Rhetorical Questions in LLM Representations: A Linear Probing Study

Louie Hong Yao[1], **Vishesh Anand**[2], **Yuan Zhuang**[3], **Tianyu Jiang**[2]
[1]Independent Researcher [2]University of Cincinnati [3]Amazon
lhyao731@gmail.com, anandvh@mail.uc.edu, zyone@amazon.com, tianyu.jiang@uc.edu

Abstract

Rhetorical questions are asked not to seek information but to persuade or signal stance. How large language models internally represent them remains unclear. We analyze rhetorical questions in LLM representations using linear probes on two social-media datasets with different discourse contexts, and find that rhetorical signals emerge early and are most stably captured by last-token representations. Rhetorical questions are linearly separable from information-seeking questions within datasets, and remain detectable under cross-dataset transfer, reaching AUROC around 0.7-0.8. *However, we demonstrate that transferability does not simply imply a shared representation.* Probes trained on different datasets produce different rankings when applied to the same target corpus, with overlap among the top-ranked instances often below 0.2. Qualitative analysis shows that these divergences correspond to distinct rhetorical phenomena: some probes capture discourse-level rhetorical stance embedded in extended argumentation, while others emphasize localized, syntax-driven interrogative acts. Together, these findings suggest that rhetorical questions in LLM representations are encoded by multiple linear directions emphasizing different cues, rather than a single shared direction.

1 Introduction

Rhetorical language is a common and important part of everyday communication. One of its most typical forms is the rhetorical question, which speakers use not to seek information, but to persuade, challenge, or signal stance. For example, the question "Do you really believe that?" is often used to express doubt rather than to elicit an answer, and "Who would ever agree with this?" functions as a critique rather than a genuine inquiry. In contrast to domains involving formal problem-solving or factual inquiry, which rely on literal interpretation, rhetorical questions convey non-literal meaning shaped by context, speaker intent, and discourse structure. As a result, understanding rhetorical questions provides an important perspective on how large language models interpret and generate language in real-world communicative settings.

While rhetorical and informational questions have been studied computationally for many years (Bhattasali et al., 2015; Oraby et al., 2017; Zhuang and Riloff, 2020; Kikteva et al., 2024), how large language models internally represent rhetorical questions has received far less attention. Existing work on rhetorical question understanding has largely framed the problem as a question-answering or classification task with explicit labels (Ikumariegbe et al., 2025). While convenient, this formulation focuses on prediction accuracy, whereas our work takes a step forward in understanding the representational basis of rhetorical behavior within the model.

To achieve this, we adopt a perspective that examines how these models encode rhetorical question within their internal *representational space*. Using multiple linear probes (Park et al., 2023), we investigate where rhetorical signals emerge, whether probes learned within the same dataset capture similar structure, and how probes learned from different data sources compare and transfer across datasets. We find that rhetorical content is linearly separable from information-seeking questions within a given context and remains detectable under cross-dataset transfer. However, probing directions differ substantially within and across datasets. These differences reflect distinct rhetorical properties and lead to different rankings on the same corpus. *This shows that strong probing accuracy or transferability does not imply that a property is captured by a single shared representational direction.* Instead, rhetorical questions are not organized along a single linear axis, but reflect multiple linguistic features that are emphasized differently depending on context and data. Our code

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 155–172

is publicly available.[1] We summarize our contributions as follows:

1. We present a systematic, data-driven analysis of rhetorical question in LLM representations across real-world contexts, showing that rhetorical signals emerge early and are most stably captured by last-token representations in decoder-only models.
2. We reveal a divergence between discriminative performance and representational alignment: although rhetorical questions are consistently linearly separable and transferable, probes learned from different data distributions induce substantially different rankings with little overlap in the upper and lower ends of the ranking.
3. Through qualitative analysis, we demonstrate that rhetorical meaning is inherently heterogeneous, spanning discourse-level rhetorical stance and localized, syntax-driven interrogative acts rather than a single, unified representational dimension.

2 Related Work

Rhetorical Questions. Rhetorical questions have long been studied in linguistics and NLP, with early work focusing on their pragmatic and discourse functions (Jurafsky et al., 1997; Frank, 1990; Roberts and Kreuz, 1994; Han, 2002; Špago, 2016). More recent computational studies frame rhetorical questions as a classification or detection task using linguistic features and contextual cues (Bhattasali et al., 2015; Oraby et al., 2017; Zhuang and Riloff, 2020; Kikteva et al., 2024). Recently, Ikumariegbe et al. (2025) examined rhetorical questions in the context of large language models using QA-style judgments of whether a question in context is rhetorical or informational.

Beyond rhetorical questions specifically, recent work has examined rhetorical behavior in LLMs at broader stylistic or strategic levels. Qiu et al. (2025) introduce a counterfactual framework for measuring rhetorical style independently of substantive content, and Ji et al. (2025) develop LLM-based models for annotating rhetorical strategies across domains. In a different vein, Reinhart et al. (2025) analyze grammatical and rhetorical variation in LLM-generated text at the level of model outputs. By contrast, our work focuses on rhetorical questions at the level of internal representations, examining how rhetorical intent is encoded, organized, and transferred across layers and datasets.

Related work on sarcasm and irony (Zhang et al., 2025; Lee et al., 2025) likewise addresses non-literal language, but centers on distinct pragmatic phenomena and does not directly examine the representation of rhetorical question intent in LLMs.

Interpretability of Representations. Linear probing, which evaluates whether a target property is linearly decodable from a model's internal representations, has long been used to analyze neural networks (Alain and Bengio, 2017). A broad line of work studies neural language models through interpretability and probing methods, using linear probes as diagnostic tools to assess whether linguistic or semantic properties are accessible from model representations. Recent work emphasizes population-level and training-free approaches, such as diffMean directions (Marks and Tegmark, 2024; Vennemeyer et al., 2025), as well as sparse autoencoders (Cunningham et al., 2024; Gao et al., 2025; Farnik et al., 2025; Heap et al., 2025) that linearly decompose activations into interpretable feature directions. Complementary geometric and information-theoretic analyses have also been applied to study representation spaces in LLMs (Hosseini and Fedorenko, 2023; Skean et al., 2025). Related work has also increasingly explored causal intervention-based methods (Meng et al., 2022; Ghandeharioun et al., 2024), such as patching and model editing, to test how specific internal states contribute to downstream behavior. Our work builds on this representation-centric literature by applying linear probing and geometric analysis to rhetorical question intent across contexts.

3 Methods

In this section, we describe the datasets, representation choices, and linear probing framework used in our analysis.

3.1 Datasets

We conduct our analysis on two real-world rhetorical question datasets drawn from social media, which differ in domain, annotation protocol, and availability of contextual information.

RQ. The RQ dataset introduced by Zhuang and Riloff (2020) consists of Twitter questions annotated as rhetorical or informational. The dataset

[1] https://github.com/ruyi101/rq-representation-probing

contains 4,997 instances, with 2,332 labeled as rhetorical, and is split into 3,200 training, 797 validation, and 1,000 test samples. Each instance includes a target question and a prior tweet providing conversational context. In the main analysis, we focus on the *question-with-context* formulation, which concatenates the prior tweet and the target question. Under this formulation, instances average 38.9 tokens.[2]

SRAQ. The SRAQ dataset proposed by Ikumariegbe et al. (2025) draws from Reddit conversation threads, where each example is built around a target question found within a single user's comment. Because a comment can span multiple paragraphs, the dataset provides context at different levels of granularity. We consider two: the *Paragraph* formulation, which retains only the paragraph containing the target question, and the *Full_turn* formulation, which contains the entire comment. All instances are annotated as rhetorical or informational. The dataset includes 971 samples, split into 384 training, 103 validation, and 484 test instances, with 609 questions labeled as rhetorical. In the main paper, we use the *Paragraph* formulation, which averages 58.5 tokens per instance. Results using the *Full_turn* formulation show similar trends and are reported in the Appendix G.

Across both datasets, we use the original train-validation-test splits provided by the respective authors and do not modify the annotation labels.

3.2 Representations

Given an input sequence of tokens $\{x_1, \ldots, x_T\}$, we extract hidden representations from a pretrained language model. Unless otherwise specified, we use the *last-token representation* $h_T \in \mathbb{R}^d$, which is commonly used as a sequence-level summary in decoder-style models.

To examine the effect of token aggregation, we also consider mean-pooled representations

$$\bar{h} = \frac{1}{T} \sum_{t=1}^{T} h_t, \tag{1}$$

where h_t denotes the hidden state at token position t.

For fair comparison across probes, we primarily project representations into a PCA space with $k = 64$ dimensions, defined separately for each dataset and model, and for each input formulation (with and without additional context). This choice is motivated by the widely adopted view that high-dimensional language model representations concentrate near a low-dimensional manifold, such that most task-relevant variation is captured by a relatively small number of principal components (Skean et al., 2025). Projecting into a shared low-dimensional subspace reduces noise and stabilizes comparisons across probes while preserving the dominant structure of the representations.

Concretely, for a fixed dataset–model–input setting (e.g., paragraph vs. full_turn), a single PCA transformation is applied to all examples, and all probes within that setting operate in the same projected space. Different datasets, models, or input formulations use different PCA projections. In addition to this shared PCA space, we also report selected results computed directly in the original embedding space, and verify that key diffMean trends remain similar (Appendix A).

3.3 Linear Probes

We analyze rhetorical separability using three linear probes applied to fixed representations: a training-free population-level diffMean probe (Marks and Tegmark, 2024), and two trained discriminative probes based on logistic regression and linear support vector machine.

DiffMean. The diffMean probe estimates a population-level direction by subtracting class-conditional means. Let $\mathcal{D}_+$ and $\mathcal{D}_-$ denote rhetorical and informational examples, respectively. The direction is

$$w_{\text{DM}} = \mu_+ - \mu_-, \tag{2}$$

where $\mu_\pm = \mathbb{E}_{x \in \mathcal{D}_\pm}[h(x)]$. Each example is scored by the inner product $w_{\text{DM}}^\top h(x)$. Larger values indicate stronger alignment of the representation $h(x)$ with the diffMean direction.

Discriminative Linear Probes. We additionally train linear classifiers using logistic regression and hinge loss. Logistic regression learns a weight vector by minimizing the cross-entropy loss, which encourages probabilistic separation between rhetorical and informational examples. The hinge-loss probe, on the other hand, optimizes a margin-based objective corresponding to a linear support vector machine. In both cases, the learned weight vector w defines a linear scoring function $w^\top h(x)$. Larger scores indicate stronger alignment with the learned separator, analogous to diffMean. The two probes

[2] All token counts for both RQ and SRAQ use the Llama-3.3-70B-Instruct tokenizer.

differ only in their optimization objective while sharing the same linear hypothesis.

In our experiments, the diffMean direction is computed using the training split only, and the discriminative probes are trained on the training split and selected using validation performance. Unless otherwise noted, all results below are reported on the test split.

3.4 Evaluation Metrics

We use multiple evaluation metrics to characterize rhetorical separability and to compare the behavior of different linear probes. AUROC is used to evaluate each probe direction individually, measuring how well its scores separate rhetorical from informational questions on held-out data. To compare probes across datasets or objectives, we analyze both the similarity between probe directions and the agreement between their induced orderings.

Rank Agreement. Let $s_i^{(a)}$ and $s_i^{(b)}$ denote the scores assigned to example i by two probes (or by the same probe trained on two datasets). Ranking agreement is measured using the *Spearman's rank correlation*:

$$\rho_s = \mathrm{corr}\Big(\mathrm{rank}(s^{(a)}),\, \mathrm{rank}(s^{(b)})\Big), \qquad (3)$$

where $\mathrm{rank}(\cdot)$ maps scores to ranks and $\mathrm{corr}(\cdot,\cdot)$ is the Pearson correlation applied to the rank vectors.

Overlap at the tails. To assess agreement in the upper and lower ends of the ranking, for a fraction $p \in (0,1)$ we form the top-p (or bottom-p) sets A_p and B_p under each scoring function, and compute the *Jaccard index*:

$$J(A_p, B_p) = \frac{|A_p \cap B_p|}{|A_p \cup B_p|}. \qquad (4)$$

These metrics quantify whether different probes induce similar orderings globally (ρ_s) and whether they retrieve similar top- and bottom-ranked examples (J).

4 Representation Choices for Rhetorical Probing

We first examine how representation choice affects rhetorical probing in decoder-only models.[3] In the main text, we focus on two sequence-level representations: the final-token representation and mean pooling over all tokens. These choices reflect two contrasting assumptions about where rhetorical intent is encoded, namely whether it is concentrated near the end of the sequence or distributed across contextual cues. We therefore compare last-token and mean-pooled representations as our primary sequence-level views. For completeness, we also explore more fine-grained pooling variants, including pooling over the last few tokens and over the question span alone, and report these supplementary results in Appendix D.

Figure 1 reports layer-wise test-set AUROC for rhetorical probing using PCA-reduced embeddings with 64 components, across two datasets (RQ and SRAQ) and two decoder-only models: Qwen3-32B (Yang et al., 2025) and Llama-3.3-70B (Grattafiori et al., 2024). Results are shown for mean-pooled and last-token representations. For each representation, we evaluate three linear probing directions: the diffMean vector, a hinge-loss classifier, and a logistic classifier.

DiffMean. For the diffMean direction, mean-pooled representations achieve AUROC comparable to last-token representations at early layers ($\approx 0.60 - 0.65$). One plausible explanation is that mean pooling aggregates information from multiple tokens, making contextual signals available earlier, while last-token representations require deeper layers to accumulate similar context. At deeper layers, last-token representations generally achieve higher AUROC (≈ 0.8), suggesting that mean pooling is less effective because it aggregates token representations that are less informative for the rhetorical distinction.

Discriminative Probes. Hinge and logistic classifiers typically outperform diffMean across layers, with AUROC values around 0.85–0.9 on RQ and 0.8–0.85 on SRAQ for last-token representations, suggesting that learned linear decision boundaries better leverage the available representations. However, for both classifiers, mean-pooled representations do not provide an advantage over last-token representations. In particular, hinge and logistic probes applied to mean-pooled embeddings fail to surpass their counterparts operating on last-token embeddings, suggesting that, at the linear level with 64 PCA components, aggregating information across all tokens offers no additional benefit beyond the final token representation.

Dataset-dependent effects remain evident when using mean-pooled representations. While results on RQ remain relatively smooth across layers, per-

[3] Our main analysis focuses on decoder-only models because they are the dominant setting for interactive language use. For completeness, we include a brief comparison with an encoder-based model in Appendix E.

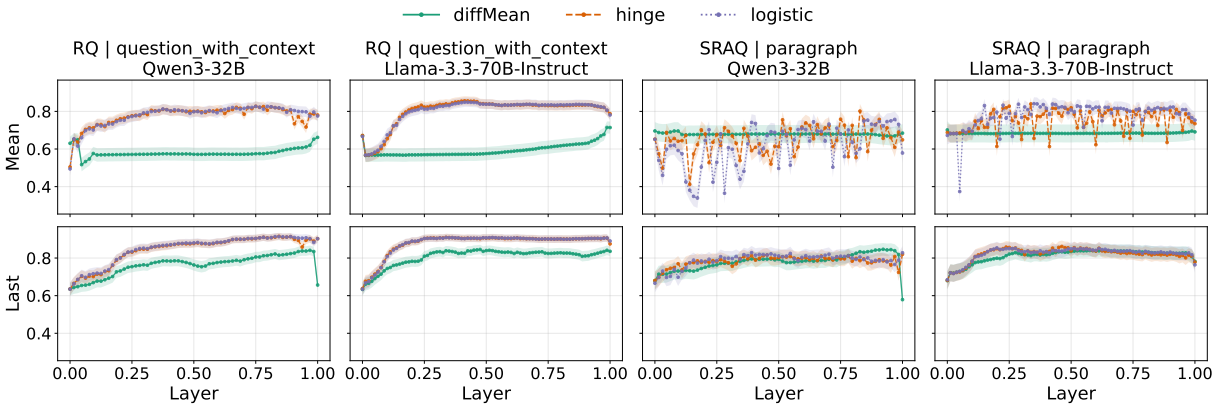


Figure 1: AUROC across layers and representations. Test AUROC across normalized layers using PCA-reduced representations. Rows compare mean-pooled and last-token embeddings; columns vary by dataset (RQ, SRAQ) and model (Qwen3-32B, Llama-3.3-70B).

formance on SRAQ is markedly more unstable across layers and across probing methods. This pattern suggests that, for more context-rich inputs, averaging across tokens can dilute rhetorical signals throughout the model.

Overall, mean pooling appears to retain useful lexical information at early layers, but its probing performance becomes noticeably less stable, particularly on SRAQ. Because mean-pooled representations neither outperform last-token representations under linear probes with 64 PCA components nor offer comparable stability, we focus the remainder of our analysis on last-token representations.

5 Rhetorical Separability Across Linear Probes

Having established last-token representations as a stable choice, we next examine whether rhetorical questions are linearly separable and how different linear probes capture this structure.

AUROC. Figure 1 (second row) shows the AUROC achieved by different linear probes across layers and datasets. On RQ, hinge and logistic classifiers consistently achieve higher AUROC than the diffMean direction across a broad range of layers, with peak values approaching 0.9 at intermediate depths. This behavior is not surprising, as hinge and logistic probes are trained to optimize class separation, whereas diffMean is a training-free baseline. Across layers, hinge and logistic classifiers behave similarly, with no systematic difference in their peak performance.

In contrast, SRAQ shows a different pattern. All probes achieve AUROC well above chance. However, hinge and logistic classifiers provide only small improvements over the diffMean direction. Their performance closely tracks the training-free baseline across layers. Compared to RQ, the gap between trained probes and diffMean is much smaller.

Overall, AUROC results indicate that rhetorical questions are linearly separable from informational questions across models and datasets, though separability is not perfect. Appendix F provides a steering analysis showing that perturbations along learned probing directions induce coherent changes in rhetorical behavior, consistent with these directions capturing a rhetorical signal.

Probing Alignment. Now we examine the alignment between probing directions and the rankings they induce across layers to better understand the remaining gaps. Figure 2 shows the relationships among linear probing directions across layers, models, and datasets. Across all settings, hinge and logistic probes are nearly identical. Their cosine similarity remains close to 1 across layers, with near-perfect Spearman correlation and high Jaccard overlap of top- and bottom-ranked examples. Despite different loss functions, both probes recover essentially the same linear direction and induce highly similar rankings.

In contrast, the relationship between diffMean and the trained probes is weaker and varies across datasets and layers. On RQ, diffMean shows moderate cosine similarity with hinge and logistic probes, mostly below 0.7, especially at intermediate and later layers. Spearman correlations range between 0.6 and 0.8, indicating only partial agreement in the induced rankings. This alignment is consistently lower and more layer-dependent than

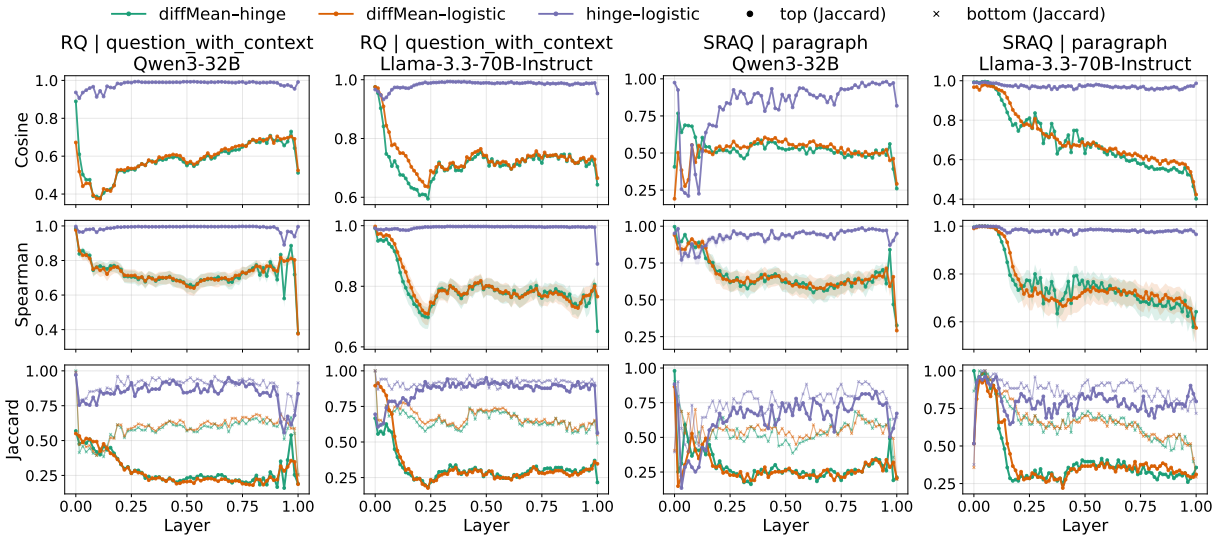


Figure 2: Alignment and ordering agreement between linear probes. Top: Cosine similarity between probing directions across layers. Middle: Spearman rank correlation between probe-induced scores, with shaded regions indicating confidence intervals. Bottom: Jaccard overlap between the top and bottom 20% of examples ranked by each probe. Results are shown across datasets (RQ, SRAQ), models (Qwen3-32B, Llama-3.3-70B-Instruct), and layers.

that observed between hinge and logistic probes.

On SRAQ, divergence is stronger: cosine similarity between diffMean and trained probes is lower and less stable (≈ 0.5), with reduced rank agreement (Spearman ≈ 0.6 at middle to late layers).

This pattern is also reflected in the Jaccard overlap of highest- and lowest-ranked examples, shown in the third row of Figure 2. Across models and datasets, overlap among top-ranked examples is low, around 0.25, while overlap among bottom-ranked examples is higher, typically around 0.5. This asymmetry suggests that informational instances are ranked more consistently across probes, whereas highly rhetorical instances show greater variability. In this sense, rhetorical status appears more heterogeneous than informational status.

This asymmetry helps explain the AUROC results on SRAQ. Probes with similar AUROC can induce divergent rankings and little overlap in their top- and bottom-ranked examples, hinting that comparable discriminative performance does not necessarily imply shared representational properties. Instead, probes with similar AUROC may capture different aspects of rhetorical signal.

6 Transferability of Rhetorical Separators

So far, our analysis has focused on within-dataset settings, where probe directions are learned and evaluated on the same data distribution. If rhetorical intent corresponded to a robust linear structure in representation space, such structure should generalize under distributional shift. We therefore examine cross-dataset transfer by learning probe directions on one dataset and applying them to representations from the other. We evaluate both separability and agreement relative to probes learned directly on the target dataset. Because probes are learned in dataset-specific PCA subspaces, we map directions back to the full embedding space prior to comparison (see Appendix B).

AUROC under transfer. We first examine cross-dataset transfer in terms of discriminative performance. As shown in the first row of Figure 3, probe directions learned on one dataset exhibit a modest drop in AUROC when applied to the other, but still achieve values around 0.7–0.8 across models and layers. This indicates that rhetorical intent contains a partially shared linear component across datasets.

Ranking agreement. In contrast to AUROC, ranking agreement under transfer is much weaker. For each target dataset, we score the same examples using both the in-domain direction and the transferred direction, rank the examples by their projection scores, and compute Spearman correlation between the two resulting rankings. As shown in the second row of Figure 3, these correlations are only moderate overall. For trained probes, they often fall toward 0.5 or lower at deeper layers. The diffMean directions follow a similar trend, with

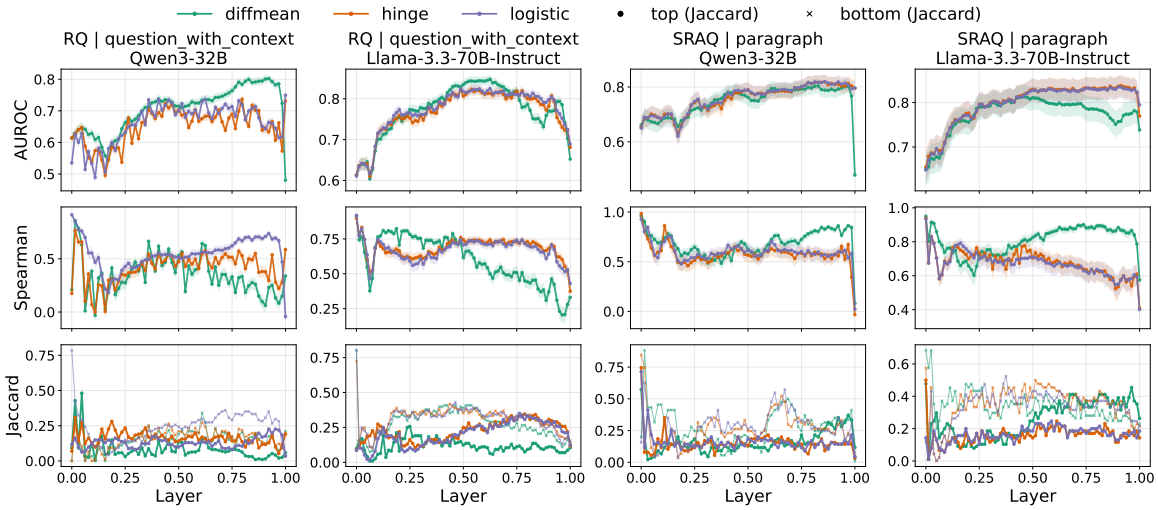


Figure 3: Transferability of rhetorical probing directions across datasets. For RQ panels (left), directions are learned on SRAQ and applied to RQ; for SRAQ panels (right), directions are learned on RQ and applied to SRAQ. Rows report test AUROC of transferred directions (top), Spearman rank correlation between rankings induced by transferred directions and rankings induced by directions learned on the target dataset (middle), and Jaccard overlap between the top and bottom 20% of examples under these two rankings (bottom).

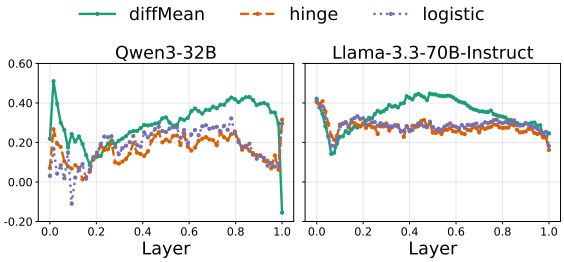


Figure 4: Cross-dataset alignment of rhetorical probing directions. Cosine similarity between probing directions learned on RQ and SRAQ, measured layer-wise for two models and three different probings.

some deviations across datasets and layers, but they likewise do not show strong agreement. This shows that transferability in AUROC does not imply close alignment in the induced rankings.

Agreement among top- and bottom-ranked examples deteriorates further under transfer. As shown in the third row of Figure 3, Jaccard overlap between the top and bottom 20% of examples is consistently low, particularly for top-ranked instances, where it often falls below 0.2. Overlap among bottom-ranked examples is higher but remains limited, around 0.3–0.4. As in the within-dataset setting, this asymmetry suggests that informational instances are identified more consistently across datasets, whereas highly rhetorical instances are more heterogeneous and dataset-dependent.

Directional alignment. To better understand these transfer behaviors, we analyze *directional alignment under transfer*. Figure 4 reports cosine similarity between probe directions learned on RQ and SRAQ across layers. Across models, all probes show low similarity, typically around 0.2–0.4, indicating that directions learned on different datasets are not collinear. Hinge and logistic probes exhibit relatively stable alignment across layers. In contrast, the diffMean direction is more variable, reaching higher similarity at intermediate layers but remaining limited to around 0.4 for Llama-3.3-70B-Instruct and at later layers for Qwen3-32B, with lower alignment elsewhere.

The transfer results provide clear evidence about the structure of rhetorical questions in representation space. Under distributional shift, probe directions retain meaningful separability while showing limited alignment, moderate rank agreement, and low overlap at the rank ends. Together, these observations suggest that, although rhetorical signal is present, it is not organized along a single linear direction. Instead, rhetorical questions appear to be expressed heterogeneously, with different non-collinear directions capturing distinct aspects of rhetorical usage. Similar patterns in the within-dataset analyses support this interpretation.

7 Qualitative Insights into Rhetorical Probing

To move beyond aggregate performance metrics, we conduct a qualitative analysis of the rankings in-

SRAQ rank	RQ rank	Paragraph	Gold Label
1	15	*I find our infatuation with ourselves to be a bit self serving.* ***Who but us cares?*** *It's in our own self interest to think we're great. But* ***what does it accomplish?*** *Nothing. . . . We think we're so great because we can manipulate things. But more often than not this manipulation simply creates major problems . . . "We can think of magnificent things; just look at the incredible architecture we've created."* ***So?*** *Spiders spin incredible webs. Birds weave beautiful nests. Bugs create towers that they live in. Complexity does not necessarily mean superiority.*	rhetorical
2	108	*. . . What anti-copyright fanfic enthusiasts like yourself don't realize is that you're just trying to have your cake and eat it too. . . . You can do that already with public domain material or with your own distilled characters and settings. But you don't want to.* ***Why not?*** *Because you REALLY WANT the popularity of these modern franchises to hold you up and make your story make sense. You're not asking for the ideas, you already have them. You're asking for the popularity and depth of concept; and nobody promised you that.*	rhetorical
3	240	*. . . If you scrutinise the majority of Reykjavik sized "hellhole" cities, you will find that they are often an extension of a greater metropolitan area. For example, Watford is the same size as Reykjavik and Exeter is the same size as Reykjavik. Exeter has a lower crime rate than Watford —* ***why?*** *Because Watford is essentially a city within the greater metropolitan area of London.*	rhetorical
230	1	*My guess is you'll have to go to or phone their office to apply . . . In most states the initial court appearance is considered a "critical stage of the process" where you are entitled to Counsel . . . If this is your first offense most places have a pre-trial diversion program . . . Talk to your lawyer . . . Whatever happens this will still likely be expensive. A "social thing."* ***What? you only drink to make other people more interesting?*** *:)*	rhetorical
146	2	*. . . so* ***why would they even try*** *and be granted legal visitation when they know the issue will be looked at in court? Are they trying to somehow go around this issue? And if so, how? . . . They just closed the suit with his first wife with 3 kids and that one lasted since 1994–2017 . . . It's not for custody, they are asking for visitation rights, which is ridiculous.*	informational
237	3	***Wait, what?*** *Not a single higher spot in Estonia, but a few bumps in Latvia and Lithuania. Do your research, we have huge "mountains" compared to them* =)	rhetorical

Table 1: SRAQ instances ranked by inner product with the SRAQ-derived diffMean direction (top) and RQ-derived (bottom) at layer 48 of Qwen3-32B. Each row reports both ranks. Bold text marks the target question; ellipses indicate truncated text for readability.[4]

duced by diffMean directions at layer 48 of Qwen3-32B. We compute one diffMean direction from SRAQ and one from RQ, then score every SRAQ instance by the inner product of its representation with both directions respectively ($w_{\text{DM}}^{\top} h(x)$; see Section 3.3). For each direction, we rank all SRAQ instances by this score and retain the three highest-ranked examples. Table 1 shows the results: the top rows list the three examples ranked highest by the SRAQ-derived direction, and the bottom rows list those ranked highest by the RQ-derived direction. Each row also reports the same example's rank under the other direction.

The top rows of Table 1 show that the SRAQ-derived direction prioritizes passages in which rhetorical questions serve as structural scaffolding for extended arguments. In the top-ranked example, three successive questions ("Who but us cares?", "what does it accomplish?", "So?") each open a new stage of a philosophical argument about human self-importance; in the second, "Why not?" sets up a multi-sentence explanation of why fanfiction writers seek borrowed popularity rather than building their own. In each case, the rhetorical question drives the discourse forward and organizes the surrounding argument.

In contrast, the bottom rows show that the RQ-derived direction favors short, syntax-driven interrogative forms whose rhetorical force is localized. The top-ranked example is a paragraph of legal advice in which a rhetorical question appears only as a throwaway joke in the final sentence; the third-

[4]Some top-ranked instances are omitted here because they contain socially sensitive content; the complete rankings are reproducible from the released code.

Subset	SRAQ direction	RQ direction
Top 1%	188.8	126.4
Top 3%	150.9	116.4

Table 2: Mean token length of top-ranked SRAQ examples selected by each diffMean direction under the ranking setup used in Section 7. Token counts are computed using the Llama-3.3-70B-Instruct tokenizer.

ranked instance is a two-sentence expression of surprise. Most surprisingly, the second-ranked example is labeled *informational* in the gold annotations: it contains genuine requests for clarification ("why would they even try…?", "Are they trying to somehow go around this issue?") that the RQ-derived direction ranks highly because of their surface interrogative form, not because of any rhetorical intent.

To make this qualitative contrast more concrete, we consider input length as a simple quantitative proxy. While discourse-level properties such as stance-taking or counterargument structure are difficult to measure reliably without additional annotation, length provides a lightweight way to test whether the two directions emphasize different types of examples. As shown in Table 2, under the same ranking setup the top-ranked SRAQ examples selected by the in-domain SRAQ diffMean direction are substantially longer on average than those selected by the transferred RQ diffMean direction. This pattern is consistent with our interpretation that the SRAQ direction more often captures broader discourse context, whereas the RQ direction more often emphasizes more localized rhetorical cues.

Ultimately, this evidence suggests that rhetorical meaning does not manifest as a single dominant direction, but rather as a set of *distinct, emergent rhetorical properties*. These properties span different rhetorical functions, ranging from localized discourse repair to global rhetorical stance, indicating that rhetoric in LLMs is inherently heterogeneous and context-sensitive.

8 Conclusion and Future Work

In this work, we study how rhetorical questions are encoded in large language models using linear probes across two social media datasets. We find that rhetorical content is reliably linearly separable using a single embedding for each input sequence, and that last-token representations provide more stable signals than mean pooling. Although probing directions transfer across datasets, discriminative performance and representational alignment do not always coincide. Probes with similar AUROC can induce substantially different rankings on the same data, with little overlap in the top- and bottom-ranked examples, indicating that similar accuracy can reflect different underlying representations. Qualitative analysis suggests that this divergence reflects the heterogeneous nature of rhetorical questions, which range from discourse-level stance-taking in extended arguments to localized, turn-level interrogative acts.

More broadly, these results caution against treating strong probing performance or successful cross-dataset transfer as evidence of a single shared representational dimension. Linear probes can instead recover different directions that are all effective for discrimination but are not aligned with one another. This suggests that rhetorical questions are encoded in a context-sensitive and heterogeneous manner, rather than along a single linear feature. Future work should clarify how representational features associated with individual directions can be defined and validated, and how such features can be distinguished from collections of directions that perform similarly but capture different aspects of the phenomenon.

Another important direction is to study whether the rhetorical signals identified here are not only encoded, but also controllable. Although we find that rhetorical intent is linearly separable in representation space and our preliminary steering experiments suggest that the identified directions can induce changes in rhetorical behavior, we emphasize that linear separability does not itself imply linear controllability. Extending this analysis with more systematic causal interventions is an important direction for future work.

Limitations

Our empirical analysis is restricted to two datasets drawn from social media domains. While other datasets related to rhetorical or discourse phenomena are available, they generally lack labeling of comparable reliability, consistency, or granularity, which limits their suitability for the type of fine-grained probing analysis conducted in this work. Moreover, the datasets considered here are drawn from real-world data and therefore exhibit inherent noise, which may further obscure or attenuate

underlying signals. For these reasons, we limit our evaluation to these datasets, and our findings should be interpreted within this scope.

In addition, although we analyze representation dynamics across layers, our methodology focuses exclusively on linear probing of representation spaces. This choice emphasizes interpretability and analytical clarity, but it necessarily excludes signals that may be encoded through nonlinear interactions, alternative activation pathways, or mechanisms that are not linearly separable in the representation space considered here.

Acknowledgements

This work is supported in part by a URC Faculty Scholars Research Award from the Office of Research at the University of Cincinnati. We thank the CincyNLP group for helpful discussions, and the anonymous reviewers for their valuable feedback.

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, and others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes.

Shohini Bhattasali, Jeremy Cytryn, Elana Feldman, and Joonsuk Park. 2015. Automatic identification of rhetorical questions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2015)*.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*.

Lucy Farnik, Tim Lawson, Conor Houghton, and Laurence Aitchison. 2025. Jacobian sparse autoencoders: Sparsify computations, not just activations. In *Proceedings of the 42nd International Conference on Machine Learning (ICML 2025)*.

Jane Frank. 1990. You call that a rhetorical question?: Forms and functions of rhetorical questions in conversation. *Journal of Pragmatics*, 14:723–738.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2025. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Chung-hye Han. 2002. Interpreting interrogatives as rhetorical questions. *Lingua*, 112:201–229.

Thomas Heap, Tim Lawson, Lucy Farnik, and Laurence Aitchison. 2025. Sparse autoencoders can interpret randomly initialized transformers. *arXiv preprint arXiv:2501.17727*.

Eghbal A. Hosseini and Evelina Fedorenko. 2023. Large language models implicitly learn to straighten neural sentence trajectories to construct a predictive representation of natural language. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS 2023)*.

Oghenevovwe Ikumariegbe, Eduardo Blanco, and Ellen Riloff. 2025. Studying rhetorically ambiguous questions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*.

Shiyu Ji, Farnoosh Hashemi, Joice Chen, Juanwen Pan, Weicheng Ma, Hefan Zhang, Sophia Pan, Ming Cheng, Shubham Mohole, Saeed Hassanpour, Soroush Vosoughi, and Michael Macy. 2025. A generalizable rhetorical strategy annotation model using LLM-based debate simulation and labelling. In *Findings of the Association for Computational Linguistics: EMNLP 2025 (Findings of EMNLP 2025)*.

Daniel Jurafsky, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries, Elizabeth Shriberg, Andreas Stolcke, Paul Taylor, and Carol Van Ess-Dykema. 1997. Automatic detection of discourse structure for speech recognition and understanding. In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 88–95. IEEE.

Zlata Kikteva, Alexander Trautsch, Steffen Herbold, and Annette Hautli-Janisz. 2024. Question type prediction in natural debate. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2024)*.

Joshua Lee, Wyatt Fong, Alexander Le, Sur Shah, Kevin Han, and Kevin Zhu. 2025. Pragmatic metacognitive prompting improves LLM performance on sarcasm detection. In *Proceedings of the 1st Workshop on Computational Humor (CHum)*.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling (COLM 2024)*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS 2022)*.

Shereen Oraby, Vrindavan Harrison, Amita Misra, Ellen Riloff, and Marilyn Walker. 2017. Are you serious?: Rhetorical questions and sarcasm in social media dialog. In *Proceedings of the 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2017)*.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2023. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*.

Jingyi Qiu, Hong Chen, and Zongyi Li. 2025. Counterfactual llm-based framework for measuring rhetorical style. *arXiv preprint arXiv:2512.19908*.

Alex Reinhart, Ben Markey, Michael Laudenbach, Kachatad Pantusen, Ronald Yurko, Gordon Weinberg, and David West Brown. 2025. Do llms write like humans? variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences*, 122(8):e2422455122.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*.

Richard M Roberts and Roger J Kreuz. 1994. Why do people use figurative language? *Psychological science*, 5:159–163.

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 *Proceedings of the 42nd International Conference on Machine Learning (ICML 2025)*.

Džemal Špago. 2016. Rhetorical questions or rhetorical uses of questions? *Explorations in English Language and Linguistics*, 4:102–115.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Daniel Vennemeyer, Phan Anh Duong, Tiffany Zhan, and Tianyu Jiang. 2025. Sycophancy is not one thing: Causal separation of sycophantic behaviors in llms. *arXiv preprint arXiv:2509.21305*.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, and 1 others. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*.

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. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yazhou Zhang, Chunwang Zou, Zheng Lian, Prayag Tiwari, and Jing Qin. 2025. Sarcasmbench: Towards evaluating large language models on sarcasm understanding. *IEEE Transactions on Affective Computing*, 16(4):2560–2578.

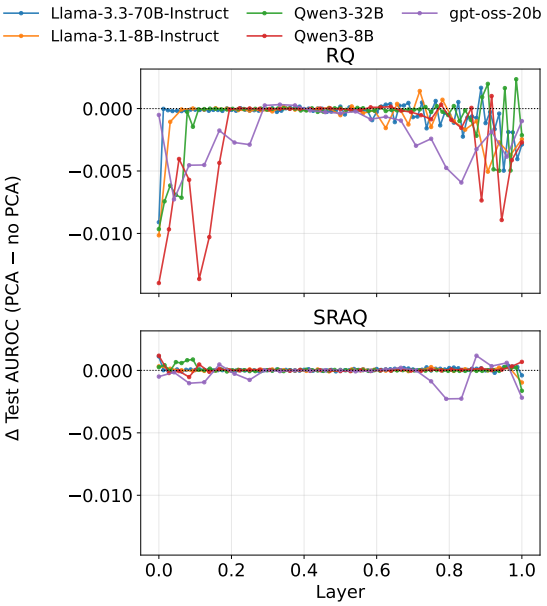
Yuan Zhuang and Ellen Riloff. 2020. Exploring the role of context to distinguish rhetorical and information-seeking questions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*.

A Effect of PCA Truncation on Linear Probing

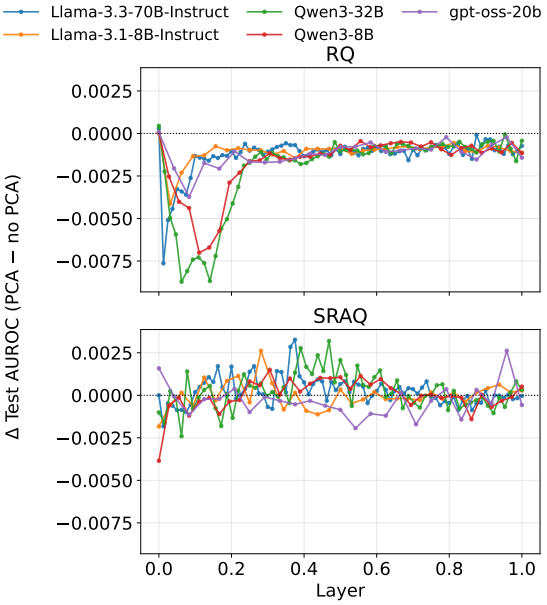
In this section, we compare diffMean probe performance using the first 64 PCA components versus the full embedding spaces. As shown in Figure 5, the resulting AUROC differences are consistently small across models and layers, remaining below one percentage point in magnitude. This indicates that, for diffMean probes, projecting representations onto a low-dimensional PCA subspace does not meaningfully alter discriminative performance relative to using the full embedding space.

To further contextualize this result, Figure 6 reports the per-layer explained-variance ratio of the 64th principal component for both mean-token and last-token representations. Across models and layers, the variance explained by the 64th component remains well below 1%, indicating that even the highest-index component retained in our PCA truncation captures only a very small fraction of the total variance. Consequently, all subsequent components beyond the 64th are expected to contribute even less. This provides a geometric explanation for the negligible AUROC differences observed above: restricting representations to the leading 64 principal components preserves nearly all variance relevant to diffMean linear probes, while discarding directions that collectively account for only a minor fraction of the embedding space.

This observation motivates our choice to restrict representations to the first 64 principal components in subsequent experiments involving other linear probes, including logistic and hinge classifiers. Since higher-index components explain only a negligible fraction of the total variance, retaining them is unlikely to add meaningful discriminative signal and may instead introduce noise or numerical instability. In practice, projecting onto the leading 64 components yields more stable training and evaluation behavior for these probes without materially affecting performance. We therefore focus on this dimensionality throughout the remainder of our linear probing analyses.



(a) Mean token.



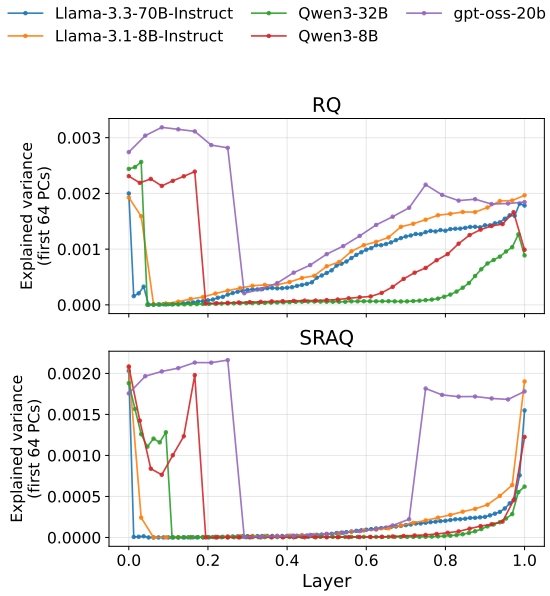
(b) Last token.

Figure 5: AUROC differences (PCA minus no PCA) for diffMean linear probes under mean-token and last-token pooling.

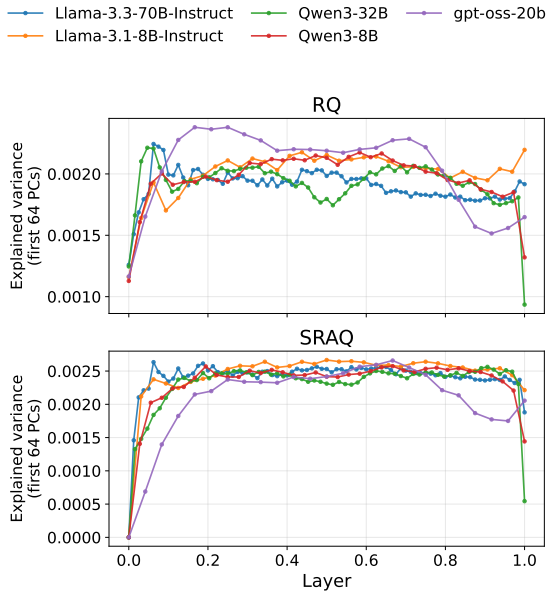
B Mapping Linear Probes from PCA to Embedding Space

In our main experiments, logistic-regression and hinge-loss probes are trained in a PCA-reduced space of dimension k (here k=64) for numerical stability and consistent comparisons across probes. Because PCA projections differ across datasets, expressing a learned probe in the original embedding space of dimension d allows for consistent interpretation and comparison.

PCA projection. Let $x \in \mathbb{R}^d$ denote an original (sequence-level) representation, and let $\mu \in \mathbb{R}^d$ be the PCA mean computed on the training split for a fixed dataset–model–input setting. Let $W \in \mathbb{R}^{k \times d}$ be the PCA loading matrix whose rows are the top-k principal directions (orthonormal). The PCA



(a) Mean-token representations.



(b) Last-token representations.

Figure 6: Per-layer explained-variance ratio of the 64th principal component (i.e., the marginal variance explained by the 64th principal component) for mean-token vs. last-token representations.

coordinates are

$$z \;=\; (x-\mu)W^\top \in \mathbb{R}^k. \tag{5}$$

Linear probe in PCA space. Both logistic regression and hinge-loss (linear SVM) define an affine scoring function in PCA space,

$$s(z) \;=\; w_z^\top z + b, \tag{6}$$

where $w_z \in \mathbb{R}^k$ is the learned weight vector and $b \in \mathbb{R}$ is the bias/intercept.

Mapping weights back to the original space. Substituting $z = (x-\mu)W^\top$ yields

$$s(x) = w_z^\top (x-\mu)W^\top + b \tag{7}$$

$$= (W^\top w_z)^\top x \;+\; \big(b - (W^\top w_z)^\top \mu\big). \tag{8}$$

Therefore, the equivalent probe in the original embedding space has weight vector

$$w_x \;=\; W^\top w_z \in \mathbb{R}^d \tag{9}$$

and bias

$$b_x \;=\; b - w_x^\top \mu. \tag{10}$$

This mapping preserves scores exactly for any x under the same PCA transform. Intuitively, the mapped-back classifier places all its mass in the PCA subspace; components orthogonal to $\mathrm{span}(W)$ have zero weight.

C Geometric Characterization of Dataset Differences

Throughout the paper, we observe that RQ and SRAQ exhibit systematically different rhetorical behaviors, both in qualitative examples and in the linear directions identified by probing. Motivated by these behavioral differences, we further examine whether they are reflected in the geometry of the representation spaces already analyzed in the main text.

Specifically, we reuse the same PCA-based subspace construction employed throughout the paper. For each model and layer, we form a 64-dimensional subspace using the top-64 principal components of the embedding distributions induced by each dataset. This allows us to directly compare the geometric structure of the dominant variance directions underlying the observed rhetorical signals, without introducing a new representation.

We compare the resulting subspaces across datasets using two complementary alignment measures, which capture different geometric aspects. The first is the geodesic distance between subspaces on the Grassmann manifold, computed from their principal angles. This metric depends only on the subspaces themselves and is invariant to the ordering or orientation of individual basis vectors. As a result, it provides a global measure of subspace alignment, quantifying how similarly the two datasets organize their dominant variance directions as a whole.

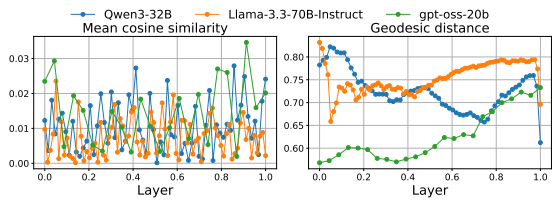


Figure 7: Subspace alignment between RQ and SRAQ across layers for multiple models. We report cosine similarity and geodesic distance between the corresponding layer-wise subspaces, using normalized layer index on the x-axis. Higher distances indicate weaker cross-dataset alignment.

The second measure is the mean cosine similarity between corresponding principal components. Unlike geodesic distance, this metric is sensitive to the alignment of individual PCA directions and implicitly depends on their ordering. It therefore captures whether not only the subspaces overlap, but also whether their leading directions of variation are aligned in a consistent manner. In practice, the mean cosine similarity remains close to zero across layers, indicating that even when subspaces are not maximally separated, their dominant directions are largely misaligned.

Figure 7 reports both metrics across normalized layer depth for multiple models. Together, these measurements provide a geometric characterization of dataset differences from two perspectives: overall subspace alignment, and fine-grained alignment of leading variance directions, complementing the behavioral differences observed throughout the paper.

In addition, we observe that `GPT-OSS-20B` (Agarwal et al., 2025) exhibits qualitatively different behavior from the other models considered. Across layers, this model shows distinct trends in geodesic distance, a pattern that is consistent with its behavior observed under multiple other analyses in the paper. While we do not attribute this difference to a specific architectural or training factor, the consistency of this effect across independent metrics suggests that it reflects a systematic property of the model's representations rather than measurement noise.

D DiffMean Probing with Alternative Pooling

To further examine the effect of representation choice, we run additional training-free diffMean experiments with alternative pooling strategies,

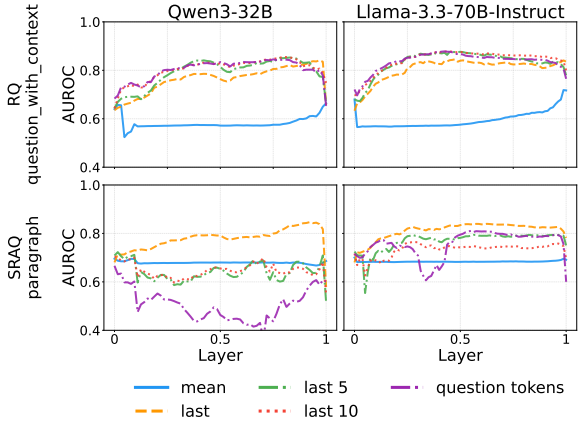


Figure 8: Test AUROC of training-free diffMean directions across layers under alternative pooling strategies. Results are shown for RQ (*question_with_context*) and SRAQ (*paragraph*) using Qwen3-32B and Llama-3.3-70B-Instruct. We compare mean pooling over all tokens, last-token pooling, pooling over the last 5 or 10 tokens, and mean pooling over question tokens only.

shown in Figure 8. In addition to mean pooling over all tokens and standard last-token pooling, we consider pooling over the last 5 or 10 tokens and mean pooling restricted to the question span only (*question tokens*). We evaluate these variants across layers for Qwen3-32B and Llama-3.3-70B-Instruct on RQ (*question_with_context*) and SRAQ (*paragraph*). On RQ, pooling over the last few tokens and question-token pooling are often competitive with, and sometimes slightly better than, last-token pooling at middle layers. On SRAQ, these alternatives are generally less stable and often underperform last-token pooling, especially for Qwen3-32B. Overall, these results support the main-text focus on mean pooling and last-token representations as the two standard sequence-level settings for analyzing how rhetorical intent is encoded.

E Encoder-Based Model Results

As a comparison to the decoder-only models in the main text, we also evaluate an encoder-based model, ModernBERT-large (Warner et al., 2025), using the [CLS] representation across layers. The results are shown in Figure 9. On RQ, AUROC improves gradually in the earlier and middle layers, reaching a peak of 64.8 at layer 17, and then declines toward later layers. On SRAQ, performance is more stable across the network, remaining in a relatively narrow range between 67.8 and 69.9. Overall, these results remain below those of the decoder-only models reported in the main text.

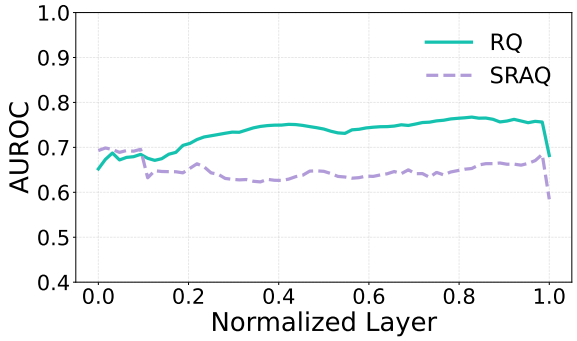


Figure 9: Test AUROC of training-free diffMean directions across normalized layers for ModernBERT-large using the [CLS] representation, shown for RQ and SRAQ.

One possible interpretation is that rhetorical information is organized differently in encoder-based models, so that a single [CLS] representation may not provide the most effective readout for this task. More generally, the contrast suggests that representation choice may play a more important role for encoder-based models, where rhetorical cues could be distributed across tokens rather than concentrated in a single sequence-level summary. A broader encoder-decoder comparison, including alternative token aggregation strategies, is left for future work.

F Causal Steering

To validate that the identified directions capture rhetorical intent, we conduct causal steering experiments and report the results in this section, following the activation steering methodology of Turner et al. (2023) and Rimsky et al. (2024).

Steering Mechanism. During inference, we apply steering by modifying the residual stream at a chosen layer. The steered hidden state is computed as

$$h'_l = h_l + \alpha \cdot v_l, \tag{11}$$

where h_l denotes the original residual stream at layer l, h'_l the steered residual stream, α controls the strength of the intervention, and v_l is the steering vector for that layer. In our experiments, we use the diffMean vectors identified in the main text.

Setup. We conduct our steering experiments on the RQ dataset using Qwen3-32B, where the context and the question are explicitly separated. Experiments are performed across seven layers, using 400 contexts drawn from the dataset (200 originally labeled as rhetorical and 200 as informational). To test whether the identified direction captures rhetorical question intent, we provide the model with only the context and prompt it to generate a follow-up question. The prompt template is shown in Figure 10, and the generation hyperparameters are listed in Table 3.

Parameter	Value
max_new_tokens	50
do_sample	True
temperature	0.7
top_p	0.9
repetition_penalty	1.1

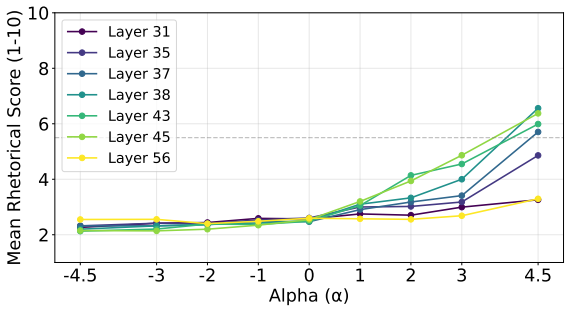
Table 3: Generation parameters used in the steering experiments.

```
{context}
Ask one concise follow-up question (ideally
under 15 words). Your entire reply should be
just that single question.
```

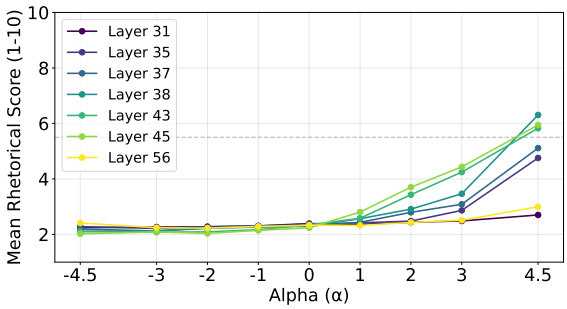
Figure 10: Prompt template used for question generation in the steering experiments.

We then feed each generated question, together with its context, into GPT-5.1 to obtain a rhetorical score on a 1–10 scale, where 1 corresponds to a purely informational question and 10 corresponds to a strongly rhetorical question. The scoring prompt used for this evaluation is shown in Figure 12.

Results. The steering results are shown in Figure 11. Rhetorical scores increase largely monotonically with the steering strength α, and for sufficiently large positive values (e.g., $\alpha = 4.5$), several layers reach scores around 7, indicating strongly rhetorical behavior. The magnitude of this effect varies across layers, with layers 38, 43, and 45 responding most strongly to steering, while others (e.g., layers 31 and 56) show weaker responses. Contexts originally paired with rhetorical questions consistently reach slightly higher scores, reinforcing the role of discourse context in shaping rhetorical interpretation. Taken together, these results provide evidence that the identified linear directions capture rhetorical question intent in the model's internal representations.



(a) Contexts paired with rhetorical questions in the original dataset



(b) Contexts paired with informational questions in the original dataset

Figure 11: Alpha sweep results on Qwen3-32B. Mean rhetorical score (y-axis) as a function of the steering coefficient α (x-axis) for different steering layers, evaluated separately on rhetorical and informational contexts.

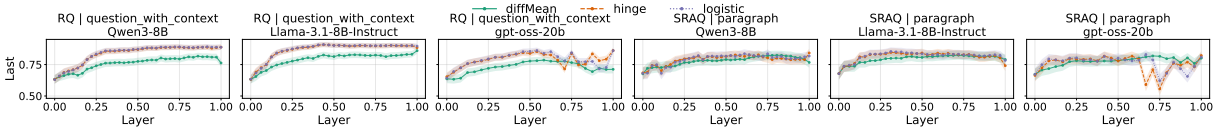
G Additional Results on AUROC and Alignment

Figures 13–15 report additional analyses that extend the main text to settings not previously shown. Specifically, we evaluate last-token probe performance on the `full_turn` input formulation for SRAQ, and examine both within-dataset and cross-dataset alignment for three additional LLMs: QWEN3-8B, LLAMA-3.1-8B-INSTRUCT, and GPT-OSS-20B.

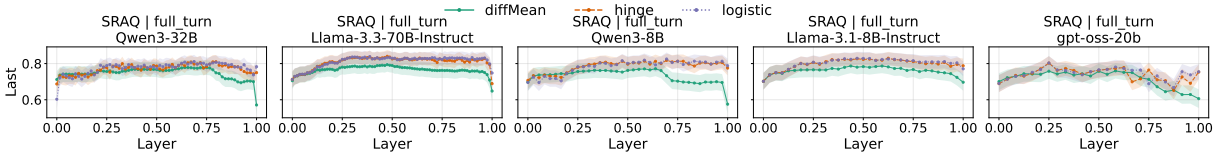
Across these settings, rhetorical questions remain linearly separable, with AUROC trends broadly consistent with those reported in the main text. However, results for GPT-OSS-20B exhibit noticeably higher variability across layers and runs compared to the other models, both in AUROC and in alignment-based measures. We do not observe comparable instability for QWEN3-8B or LLAMA-3.1-8B-INSTRUCT. While we do not investigate the source of this behavior further, this observation suggests that probe stability may vary substantially across model architectures, even when overall discriminative performance appears similar.

```
Your task is to rate a generated question on a scale of 1-10 based on how rhetorical versus
informational it is.

Scoring Scale:
10 = Most rhetorical: strong presuppositions, obvious implied answers, highly persuasive tone.
7-9 = Rhetorical: contains presuppositions, implied answers, persuasive tone, or tag questions.
4-6 = Neutral / mixed: contains both factual and rhetorical elements.
2-3 = Informational: direct, factual question seeking specific information.
1 = Most informational: purely factual question with no rhetorical elements.

Context: {context}
Generated Question: {question}

Provide only a single integer from 1-10 as your rating.
```

Figure 12: Prompt used to score generated questions on a rhetorical–informational scale.



(a) Additional LLMs. Last-token test AUROC across layers for models not shown in the main text.



(b) Additional input column. Last-token test AUROC across layers for the `full_turn` formulation (not shown in the main text).

Figure 13: Additional last-token probe results. Both panels report within-setting performance (training and evaluation within the same dataset, model, and input formulation) as a function of layer. (a) Results for additional LLMs beyond those highlighted in the main figures. (b) Results for the `full_turn` input formulation. Across settings, trends are consistent with the main text: rhetorical question intent remains linearly separable across layers, and last-token representations provide stable signals.

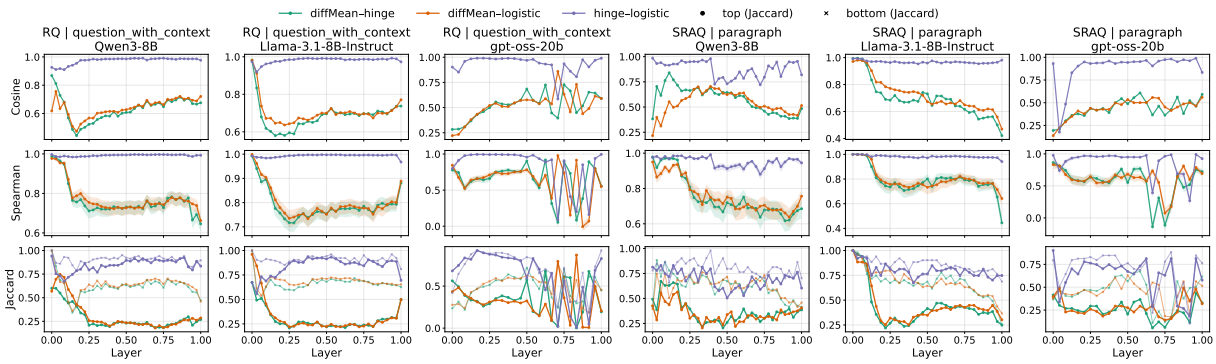


Figure 14: Within-dataset alignment across additional LLMs. Comparison of probe alignment metrics computed within the same dataset for models not shown in the main text. Although probes achieve similar discriminative performance, their learned directions exhibit limited alignment, indicating that multiple, non-collinear directions can support linear separability even within a fixed dataset. Trends are consistent with the main results.

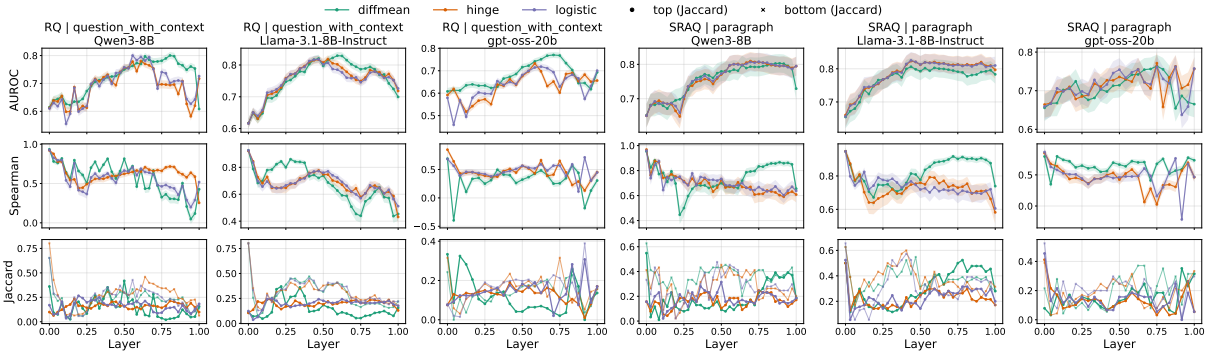


Figure 15: Cross-dataset alignment across additional LLMs. Comparison of probe alignment metrics when probe directions learned on one dataset are applied to another, for models not shown in the main text. Despite partial transferability in discriminative performance, alignment between probe directions remains limited, reflecting divergence in the induced rankings across datasets. These results are consistent with the trends reported in the main text.