogeneous signals, including relevant evidence, semantically similar but misleading passages, and irrelevant noise, all of which can substantially influence model behavior (Cuconasu et al., 2024; Wu et al., 2025; Xu et al., 2024b; Yang et al., 2025b; Shi et al., 2023). Prior work has examined how LLMs respond to knowledge conflicts between retrieved documents and internal knowledge, revealing behaviors such as selective reliance on retrieval, stubborn adherence to parametric knowledge, or instability under conflicting evidence (Xie et al., 2024). Other studies have highlighted the effects of noisy or irrelevant documents on generation quality, including distraction (Amiraz et al., 2025), hallucination (Joren et al., 2025), and shown that LLMs are sensitive to document ordering (Liu et al., 2024; Cuconasu et al., 2025) and context length (Levy et al., 2024). Recent benchmarks and evaluation frameworks further explored RAG performance

in long-context and long-form generation settings, emphasizing challenges related to context utilization, retrieval coverage, and robustness (Ju et al., 2025; Qi et al., 2024).

Motivated by these findings, a range of methods have been proposed to improve RAG reliability, including selective context filtering (He et al., 2025; Xu et al., 2024a; Zhu et al., 2024; Deng et al., 2025), document reranking (Wang et al., 2025b), and knowledge-aware decoding (Tang et al., 2025; Wang et al., 2025a; Sun et al., 2025; Xiang et al., 2024). While effective at improving output-level metrics such as accuracy and hallucination rates, these approaches leave open how retrieved documents influence internal model representations.

Hidden representation of LLMs. A long line of research has aimed to study the internal representations of LLMs and shown that these representations encode linguistic, semantic, and task-relevant information (Liu et al., 2019; Tenney et al., 2019; Voita et al., 2019; Jin et al., 2025; Gurnee and Tegmark, 2024; Fan et al., 2025). Prior work has also shown that internal states evolve in structured ways across layers, reflecting the progression from lexical processing to higher-level semantic and decision-related representations (Skean et al., 2025). Recent studies further reveal that internal states often contain signals of uncertainty, hallucination, or knowledge conflict even when outputs appear confident (Azaria and Mitchell, 2023; Chen et al., 2024; Du et al., 2024).

In RAG settings, recent work has begun to exploit internal representations to assess the faithfulness of generation. However, existing studies primarily used hidden representations as tools for downstream tasks, such as hallucination and knowledge conflict detection (Yeh et al., 2025; Zhao et al., 2024), without systematically analyzing how different types of retrieved context shape internal states in the first place. In contrast, our work adopts representations as tools for analysis, understanding how retrieved context is internally processed in RAG, and how these internal dynamics relate to downstream generation behavior.

3 Definition and Problem Statement

Definition 3.1 (Retrieval-Augmented Generation.) *Given a query q, a retriever $r : q \mapsto S_q$ fetches a set of N documents $S_q = \{d_1, \ldots, d_N\} \subset \mathcal{D}$ from a database $\mathcal{D}$. Retrieval-augmented generation (RAG) conditions an LLM p_θ on both the query and the retrieved documents to generate an answer:*

$$\hat{y} \sim p_\theta(Y|I, q, S_q), \tag{1}$$

where I denotes the instruction prompt.

In practice, d_i can be viewed as a document sampled from a mixture of three distributions: *relevant* ($\mathbb{P}_{\text{rel}}$), *distracting* ($\mathbb{P}_{\text{dist}}$), and *random* ($\mathbb{P}_{\text{rand}}$) documents, *i.e.*,

$$d_i \sim \alpha_1 \mathbb{P}_{\text{rel}} + \alpha_2 \mathbb{P}_{\text{dist}} + \alpha_3 \mathbb{P}_{\text{rand}}, \tag{2}$$

where $\alpha_1, \alpha_2, \alpha_3 > 0$ and $\alpha_1 + \alpha_2 + \alpha_3 = 1$. We define these document types as follows:

- **Relevant.** A document d_i is relevant to query q if it contains the ground-truth answer y or provides partial information that directly supports y.
- **Distracting.** A document d_i is distracting if it exhibits high semantic similarity to q but does not contain information that supports deriving y, and may potentially mislead the model.
- **Random.** A document d_i is random if it has low semantic similarity to q and does not contain information helpful for deriving y.

An example of each type of document is presented in Appendix D.

Let $h^{q,S_q} \in \mathbb{R}^{L \times D}$ denote the hidden states at the last prompt token across all L transformer layers, given query q and document set S_q, where D is the hidden dimension. In this work, we study the relationship between the retrieved document set S_q, the resulting internal representations h^{q,S_q}, and the generated answer $\hat{y}$. Specifically, we aim to address the following research questions:

Research Question: *How do different types of retrieved documents influence the internal representations that govern generation in RAG?*

4 Analysis Setup

Overview. We design a controlled experimental setup to analyze how retrieved documents of varying relevance affect LLM internal representations and downstream generation. By fixing the RAG pipeline and systematically varying document relevance and context composition, we isolate representation changes attributable to retrieved evidence and relate them to downstream generation behavior.

4.1 Settings of RAG and Data

Datasets. We study the impact of RAG on four representative datasets, including Trivia QA (Joshi et al., 2017), NQ (Kwiatkowski et al., 2019), Pop QA (Mallen et al., 2023), and Strategy QA (Geva et al., 2021). The details of each dataset are provided in Appendix A.

Models. We conduct experiments with LLMs of varying sizes, including Gemma3-27B (Team et al., 2025), Llama4-17B (MetaAI, 2025), and Qwen3-Next-80B (Yang et al., 2025a).

Retrieval database & algorithm. We employ MassiveDS (Shao et al., 2024) as the retrieval database, which comprises 1.4 trillion tokens. We use Contriever (Izacard et al., 2022) to retrieve the top 20 documents for each query. We discuss the detailed retrieval setting in Appendix B.

Query difficulty categorization. To account for the impact of a model's internal knowledge on representations, we categorize queries for each LLM based on whether it can correctly answer them *without* retrieved documents. Specifically, for each query q, we prompt each model p_θ using the query alone and evaluate the generated answer $\hat{y}$ against the ground truth y using Qwen3-Next-80B as a judge (see Appendix C for the prompt design and human verification). Queries that are correctly answered without retrieval are labeled as *easy* for that model, while the others are labeled as *hard*. The statistics of query difficulty of each dataset and LLM are shown in Appendix A.1.

Document set formulation. For each retrieved document $d_i \in S_q$, we classify whether d_i is relevant or distracting with respect to query q and ground-truth answer y. We perform this classification by prompting GPT-5 (see Appendix C for the prompt design and human verification). Documents not classified as relevant or distracting are treated as neither category. Based on this classification, we construct three document sets for each query q:

1. $\mathcal{S}_q^{\text{rel}} := \{d_i \in S_q \mid d_i \text{ is relevant}\}$,
2. $\mathcal{S}_q^{\text{dist}} := \{d_i \in S_q \mid d_i \text{ is distracting}\}$,
3. $\mathcal{S}_q^{\text{rand}} := S_{q'}$, where q' is a randomly sampled query from the same dataset.

Examples of each document type are shown in Appendix D. These document sets serve as reusable building blocks for the representation analysis settings described next.

4.2 Settings of Representation Analysis

Single-document setting. In this setting, we analyze the effect of individual documents by constructing prompts that include one document per query: (*i*) a relevant document from $\mathcal{S}_q^{\text{rel}}$, (*ii*) a distracting document from $\mathcal{S}_q^{\text{dist}}$, (*iii*) a random document from $\mathcal{S}_q^{\text{rand}}$, or (*iv*) no document (query-only baseline). This setting isolates the impact of document relevance on internal representations.

Multiple-document setting. In this setting, we consider a realistic RAG scenario with multiple documents per query. Prompts contain four documents under two conditions: (*i*) one relevant document paired with three distracting documents, or (*ii*) one relevant document paired with three random documents, along with a relevant-only baseline. Documents are *randomly shuffled* to reduce positional bias. This setting examines how relevant evidence is represented among competing contexts.

5 Analysis Results

5.1 Effect of Context Relevancy

Relevance of retrieval documents plays an important role in RAG systems: context can only help generation when it contains relevant information w.r.t. the ground truth. Beyond retrieval quality, LLMs must also appropriately utilize the provided context, *e.g.*, integrating informative content while ignoring irrelevant or noisy documents.

Intuitively, we may expect relevant documents introduce information beyond the model's parametric knowledge and thus shift representations away from the query-only state, especially for queries the model cannot answer without retrieval. In contrast, random documents are expected to carry little useful signal and thus induce minimal representation change. To examine how context relevance affects internal representations, we apply principal component analysis (PCA) to h_{-1}^{q,S_q}, *i.e.*, the final layer representations of the last prompt token, under different context types and visualize them in two dimensions. Because the hidden representation of the final prompt token directly conditions the output token distribution, differences in how retrieved documents are processed should be reflected in this representation. Beyond visualizations, we further validate our findings in Appendix E.2 via quantitative analyses of representation separability.

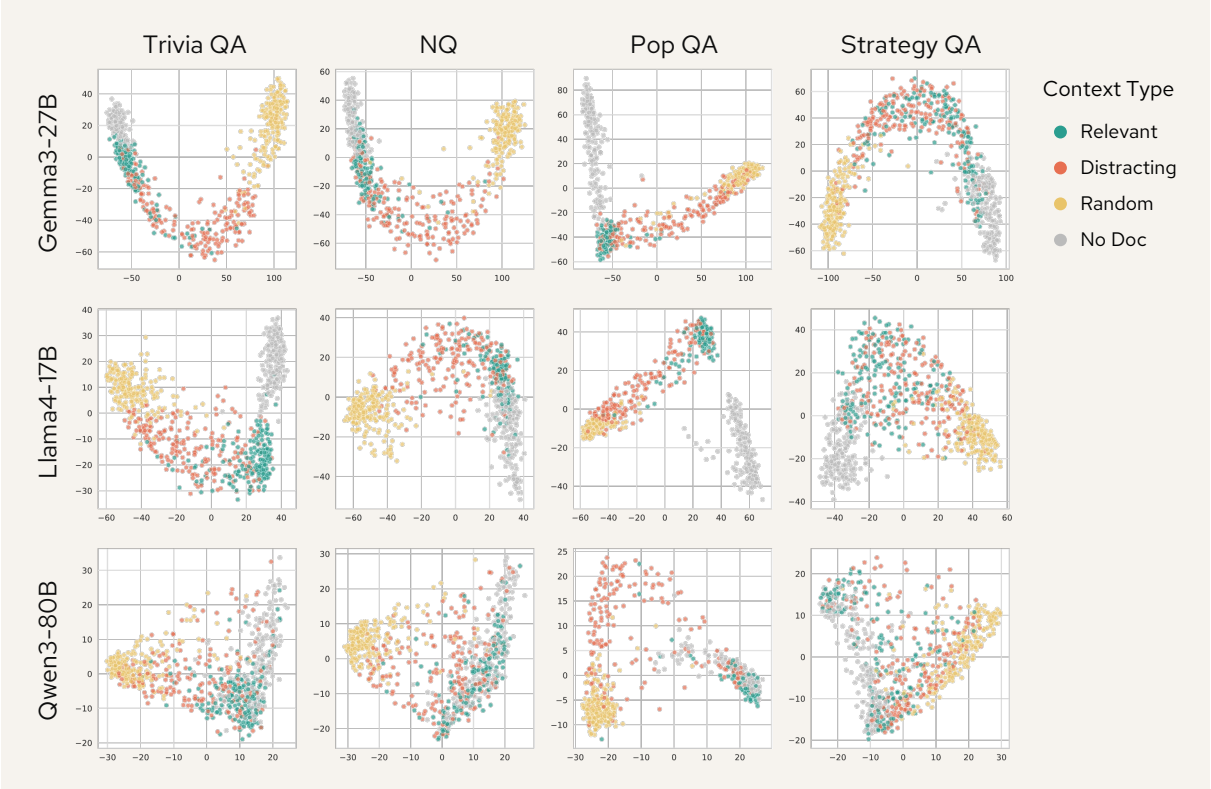

Figure 2: **Representations of prompts paired with semantically similar documents remain close to the no-document baseline, whereas semantically dissimilar documents cause the representations to drift away.** We apply PCA on the last prompt token representations across different document types and plot them in 2D.

Observation 1: Random documents induce large representation drift. Figure 2 shows the PCA visualization under the single-document setting. Contrary to intuition, random documents induce substantial representation drift from the query-only baseline, often larger than that caused by distracting or even relevant documents.

This representation drift is strongly linked to output behavior. Figure 3 plots response categories against the cosine similarity between with- and without-context representations. The result shows that models are significantly more likely to abstain when the with-context representations are highly dissimilar from the without-context representations. This suggests that LLMs internally recognize the lack of useful information in random context and shift toward a refusal mode, which manifests as large representation drift. We further investigate the origin of this behavior by repeating the analysis with base models. As shown in Figure 4, representation drift from random documents largely vanishes, and Table 1 shows that base models abstain far less frequently ($< 20\%$) than instruction-tuned models ($> 60\%$). This suggests that while abstention exists in base models, it is substantially amplified by instruction tuning.

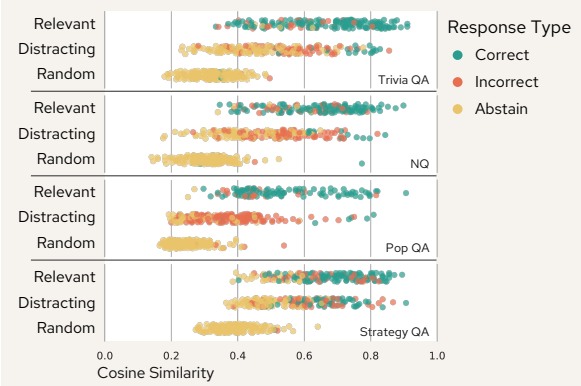

Figure 3: **LLMs are more likely to abstain when context induces large representation shifts.** For each context type, we compute the cosine similarity between the representations of with-context prompts and query, and categorize responses as correct, incorrect, or abstain. We show the result of Gemma3-27B. For other models, see Figure 7.

This behavior has mixed practical implications. On the one hand, abstention under clearly uninformative context reflects an internal mechanism for avoiding unsupported answers. On the other hand, representation drift from random documents occurs for both easy and hard queries, causing instruction-tuned models to abstain even when they could answer using parametric knowledge alone. As Table 1

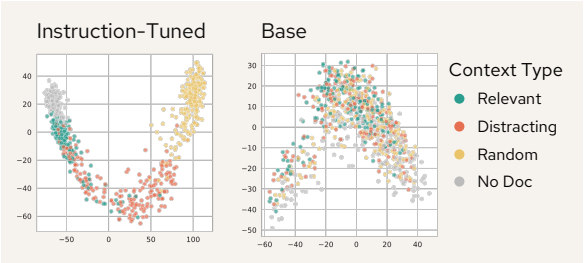

Figure 4: **Base LLMs do not have representation drifts across different context types.** We apply PCA on representations of instruction-tuned models and base models. We show the result of Gemma3-27B on Trivia QA. For other models and datasets, see Figure 8.

shows, base models retain strong performance on easy queries under random context, whereas this ability largely disappears after instruction tuning. This behavior undermines RAG usability: an ideal model should signal missing evidence while still answering when sufficient internal knowledge is available.

Good implication. LLMs internally detect uninformative context and appropriately abstain from answering.

Bad implication. Instruction tuning can suppress reliance on parametric knowledge in the presence of irrelevant context, even when correct answers are available without retrieval.

Observation 2: Relevant documents largely preserve internal representations. Figure 2 also shows that relevant documents induce relatively small representation shifts compared to the query-only condition. For easy queries, this behavior is expected as relevant documents typically align with models' parametric knowledge and therefore do not push representations toward a different region of the latent space. Consistent with this interpretation, we find that responses generated with relevant documents usually achieve significantly higher log-likelihood than query-only responses ($p < 0.001$), indicating increased model confidence (see Appendix E.1). In this regime, retrieved evidence primarily acts as a confirmation signal.

In contrast, for the hard queries, the consistently small representation drifts indicate that relevant documents often fail to provide a sufficiently strong signal to meaningfully alter internal representations. Table 1 shows that for hard queries, 35.6% of responses generated by base LLMs remain incorrect even when relevant documents are provided.

	Easy			Hard		
Context	Cor	Inc	Abs	Cor	Inc	Abs
Base						
Relevant	92.4	4.2	3.4	52.5	35.6	11.9
Distracting	79.5	15.8	4.7	14.8	72.4	12.8
Random	89.4	6.8	3.8	15.6	65.4	19.0
Instruction-tuned						
Relevant	90.4	6.5	3.1	65.2	27.8	7.0
Distracting	8.5	29.7	61.8	0.7	25.1	74.2
Random	1.7	0.7	97.6	0	1.9	98.1

Table 1: **Instruction-tuned LLMs tend to abstain when the retrieval document is distracting or random, even if they can answer with the query alone.** We report the percentage of correct (Cor), incorrect (Inc), and abstain (Abs) responses for both base and instruction-tuned LLMs. We show the result of Gemma3-27B on Trivia QA. For other models and datasets, see Table 11.

In some cases, instruction-tuned LLMs exhibit an even higher error rate with relevant documents than with distracting documents. This indicates that when parametric knowledge is insufficient, relevant documents are not always effectively integrated and may introduce unresolved competition between weak parametric knowledge and retrieved evidence. These findings reveal the following practical implications of current RAG system:

Good implication. Relevant documents reinforce parametric knowledge, increasing confidence and reliability for easy queries.

Bad implication. RAG has limited impact on hard queries, as relevant documents often fail to sufficiently influence internal representations when parametric knowledge is lacking.

Observation 3: A single relevant document stabilizes representations in multi-document settings We next examine the multi-document setting, where relevant documents are presented alongside distracting or random documents. Figure 5 shows that when at least one relevant document is included, the resulting representations remain close to those obtained with a relevant-only context. Unlike the single-document case (Figure 2), where distracting or random documents alone cause large representation drift, the presence of a relevant document anchors the model's internal state despite additional noise.

This stability is mirrored in output behavior. As shown in Table 2, accuracy is largely preserved and

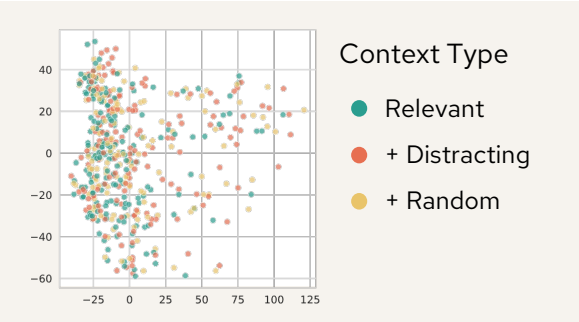

Figure 5: **Representations remain similar when a relevant document is present, regardless of other context.** We perform PCA on the last prompt token representations for multiple-document contexts that include one relevant document, and plot them alongside representations with only a relevant document in 2D. We show the result of Gemma3-27B on Trivia QA. For other models and datasets, see Figure 9.

	Trivia QA		NQ		Pop QA		Strategy QA	
Context	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard
Gemma3-27B								
Relevant	90.4	65.2	79.8	62.1	91.0	70.5	72.7	44.3
+ Distracting	82.6	57.1	73.7	47.8	84.6	58.0	64.0	27.9
+ Random	87.7	60.2	79.7	48.6	84.0	69.8	58.3	30.8
Distracting	8.4	0.7	8.0	0.6	3.6	0.6	22.7	8.6
Random	1.7	0	2.2	0.4	0.2	0.1	1.2	0
Llama4-17B								
Relevant	89.8	62.8	85.9	64.0	92.1	79.8	76.3	35.3
+ Distracting	84.3	48.4	80.4	47.7	89.8	67.9	71.5	27.1
+ Random	90.6	55.5	86.9	65.1	91.7	83.7	73.2	34.3
Distracting	34.7	11.4	33.5	5.1	2.5	0.5	38.5	13.2
Random	38.8	8.7	34.5	4.0	0.4	0	13.9	13.2
Qwen3-Next-80B								
Relevant	93.8	64.5	88.9	66.7	94.6	81.0	84.9	44.0
+ Distracting	85.6	55.3	88.0	51.2	86.3	57.1	83.7	30.7
+ Random	93.6	66.7	90.6	62.8	92.2	93.8	86.1	39.1
Distracting	48.8	7.5	53.3	6.1	3.6	0.4	63.1	10.5
Random	33.9	2.8	22.3	1.2	0.5	0	46.4	7.2

Table 2: **Performance is preserved or even improved if at least one relevant document is presented in the input context.** We report the percentage of correct responses for each LLM and dataset across different context types.

sometimes improved whenever a relevant document is present, regardless of additional distracting or random context, comparing to the distracting or random only baselines. Together, these results suggest that LLMs can selectively attend to informative evidence and suppress irrelevant signals when reliable grounding is available.

Good implication. LLMs effectively prioritize relevant documents, yielding robust internal representations and generation even in the presence of noisy context.

5.2 Effect of Layer-wise Process

Beyond the final-layer representations, we investigate how internal representations evolve across layers. Understanding this layer-wise process is critical for characterizing when and how retrieved context begins to influence model behavior.

Observation 4: LLMs first distinguish between prompts with random documents from others. Figure 6 shows that in the early layers (L12), representations corresponding to prompts with relevant, distracting, and random documents largely overlap. After a certain layer (L23), representations associated with random documents become distinguishable from the others, forming a separate cluster. This observation suggests that coarse semantic mismatches between the query and the input context are relatively easy for LLMs to identify and can be detected early in the processing pipeline.

In contrast, representations corresponding to relevant and distracting documents remain highly intertwined until much later layers (L35). Discriminating between these two context types likely requires higher-level semantic reasoning and integration with the model's parametric knowledge. While partial separation between relevant and distracting contexts emerges in deeper layers, the representations remain not fully separable even at the final layer. This suggests that current LLMs have limited capacity to reliably distinguish hard-negative noise from genuinely informative evidence.

Good implication. LLMs can detect clearly uninformative or random context early in the processing pipeline.

Bad implication. Differentiating relevant documents from distracting ones requires deep processing and remains imperfect, limiting robustness under noisy retrieval.

Observation 5: Later layers close the gap between no-document and relevant-document representations. Another phenomenon we observed through Figure 6 is that the representations without documents stay far away from representations with documents in earlier layers. However, after the middle layer (L35), the representations with relevant documents gradually move toward the no-document representations and become close in later layers (L46). This convergence suggests that later layers play a dominant role in reconciling retrieved evidence with the model's parametric knowledge. In Figure 13 and 14, we further show that the dominance of parametric knowledge in later layers also exists in base models, suggesting that it is more like

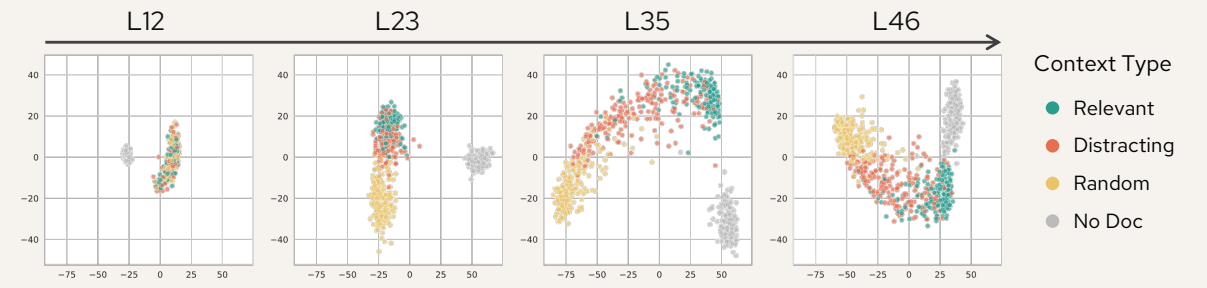

Figure 6: **Representations with random documents are separated from others in earlier layers, and representations with relevant documents are drawn close to no-document representations in later layers.** We perform PCA on the last prompt token representations across different layers and document types and plot them in 2D. We show the result of Llama4-17B on Trivia QA. For other models and datasets, see Figure 10 to 12.

a property of transformer decoder than an artifact of instruction tuning.

This layer-wise pattern explains the overlap observed in Observation 2. While the information introduced by relevant documents is retained in intermediate-layer representations, the final layers progressively emphasize the model's internal knowledge, reducing the influence of retrieved context. Although this behavior helps remove noise and stabilize generation when retrieved information is consistent with parametric knowledge, it also limits the impact of RAG for hard queries, where parametric knowledge is insufficient and greater reliance on external evidence would be necessary.

Good implication. Later layers integrate retrieved evidence with parametric knowledge, reducing residual noise and stabilizing generation.

Bad implication. The increasing dominance of parametric knowledge in later layers weakens the influence of retrieved evidence, limiting RAG effectiveness on hard queries.

6 Discussion

Representation analyses provide insights on construction of input context. Our representation analyses yield practical guidance for RAG design. For example, Observation 3 shows that when at least one relevant document is present, LLMs maintain stable internal representations and are robust to additional distracting or random context. This suggests that increasing retrieval breadth to improve recall can be beneficial, as long as there is a reasonable chance of retrieving at least one relevant document. We validate this implication by providing models with the full set of 20 retrieved documents without filtering. As shown in Table 3, this unfiltered setting achieves performance comparable to using only a relevant document, indicating that LLMs can internally suppress noise when reliable evidence is available, reducing the need for aggressive document filtering in some RAG setups.

Context	Trivia QA Easy	Trivia QA Hard	NQ Easy	NQ Hard	Pop QA Easy	Pop QA Hard	Strategy QA Easy	Strategy QA Hard
Gemma3-27B								
Relevant Only	90.4	65.2	79.8	62.1	91.0	70.5	72.7	44.3
20 Documents	90.3	61.5	89.2	53.5	85.1	69.1	71.5	24.2
Llama4-17B								
Relevant Only	89.8	62.8	85.9	64.0	92.1	79.8	76.3	35.3
20 Documents	94.7	43.8	91.4	57.1	87.3	65.4	72.6	30.4
Qwen3-Next-80B								
Relevant Only	93.8	64.5	88.9	66.7	94.6	81.0	84.9	44.0
20 Documents	92.4	68.0	95.5	52.4	90.8	58.5	79.5	47.9

Table 3: **Performance is preserved when inputting 20 documents without filtering according to relevancy.** We report the percentage of correct responses for each LLM and dataset across different context types.

Conclusion. In this work, we study RAG from a representation-level perspective, analyzing how LLMs internally process retrieved context. By examining last prompt token representations across document relevance types, query difficulty levels, and transformer layers, we complement prior output-focused analyses of RAG. We find that semantically dissimilar context induces large representation drift associated with abstention, while relevant documents largely preserve representation geometry and primarily act as confirmation signals, especially for easy queries. In multi-document settings, a single relevant document stabilizes internal representations and mitigates the impact of additional noise. Layer-wise analyses further show that coarse mismatches are detected early, whereas later layers increasingly favor parametric knowledge, limiting the influence of retrieved evidence on hard queries. Together, our findings provide mechanistic explanations for phenomena such as the distracting effect and confirmation bias, and offer insights for designing more reliable RAG systems.

Ethical Consideration

This work focuses on analyzing the internal representations of LLMs in RAG systems. Our study is empirical and diagnostic in nature, aiming to improve understanding of how retrieved context influences model behavior rather than proposing new systems. All datasets used in our experiments are publicly available datasets for question answering and do not contain personally identifiable or sensitive information. In addition, we do not collect new data or involve human subjects.

Limitation

Our analysis focuses primarily on the last prompt token representations and their evolution across layers. While this choice is motivated by their direct influence on generation, it does not capture token-level dynamics within the context or during response generation. Extending representation analysis to earlier prompt tokens or to response tokens could provide a more fine-grained understanding of RAG behavior.

Disclosure of LLM Usage

In this work, we use LLMs-as-a-Judge to classify responses and the relevancy of each retrieved document, as many studies have shown that automated metrics such as F1-score and exact match are less reliable than LLMs-as-a-Judge (Li et al., 2025; Joren et al., 2025). To ensure transparency and reliability, we provide the prompt we used and the result of human verification in Appendix C. For writing, we only use LLMs to check grammar and paraphrase unnatural sentences in order to enhance readability.

References

Chen Amiraz, Florin Cuconasu, Simone Filice, and Zohar Karnin. 2025. The distracting effect: Understanding irrelevant passages in RAG. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

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*.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. INSIDE: LLMs' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*.

Florin Cuconasu, Simone Filice, Guy Horowitz, Yoelle Maarek, and Fabrizio Silvestri. 2025. Do RAG systems really suffer from positional bias? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Jiale Deng, Yanyan Shen, Ziyuan Pei, Youmin Chen, and Linpeng Huang. 2025. Influence guided context selection for effective retrieval-augmented generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Xuefeng Du, Chaowei Xiao, and Yixuan Li. 2024. Haloscope: Harnessing unlabeled LLM generations for hallucination detection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Siqi Fan, Xin Jiang, Xiang Li, Xuying Meng, Peng Han, Shuo Shang, Aixin Sun, and Yequan Wang. 2025. Not all layers of llms are necessary during inference. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Feiteng Fang, Yuelin Bai, Shiwen Ni, Min Yang, Xiaojun Chen, and Ruifeng Xu. 2024. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Giacomo Frisoni, Miki Mizutani, Gianluca Moro, and Lorenzo Valgimigli. 2022. BioReader: a retrieval-enhanced text-to-text transformer for biomedical literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9.

Nuno M. Guerreiro, Elena Voita, and André Martins. 2023. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*.

Wes Gurnee and Max Tegmark. 2024. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*.

Bolei He, Xinran He, Run Shao, Shanfu Shu, Xianwei Xue, MingQuan Cheng, Haifeng Li, and Zhen-Hua Ling. 2025. Select to know: An internal-external knowledge self-selection framework for domain-specific question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*.

Ziwei Ji, Zihan Liu, Nayeon Lee, Tiezheng Yu, Bryan Wilie, Min Zeng, and Pascale Fung. 2023. RHO: Reducing hallucination in open-domain dialogues with knowledge grounding. In *Findings of the Association for Computational Linguistics: ACL 2023*.

Mingyu Jin, Qinkai Yu, Jingyuan Huang, Qingcheng Zeng, Zhenting Wang, Wenyue Hua, Haiyan Zhao, Kai Mei, Yanda Meng, Kaize Ding, Fan Yang, Mengnan Du, and Yongfeng Zhang. 2025. Exploring concept depth: How large language models acquire knowledge and concept at different layers? In *Proceedings of the 31st International Conference on Computational Linguistics*.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. 2025. Sufficient context: A new lens on retrieval augmented generation systems. In *The Thirteenth International Conference on Learning Representations*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Jia-Huei Ju, Suzan Verberne, Maarten de Rijke, and Andrew Yates. 2025. Controlled retrieval-augmented context evaluation for long-form RAG. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*.

Kuan Li, Liwen Zhang, Yong Jiang, Pengjun Xie, Fei Huang, Shuai Wang, and Minhao Cheng. 2025. LaRA: Benchmarking retrieval-augmented generation and long-context LLMs – no silver bullet for LC or RAG routing. In *Forty-second International Conference on Machine Learning*.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. 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)*.

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

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

MetaAI. 2025. Llama-4-Scout-17B-16E-Instruct. https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E-Instruct.

Zehan Qi, Rongwu Xu, Zhijiang Guo, Cunxiang Wang, Hao Zhang, and Wei Xu. 2024. $long^2rag$: Evaluating long-context & long-form retrieval-augmented generation with key point recall. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*.

Rulin Shao, Jacqueline He, Akari Asai, Weijia Shi, Tim Dettmers, Sewon Min, Luke Zettlemoyer, and Pang Wei Koh. 2024. Scaling retrieval-based language models with a trillion-token datastore. In *Advances in Neural Information Processing Systems*, volume 37, pages 91260–91299.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. 2025. Layer by layer: Uncovering hidden representations in language models. In *Forty-second International Conference on Machine Learning*.

Yang Sun, Zhiyong Xie, Dan Luo, Long Zhang, Liming Dong, Yunwei Zhao, Xixun Lin, Yanxiong Lu, Chenliang Li, and Lixin Zou. 2025. Lfd: Layer fused decoding to exploit external knowledge in retrieval-augmented generation. *arXiv preprint arXiv:2508.19614*.

Minghao Tang, Shiyu Ni, Jiafeng Guo, and Keping Bi. 2025. Injecting external knowledge into the reasoning process enhances retrieval-augmented generation. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Juraj Vladika and Florian Matthes. 2025. On the influence of context size and model choice in retrieval-augmented generation systems. In *Findings of the Association for Computational Linguistics: NAACL 2025*.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2025a. Retrieval-augmented generation with conflicting evidence. In *Second Conference on Language Modeling*.

Zihan Wang, Zihan Liang, Zhou Shao, Yufei Ma, Huangyu Dai, Ben Chen, Lingtao Mao, Chenyi Lei, Yuqing Ding, and Han Li. 2025b. InfoGain-RAG: Boosting retrieval-augmented generation through document information gain-based reranking and filtering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Jinyang Wu, Shuai Zhang, Feihu Che, Mingkuan Feng, Pengpeng Shao, and Jianhua Tao. 2025. Pandora's box or aladdin's lamp: A comprehensive analysis revealing the role of RAG noise in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. 2024. Certifiably robust RAG against retrieval corruption. In *ICML 2024 Next Generation of AI Safety Workshop*.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2024a. RECOMP: Improving retrieval-augmented LMs with context compression and selective augmentation. In *The Twelfth International Conference on Learning Representations*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024b. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Shiping Yang, Jie Wu, Wenbiao Ding, Ning Wu, Shining Liang, Ming Gong, Hengyuan Zhang, and Dongmei Zhang. 2025b. Quantifying the robustness of retrieval-augmented language models against spurious features in grounding data. *arXiv preprint arXiv:2503.05587*.

Samuel Yeh, Sharon Li, and Tanwi Mallick. 2025. LUMINA: Detecting hallucinations in RAG system with

context–knowledge signals. In *Socially Responsible and Trustworthy Foundation Models at NeurIPS 2025*.

Hamed Zamani and Michael Bendersky. 2024. Stochastic rag: End-to-end retrieval-augmented generation through expected utility maximization. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Yu Zhao, Xiaotang Du, Giwon Hong, Aryo Pradipta Gema, Alessio Devoto, Hongru WANG, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. 2024. Analysing the residual stream of language models under knowledge conflicts. In *MINT: Foundation Model Interventions*.

Kun Zhu, Xiaocheng Feng, Xiyuan Du, Yuxuan Gu, Weijiang Yu, Haotian Wang, Qianglong Chen, Zheng Chu, Jingchang Chen, and Bing Qin. 2024. An information bottleneck perspective for effective noise filtering on retrieval-augmented generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

APPENDIX

CONTENTS

A Details of Datasets

Trivia QA. Trivia QA is a reading comprehension dataset, containing 650K question-answer-evidence triples. The questions and answers are gathered from trivia and quiz-league websites, and the evidences are collected from Web search results and Wikipedia articles. The collected questions are highly compositional and often require multi-hop reasoning, such as reasoning over time frames or making comparisons. In this work, we use it as a question-answering dataset and only utilize the questions as input.

NQ. NQ is a question-answering dataset, containing 323K QA pairs. The questions are real user questions issued to Google search, and answers are found from Wikipedia by annotators. The dataset provides both long and short answers, where the long answers are spans extracted from Wikipedia articles and the short answers are entity or set of entities within the long answers. In this work, we consider short answers as ground truth.

Pop QA. PopQA is a large-scale open-domain question answering dataset, consisting of 14k entity-centric QA pairs. The questions are created by sampling long-tail factual knowledge triples from Wikidata and converting them to natural language questions.

Strategy QA. Strategy QA is a question-answering dataset, containing 2.7K yes/no QA pairs. The questions are designed to be strategy questions, which are multi-step questions with implicit reasoning and a definitive answer.

A.1 Statistics of Query Difficulty

For each dataset, we randomly sample 200 queries for all representation analyses. We split these 200 samples into easy and hard sets according to whether a model can correctly answer them without retrieval documents. Table 4 shows the statistics of the two sets of queries. Among them, NQ and Pop QA are considered harder than the others.

B Details of Retrieval Setting

In this work, we use MassiveDS as the retrieval database. MassiveDS is a vast, open-source database comprising approximately 1.4 trillion tokens. The dataset is built from a diverse collection of sources, including large-scale web crawl data as well as domain-specific corpora, to cover a wide variety of topics and writing styles. Each document in MassiveDS is a 256-word chunk of a passage, and was encoded by Contriever. For each query, we retrieve the top 20 documents using the retrieval and reranking pipeline introduced by Shao et al. (2024).

C Prompts Design

We show the two prompts used in this paper below:

Judging generated response.

You are an impartial evaluator tasked with judging the correctness of an answer. Your task: Determine if the model output is semantically and logically consistent with any of the ground truth answers.

If it conveys the same meaning or correct information (even with different wording), mark it as correct.

For yes/no questions, you only need to check whether the final [yes/no] prediction aligns with the ground truth or not.

If the question is about the date, you need to check the consistency of the year, month, and day. It is OK if the model outputs the year only, but it is unacceptable if the generated one does not match the ground truth.

If the model abstains from answering the question (e.g., saying the document does not contain sufficient information to answer the question), mark it as abstain.

Respond ONLY with one of the following JSON objects:
{"verdict": "correct"}
{"verdict": "incorrect"}
{"verdict": "abstain"}

Classifying retrieved document.

You are an objective evidence classifier. Given a user question, a list of possible answers, and a single document, decide whether the document is **relevant**, **distracting**, or **neutral** with respect to answering the query.

- Do NOT produce a chain-of-thought. Provide only the required structured output (JSON, see schema below) and a concise 1-2 sentence rationale (no internal reasoning steps).

	Trivia QA		NQ		Pop QA		Strategy QA	
LLM	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard
Gemma3-27B	150	50	117	83	66	134	155	45
Llama4-17B	158	42	118	82	93	107	140	60
Qwen3-80B	159	41	121	79	152	48	140	60

Table 4: **Statistics of easy and hard queries for each LLM and dataset.**

- Use external/world knowledge only to determine whether a document implicitly supports an answer via ordinary inference (e.g., a fact that implies resolvability, gender, date, etc.). Do not invent or hallucinate facts that are not in the document when justifying the label.
- Follow the definitions and heuristics below exactly.

Required OUTPUT (JSON)

Return a single JSON object with these fields only:

‘‘‘

{
"label": "relevant" | "distracting" | "neutral",
"confidence": 0.00-1.00,
"rationale": "⟨one- or two-sentence justification⟩",
"supporting_spans": ["⟨short excerpt(s) from the document that justify the judgment⟩"],
"inference_type": "direct" | "indirect" | "multi-hop" | "contradiction"
}

‘‘‘

- label: one of relevant, distracting, neutral.
- confidence: numeric 0-1 reflecting how certain the label is (see scoring guidance below).
- rationale: no more than two sentences, explaining why the label was chosen.
- supporting_spans: zero or more short text snippets taken verbatim from the document that most strongly support the label. If none, return [].
- inference_type:
  - direct: the document explicitly states the answer.
  - indirect: the document gives facts that strongly imply the answer.
  - multi-hop: the document provides an intermediate hop (necessary fact) that, combined with other known facts, supports the answer.
  - contradiction: the document asserts facts that contradict the correct answer.

Label definitions & heuristics

RELEVANT

- The correct answer (or parts of answer) directly appeared in the document → set inference_type = "direct".
- Or the document contains facts that clearly support the correct answer, either by single-step inference (set inference_type = "indirect") or by providing a necessary intermediate hop for a multi-hop inference (set inference_type = "multi-hop").
- If the doc contains intermediate facts that are required to get to the final answer (even though the final answer is not present), treat it as relevant (set inference_type = "multi-hop").
- Provide supporting_spans identifying the explicit sentence(s) or fact(s).

DISTRACTING

- The document asserts claims that would lead a reader away from the correct answer (*i.e.*, it contradicts the correct answer or makes claims that support an incorrect candidate). Use inference_type = "contradiction" if it explicitly contradicts.
- Or the document contains plausible but misleading facts that do not support the correct answer and could plausibly be mistaken for support (*e.g.*, plausible but irrelevant facts presented as if they answer the question). Return supporting spans that illustrate the misleading claim.
- Or the document discusses other things that are related to some entities in the query, but does not provide hints for a reader to answer the question.

NEUTRAL

- The document is unrelated to the query.

Confidence scoring guidance

- ≥ 0.90: explicit textual statement of the answer or a clear contradiction/distraction.
- $0.75 - 0.89$: strong indirect support or a strong but not explicit contradiction/distraction.
- $0.55 - 0.74$: moderate evidence (document gives facts that imply the answer but not overwhelmingly).
- $0.30 - 0.54$: weak or partial evidence, or small inconsistency; label should be conservative.
- ≤ 0.29: little or no evidence; use for neutral decisions.

Set a numeric value according to this guidance.

C.1 Human Verification

To validate the usage of LLM-as-a-Judge, we conduct human verification on LLMs' outputs. Specifically, for both response judging and retrieved document classification, we randomly select 50 data points and manually annotate them. We then compute the inter-annotator agreement between a human and LLMs. For response judging, we only observe 1 out of 50 with discrepancy between LLM and the human. For retrieved document classification, LLM achieves a 100% agreement with human in the relevant set. And in the distracting set, we observe that only less than 5% of distracting documents provide indirect information to ground answers. These results justify the data quality and validate the usage of LLM-as-a-Judge in data annotation.

D Example of Retrieved Document

We provide an example to demonstrate the differences between relevant, distracting, and random documents. In this example, the relevant document contains the exact ground truth answer, Jimi Hendrix, with sufficient context. In contrast, the distracting document does not provide any information about Jimi Hendrix, while the mentions of Monterey Pop festival and 1967 make the document have a high semantic similarity to the query.

Query. Who set fire to his guitar at the Monterey Pop festival in 1967?

Relevant. Guitar showmanship to make a big thing of breaking the guitar. I bounced all over the stage with it and I threw the bits on the stage and I picked up my spare guitar and carried on as though I really had meant to do it." Jimi Hendrix sometimes set fire to his guitar, most notably at the Monterey Pop Festival when, apparently, he felt this was the only way he could upstage the destruction by Pete Townshend and Keith Moon of The Who during their set. On March 31, 1967 at performance at London Astoria Hendrix sustained hand burns and visited

Distracting. doing them better. There's gonna be a showdown in California, at Monterey Pop, in July of 1967, where we close out the show. Put on the earbuds, and take a ride on the Magic Bus. Turn it up, smash it up, burn it up! And thanks always for your comments, questions, and reviews! Note: we are using the American release dates and label imprint for all the songs on this episode's playlist.

Random. 1. - Warren Beatty, Splendour in the Grass, 1961. Despite being championed by scenarist and playwright William Inge, Beatty nearly lost his first movie to the blonde cheeesecake. Troy had already taken Parrish from him. 2. - Richard Beymer, West Side Story, 1961. About as dumb an idea as Neil Diamond for Taxi Driver!

E Additional Experimental Results

E.1 Effect of Relevant Documents on Uncertainty

We study how relevant documents affect the uncertainty of generated responses. In particular, we focus on queries that LLMs can answer correctly in both settings of without documents and with relevant documents. We quantify uncertainty with length-normalized log-likelihood among generated tokens (Guerreiro et al., 2023). Specifically, for a generation without retrieved documents, we compute the uncertainty score

$$s_{\text{no\_doc}} = \frac{1}{|\hat{y}|} \sum_{t=1}^{|\hat{y}|} \log(p_\theta(y_t | I, q, y_{<t})). \quad (3)$$

LLM	Trivia QA	NQ	Pop QA	Strategy QA
Gemma3-27B	3.117***	-1.787	-1.274	14.375***
Llama4-17B	5.361***	-2.849	2.135**	8.449***
Qwen3-80B	9.525***	0.973	-0.663	12.216***

Table 5: **Result of paired-sample one-tailed t-test.** We report the t-statistic, where its magnitude indicates how easy we can distinguish the two distributions. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

	Trivia QA		NQ		Pop QA		Strategy QA	
LLM	Acc	AvgDist	Acc	AvgDist	Acc	AvgDist	Acc	AvgDist
Gemma3-27B	1.00 (0.00)	0.78 (0.01)	1.00 (0.00)	0.77 (0.02)	0.99 (0.01)	0.68 (0.01)	1.00 (0.00)	0.73 (0.01)
Llama4-17B	1.00 (0.00)	0.71 (0.01)	1.00 (0.00)	0.70 (0.01)	1.00 (0.00)	0.75 (0.01)	0.99 (0.00)	0.66 (0.01)
Qwen3-80B	0.98 (0.01)	0.54 (0.02)	0.97 (0.00)	0.61 (0.01)	1.00 (0.00)	0.66 (0.02)	0.97 (0.00)	0.44 (0.01)

Table 6: **In S1, the linear probe achieves both high accuracy and large distance.**

	Trivia QA		NQ		Pop QA		Strategy QA	
LLM	Acc	AvgDist	Acc	AvgDist	Acc	AvgDist	Acc	AvgDist
Gemma3-27B	0.99 (0.00)	0.31 (0.01)	1.00 (0.00)	0.37 (0.01)	0.98 (0.01)	0.41 (0.01)	1.00 (0.00)	0.40 (0.01)
Llama4-17B	1.00 (0.00)	0.43 (0.01)	0.95 (0.00)	0.40 (0.00)	1.00 (0.00)	0.53 (0.02)	1.00 (0.00)	0.44 (0.01)
Qwen3-80B	0.74 (0.01)	0.25 (0.01)	0.83 (0.01)	0.28 (0.00)	0.89 (0.01)	0.31 (0.00)	0.97 (0.02)	0.24 (0.01)

Table 7: **In S2, the linear probe achieves a high accuracy but small distance.**

And for the generation with relevant documents, we compute

$$s_{\text{rel}} = \frac{1}{|\hat{y}|} \sum_{t=1}^{|\hat{y}|} \log(p_\theta(y_t | I, q, S_q^{\text{rel}}, y_{<t})). \quad (4)$$

We hypothesize that the generations with relevant documents would have higher log-likelihood than generations without documents, *i.e.*, $s_{\text{rel}} > s_{\text{no\_doc}}$. We perform paired-sample one-tailed t-test to verify this hypothesis. Table 5 shows that t-test rejects the null hypothesis most of the time across datasets and models, suggesting that relevant documents help increase model confidence.

E.2 Representations Separability Analyses

In Section 5, we primarily present our observations through PCA visualizations. To quantitatively substantiate these findings, we conduct linear probe experiments to measure the separability of representations induced by different context types. We report both linear probe accuracy and the average distance to the decision boundary as metrics of separability.

Let $\boldsymbol{w}$ and $\boldsymbol{b}$ denote the learned weight and bias of a linear probe, and let $\{\boldsymbol{x}_i\}_{i=1}^N$ be the representations in the test set. We define the average distance to the decision boundary as

$$d = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{\boldsymbol{w}^\top \boldsymbol{x}_i + \boldsymbol{b}}{\|\boldsymbol{w}\|_2} \right|. \quad (5)$$

Based on the observations in Section 5, we evaluate the following settings:

- **S1. No_doc vs. Random**: High separability is expected (Observation 1).
- **S2. No_doc vs. Relevant**: Low separability is expected (Observation 2).
- **S3. Relevant vs. Relevant + Distracting**: Low separability is expected (Observation 3).
- **S4. Relevant vs. Random in early layer**: Low separability is expected (Observation 4).
- **S5. No_doc vs. Relevant in middle layer**: High separability is expected (Observation 5).

For each setting, we split the data into training and test sets with an 80/20 ratio, apply PCA to reduce representations to 16 dimensions, and ℓ_2-normalize each representation. We train a linear probe ten times using different random seeds and report the mean and standard deviation of each metric across runs.

Table 6 shows that in S1, the linear probe achieves both high accuracy and a large average

	Trivia QA		NQ		Pop QA		Strategy QA	
LLM	Acc	AvgDist	Acc	AvgDist	Acc	AvgDist	Acc	AvgDist
Gemma3-27B	0.66 (0.01)	0.22 (0.01)	0.62 (0.01)	0.21 (0.01)	0.65 (0.01)	0.20 (0.01)	0.98 (0.02)	0.26 (0.00)
Llama4-17B	0.70 (0.01)	0.20 (0.00)	0.70 (0.01)	0.24 (0.01)	0.74 (0.03)	0.21 (0.01)	0.93 (0.01)	0.18 (0.00)
Qwen3-80B	0.60 (0.01)	0.26 (0.00)	0.46 (0.02)	0.21 (0.01)	0.67 (0.01)	0.27 (0.02)	0.51 (0.01)	0.21 (0.00)

Table 8: **In S3, the linear probe achieves both low accuracy and small distance.**

	Trivia QA		NQ		Pop QA		Strategy QA	
LLM	Acc	AvgDist	Acc	AvgDist	Acc	AvgDist	Acc	AvgDist
Gemma3-27B	0.49 (0.01)	0.18 (0.01)	0.52 (0.01)	0.21 (0.00)	0.63 (0.01)	0.35 (0.01)	0.50 (0.02)	0.25 (0.01)
Llama4-17B	0.62 (0.01)	0.13 (0.00)	0.70 (0.01)	0.14 (0.01)	0.64 (0.01)	0.24 (0.01)	0.68 (0.01)	0.16 (0.00)
Qwen3-80B	0.64 (0.02)	0.14 (0.00)	0.76 (0.02)	0.18 (0.00)	0.77 (0.02)	0.25 (0.00)	0.71 (0.01)	0.21 (0.00)

Table 9: **In S4, the linear probe achieves both low accuracy and small distance.**

	Trivia QA		NQ		Pop QA		Strategy QA	
LLM	Acc	AvgDist	Acc	AvgDist	Acc	AvgDist	Acc	AvgDist
Gemma3-27B	1.00 (0.00)	0.54 (0.01)	1.00 (0.00)	0.56 (0.00)	1.00 (0.00)	0.73 (0.00)	1.00 (0.00)	0.57 (0.00)
Llama4-17B	0.99 (0.01)	0.57 (0.01)	1.00 (0.00)	0.46 (0.01)	1.00 (0.00)	0.61 (0.01)	1.00 (0.00)	0.46 (0.01)
Qwen3-80B	1.00 (0.00)	0.44 (0.00)	1.00 (0.00)	0.48 (0.01)	1.00 (0.00)	0.59 (0.02)	1.00 (0.00)	0.50 (0.01)

Table 10: **In S5, the linear probe achieves both high accuracy and large distance.**

distance, indicating strong separability between representations without documents and those with random documents, which supports Observation 1. In Table 7, although the probe also attains high accuracy in S2, the average distance is substantially smaller than in S1. This suggests that representations without documents and with relevant documents are separable but remain close in the latent space, validating Observation 2. Tables 8 and 9 show that in both S3 and S4, the probe exhibits low accuracy and small average distance, indicating poor separability and supporting Observations 3 and 4. Finally, Table 10 shows that in S5, the probe again achieves high accuracy with a large average distance, confirming Observation 5 that in middle layers, representations without documents and with relevant documents are strongly separated.

E.3 Additional Figures and Tables

This section contains additional figures and tables that complement the results discussed in Section 5. Due to page limitations, these materials are presented here rather than in the main text.

Observation 1 & 2. See Figure 7, 8, and Table 11.

Observation 3. See Figure 9.

Observation 4 & 5. See Figure 10 to 14.

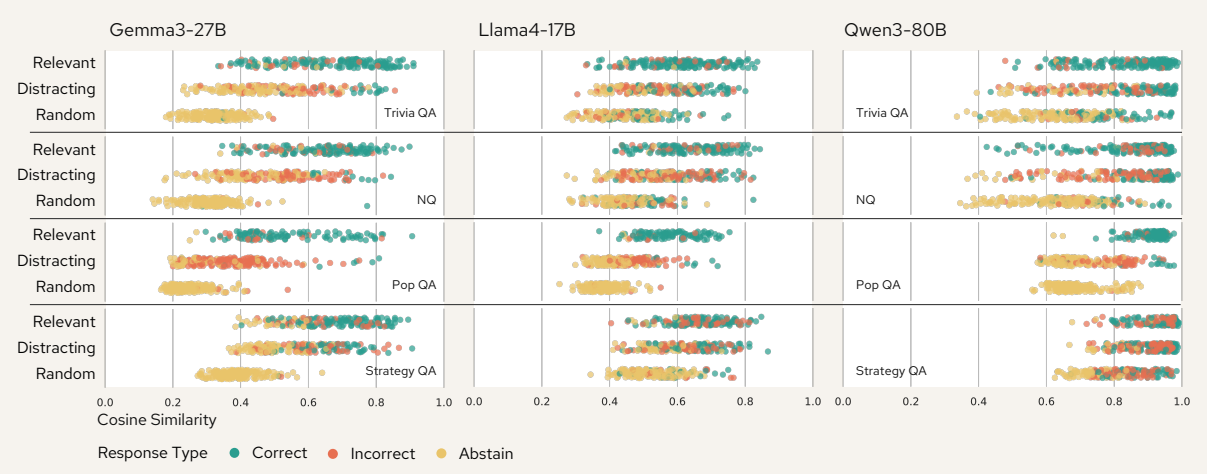

Figure 7: **The complete result of Figure 3—relationship between cosine similarity and response type.** For each context type, we compute the cosine similarity between the representations of with-context prompts and query, and categorize responses as correct, incorrect, or abstain. The result shows that LLMs are more likely to abstain when context induces large representation shifts.

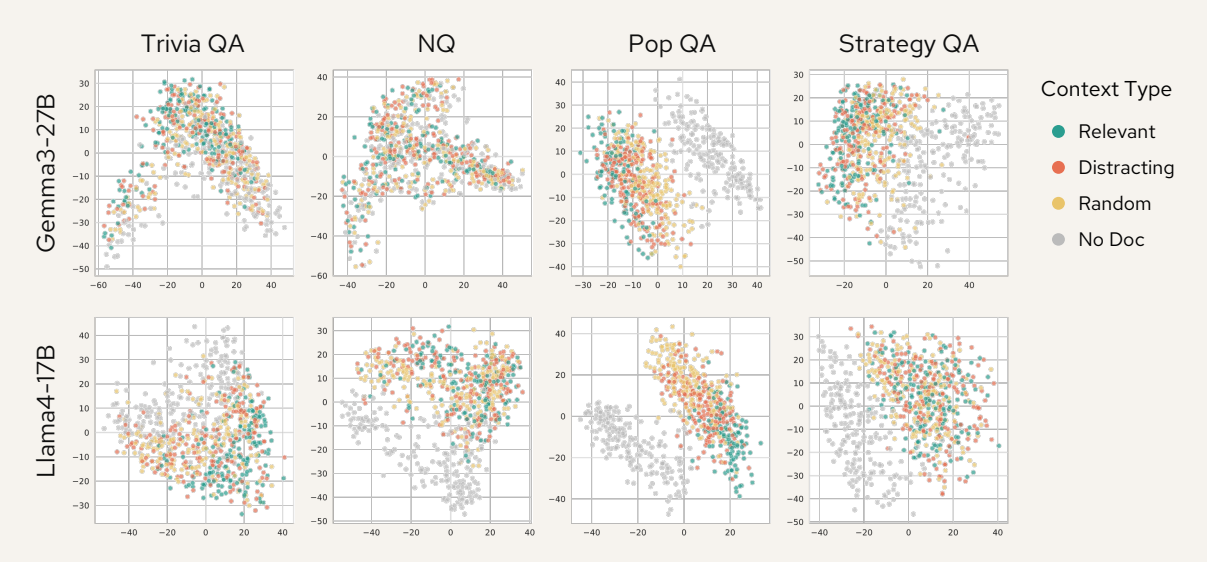

Figure 8: **The complete result of Figure 4—PCA of base models.** We apply PCA on representations of base models. The result shows that base LLMs do not have representation drift across different context types. Note that we only show the result of Gemma3-27B and Llama4-17B as Qwen3-80B did not release its base model.

	Trivia QA						NQ						Pop QA						Strategy QA					
	Easy			Hard			Easy			Hard			Easy			Hard			Easy			Hard		
Context	Cor	Inc	Abs	Cor	Inc	Abs	Cor	Inc	Abs	Cor	Inc	Abs	Cor	Inc	Abs	Cor	Inc	Abs	Cor	Inc	Abs	Cor	Inc	Abs
Gemma3-27B																								
Base																								
Relevant	92.4	4.2	3.4	52.5	35.6	11.9	81.9	10.2	7.9	43.9	39.5	16.6	92.4	2.0	5.6	76.7	20.4	2.9	74.5	16.4	9.1	44.6	42.2	13.2
Distracting	79.5	15.8	4.7	14.8	72.4	12.8	65.7	25.5	8.8	12.4	70.7	16.9	46.2	48.1	5.7	5.2	86.4	8.4	62.0	27.6	10.4	29.8	52.2	18.0
Random	89.4	6.8	3.8	15.6	65.4	19.0	66.5	12.2	21.3	11.1	53.1	35.8	71.5	24.2	4.3	9.7	85.4	4.9	65.6	20.9	13.5	23.4	55.8	20.8
Instruction-tuned																								
Relevant	90.4	6.5	3.1	65.2	27.8	7.0	79.8	15.5	4.7	62.1	29.0	8.9	91.0	7.4	1.6	70.5	29.5	0	72.6	11.6	15.8	44.3	35.4	20.3
Distracting	8.5	29.7	61.8	0.7	25.1	74.2	8.0	39.0	53.0	0.6	34.3	65.1	3.6	52.7	43.7	0.6	61.6	37.8	22.7	19.7	57.6	8.6	23.3	68.1
Random	1.7	0.7	97.6	0	1.9	98.1	2.2	0.3	97.5	0.4	2.2	97.4	0.2	5.1	94.7	0.1	7.3	92.6	1.2	0.5	98.3	0	4.3	95.7
Llama4-17B																								
Base																								
Relevant	93.5	6.2	0.3	67.2	32.2	0.6	84.5	15.4	0.1	58.7	40.9	0.4	95.4	4.6	0	79.6	20.4	0	89.1	10.9	0	34.1	65.5	0.34
Distracting	67.7	31.1	1.2	10.9	84.1	5.0	47.5	48.3	4.2	8.6	84.9	6.5	45.2	54.3	0.5	9.2	90.1	0.7	70.5	28.7	0.8	20.7	78.9	0.4
Random	91.0	8.0	1.0	7.6	86.4	6.0	76.4	19.5	4.1	15.6	73.0	11.4	74.0	25.5	0.5	20.8	78.7	0.5	74.1	24.9	1.0	20.8	78.6	0.6
Instruction-tuned																								
Relevant	89.7	8.5	1.8	62.8	30.3	6.9	85.9	12.8	1.3	64.0	30.6	5.4	92.1	6.5	1.4	79.8	15.5	4.7	76.3	17.1	6.6	35.3	56.2	8.5
Distracting	34.6	26.4	39.0	11.4	37.0	51.6	33.5	34.5	32.0	5.1	44.3	50.6	2.5	26.1	71.4	0.5	28.0	71.5	38.5	24.8	36.7	13.2	45.6	41.2
Random	38.7	2.5	58.8	8.7	12.1	79.2	34.5	5.3	60.2	4.0	15.5	80.5	0.4	0.8	98.8	0	0.9	99.1	13.8	6.1	80.1	8.2	10.9	80.9

Table 11: **The complete result of Table 1** We report the percentage of correct (Cor), incorrect (Inc), and abstain (Abs) responses for both base and instruction-tuned LLMs. The result shows that instruction-tuned LLMs tend to abstain when the retrieval document is distracting or random, even if they can answer with the query alone. Note that we only show the result of Gemma3-27B and Llama4-17B as Qwen3-80B did not release its base model.

Figure 9: **The complete result of Figure 5—PCA of multiple-document contexts.** We perform PCA on the last prompt token representations for multiple-document contexts that include one relevant document, and plot them alongside representations with only a relevant document in 2D. The result shows that representations remain similar when a relevant document is present, regardless of other context.

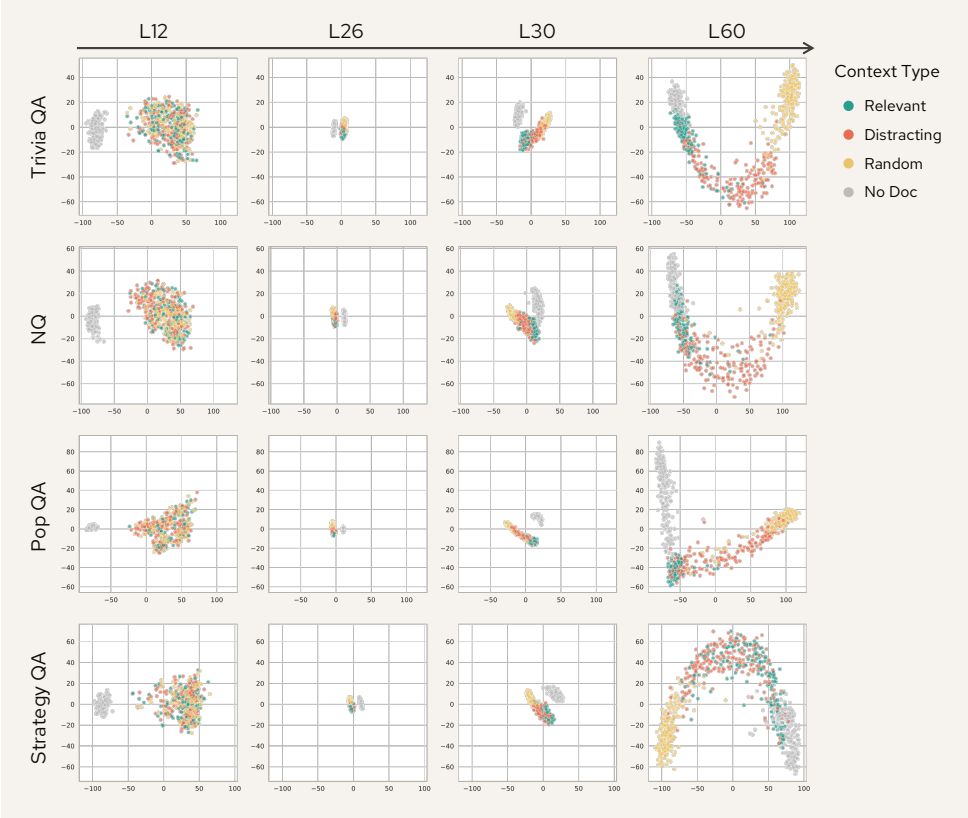

Figure 10: **The complete result of Figure 6—evolution of Gemma3-27B representations.** We perform PCA on the last prompt token representations of Gemma3-27B across different layers and plot them in 2D.

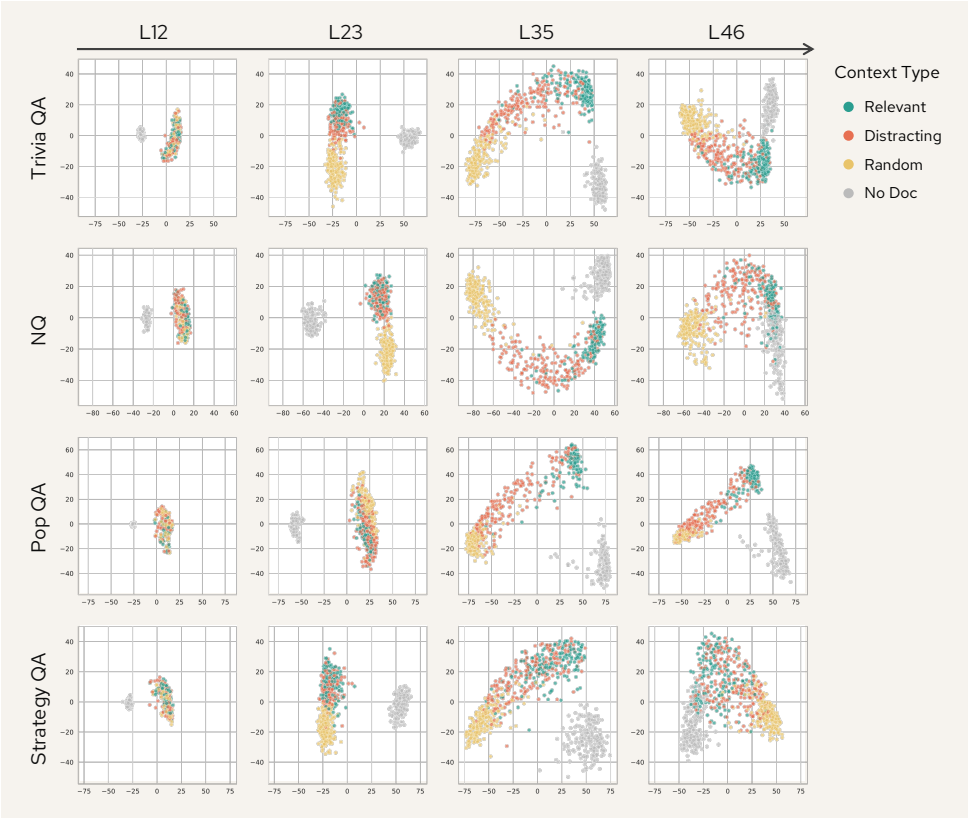

Figure 11: **The complete result of Figure 6—evolution of Llama4-17B representations.** We perform PCA on the last prompt token representations of Llama4-17B across different layers and plot them in 2D.

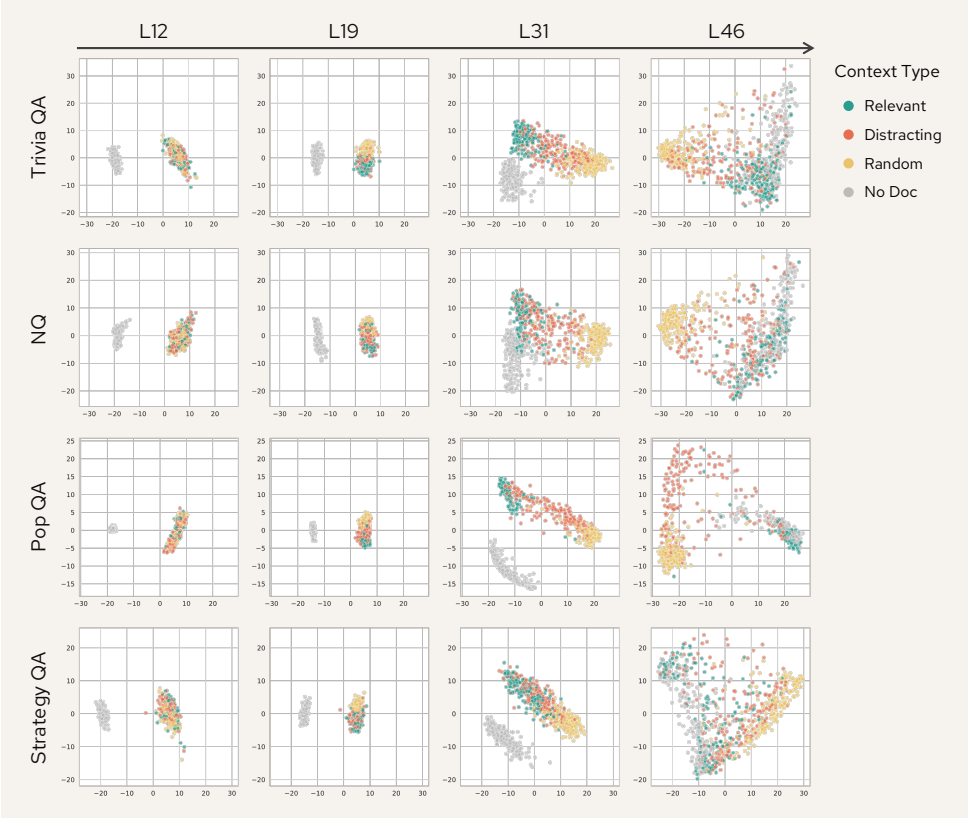

Figure 12: **The complete result of Figure 6—evolution of Qwen3-80B representations.** We perform PCA on the last prompt token representations of Qwen3-80B across different layers and plot them in 2D.

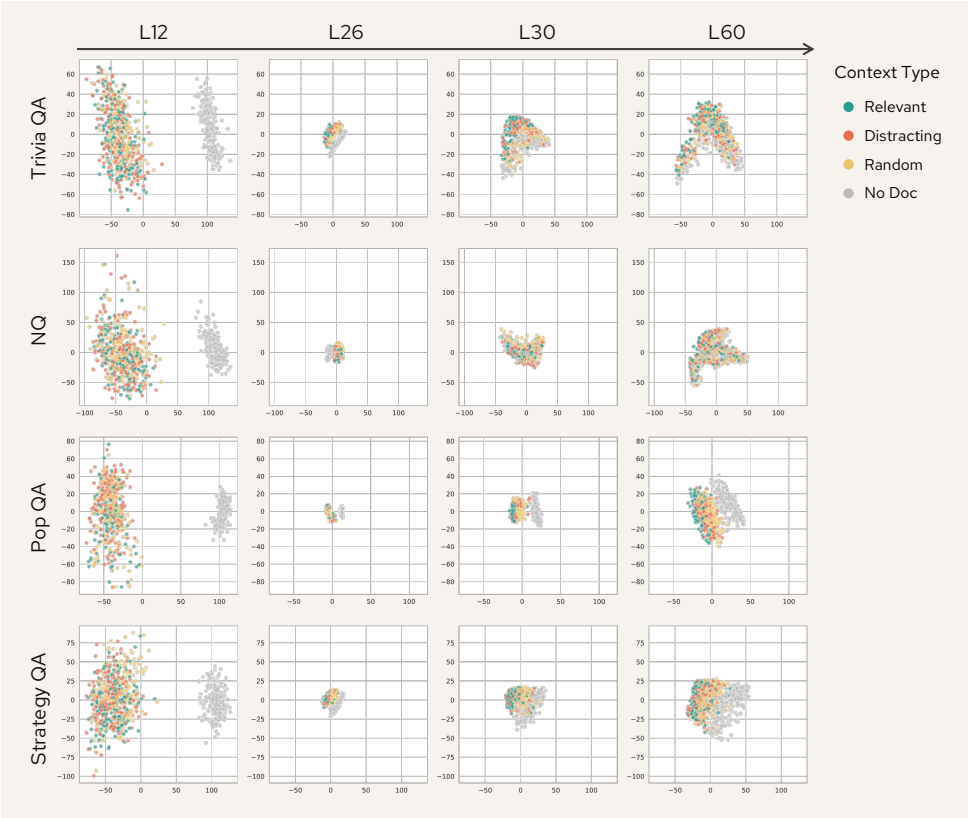

Figure 13: **Evolution of Gemma3-27B-base representations.** We perform PCA on the last prompt token representations of Gemma3-27B-base across different layers and document types and plot them in 2D.

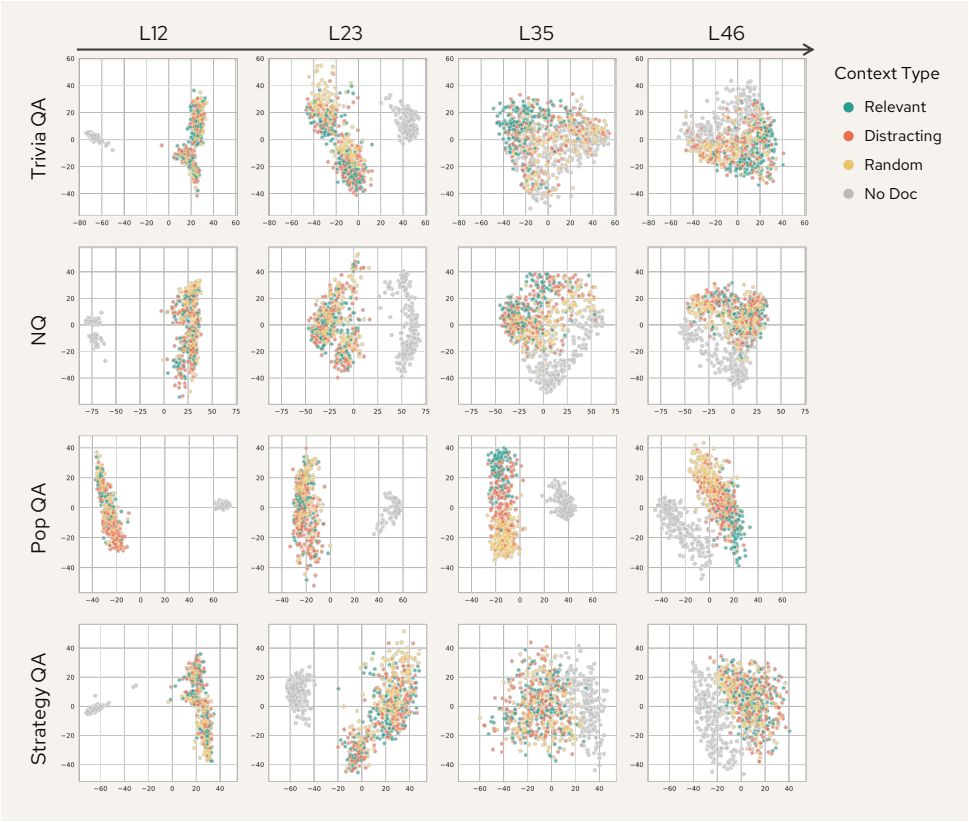

Figure 14: **Evolution of Llama4-17B-base representations.** We perform PCA on the last prompt token representations of Llama4-17B-base across different layers and document types and plot them in 2D.