THE GEOMETRY OF KNOWLEDGE ACCESSIBILITY IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) contain broad knowledge, but they cannot access all of it reliably. We study this problem through *knowledge accessibility*, which describes whether the knowledge needed for a query can be recalled from the model. We find that knowledge accessibility has a simple geometric structure in the model's representation of the query alone, before any generation. More accessible queries are closer to a center in the representation space, while less accessible queries are farther away. This geometry reveals a *knowledge boundary* that separates more accessible queries from less accessible ones. Accessibility consistently decreases with distance from the center, and this distance-based ordering transfers across datasets even when the centers differ. Controlled experiments further show that the centered geometry is more closely related to knowledge accessibility than to reasoning difficulty. The geometry also reveals when different interventions are useful. Query rewriting helps more for accessible queries, chain-of-thought reasoning helps more near the boundary, and retrieval gives larger gains beyond the boundary. These findings not only provide a new geometric view of how knowledge is organized in language models, but also suggest a useful pre-generation signal for adaptive inference.

1 INTRODUCTION

Large language models (LLMs) store a large amount of knowledge during pre-training (Petroni et al., 2019), yet this knowledge is not equally accessible for every query. This limitation is often described by the *knowledge boundary*, which separates queries that a model can answer from those that it cannot (Kadavath et al., 2022; Li et al., 2025). Recent work shows that this boundary can be estimated from the internal states of a query before any token is generated (Ni et al., 2025; Chen et al., 2026). Such generation-free signals are potentially useful for deciding how much additional computation or external information a query may require.

However, existing work often defines the knowledge boundary in terms of observed answer correctness. This conflates knowledge accessibility with reasoning, as a model may access the required knowledge but still answer incorrectly due to reasoning failures (Press et al., 2023; Yang et al., 2026). Moreover, prior methods mainly ask whether knowledge is encoded in hidden states (Dai et al., 2022; Chen et al., 2025), but they do not reveal how knowledge accessibility itself is organized in representation space.

In this work, we study the problem through *knowledge accessibility*, which describes whether the knowledge required by a query is accessible from the model's parameters. To isolate knowledge recall from reasoning, we focus on single-fact queries, such as "*The capital of France is?*", whose answers require one factual association rather than multi-step reasoning (Petroni et al., 2019; Press et al., 2023). This provides a controlled setting for studying the representation of accessibility. We then learn the geometry of knowledge accessibility in query representations and ask a different question:

What is the geometry of knowledge accessibility in an LLM's representation space?

We propose the *accessibility center hypothesis*. Inspired by center-based structures in representation learning (Wen et al., 2016; Ruff et al., 2018), we hypothesize that reliably accessible knowledge tends to occupy a localized region of representation space. We model this region with an accessibility

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

center and measure each query by its distance from the center, with smaller distances indicating higher knowledge accessibility. This distance naturally defines a *knowledge boundary* separating regions of higher and lower accessibility.

Across two LLMs and four QA datasets, we find consistent support for this centered organization. Simple centered geometries predict knowledge accessibility better than competitive probing baselines on knowledge-focused tasks, and accessibility consistently decreases with distance from the learned center. In contrast, on reasoning-heavy tasks, a flexible MLP gains an advantage by capturing additional reasoning signals, while the centered geometry remains more closely associated with factual accessibility. Controlled experiments further support this distinction between accessibility and reasoning.

The geometric view also reveals several properties that are difficult to see from correctness prediction alone. Queries tend to be less stable in their correctness as they approach the boundary, while their distance-based accessibility ordering transfers across datasets even when the precise centers differ. More importantly, the geometry suggests how different interventions benefit the model. Query rewriting is most effective for accessible queries, chain-of-thought reasoning becomes more useful near the boundary, and retrieval provides larger gains beyond the boundary. Thus, the knowledge boundary is not only a signal of whether a model may need help, its geometry also provides information about what kind of help is likely to be useful. These results provide a geometric perspective on how knowledge is organized and accessed in LLMs, and suggest that knowledge accessibility may serve as a useful signal for adaptive and more efficient inference (Yong et al., 2025a; Chen et al., 2026).

2 Knowledge Accessibility

We conceptually break down query correctness as $C \approx f(K, R, O)$, where K, R, and O denote *knowledge accessibility*, *reasoning capability*, and *other task-specific capabilities*. That is, answering a query correctly requires both access to the needed knowledge and the ability to use it appropriately. This decomposition is motivated by prior work showing that access to relevant knowledge and the ability to reason with that knowledge are distinct factors (Press et al., 2023; Yang et al., 2026). We treat knowledge accessibility (K) and reasoning capability (R) as the primary factors of interest for query correctness, while O captures other task-specific factors whose influence is expected to be limited in our setting.

In this work, we focus on knowledge accessibility, which we define as the ability to recall the factual knowledge required by a query from the model's parameters. To study knowledge accessibility separately from reasoning capability, we focus on *single-fact factual queries*. Answers of these queries mainly depend on recalling a single target fact (Petroni et al., 2019; Roberts et al., 2020). This setting allows us to isolate factual recall from multi-step reasoning.

2.1 Query Correctness and Factual Recall

For a single-fact factual query x, let $\mathbb{Y}(x)$ denote the set of valid answers and let $\hat{y}_\pi(x)$ be the model response under decoding strategy π. We define

$$c_\pi(x) \triangleq \mathbf{1}[\hat{y}_\pi(x) \in \mathbb{Y}(x)]. \tag{1}$$

In our single-fact setting, correctness mainly reflects whether the target fact is successfully recalled. We therefore use $c_\pi(x)$ as a behavioral observation of factual recall under decoding strategy π, with correctness potentially varying across decoding strategies.

More generally, we view knowledge accessibility as the model's ability to recall the required factual knowledge from its parameters under a given elicitation and decoding setting. Low accessibility may reflect either knowledge that is not reliably stored or knowledge that is stored but not successfully accessed. This work does not distinguish between these two cases.

For more complex tasks, such as mathematical or multi-step reasoning, correctness depends on more than knowledge accessibility. A model may have access to the required knowledge but still fail because of reasoning or other task-specific capabilities. This motivates our use of single-fact factual queries, where correctness provides a cleaner proxy for knowledge accessibility.

2.2 A Geometric Hypothesis of Knowledge Accessibility

Let $\boldsymbol{h}_\ell(x) \in \mathbb{R}^d$ denote the hidden representation at the final query token at layer ℓ. We compute $\boldsymbol{h}_\ell(x)$ from the query alone, before generating any answer tokens. For notational simplicity, we omit the layer index ℓ in the following. We therefore ask whether knowledge accessibility is already reflected in representation space before generation, and, if so, what geometric structure it takes.

Prior work has mostly used linear probing to study this question. These studies ask whether hidden representations can separate queries with different model outcomes (Xiao et al., 2025; Chen et al., 2026). This shows that hidden states may contain information about knowledge accessibility. However, it does not tell us how this information is organized in representation space.

A simple center-based structure has been observed in learned representations. In particular, representations with shared properties may become more concentrated around a center in representation space (Papyan et al., 2020; Wu & Papyan, 2024). We ask whether a similar effect also appears for knowledge accessibility. Similar queries may develop a shared geometric structure during training. Queries based on frequently seen knowledge may form a common center, while queries based on rarely seen or unseen knowledge receive less shared training signal and may lie farther from this center.

Accessibility Center Hypothesis. We propose that queries may be organized around an *accessibility center*, with distance from this center related to how easily the model can recall the required knowledge. Specifically, we posit an accessibility center $\boldsymbol{c} \in \mathbb{R}^d$ and define the knowledge distance as $D(x) \triangleq d(\boldsymbol{h}(x), \boldsymbol{c})$. We hypothesize that factual recall becomes less likely as knowledge distance increases. Let X denote a query drawn from the data distribution. Formally,

$$r_1 < r_2 \quad \Longrightarrow \quad \mathbb{E}[c(X) \mid D(X) = r_1] \geq \mathbb{E}[c(X) \mid D(X) = r_2] \tag{2}$$

The accessibility center should not be interpreted as a location where individual facts are stored. Rather, it summarizes a region of representation space associated with more reliable access to factual knowledge.

3 Modeling the Geometry of Knowledge Accessibility

3.1 Centered Accessibility Geometry

Center-based structures have been widely used in representation learning (Snell et al., 2017; Ruff et al., 2018). For example, center loss encourages representations sharing the same label to concentrate around a learned feature center (Wen et al., 2016). Motivated by this view, we instantiate the accessibility-center hypothesis with a simple hyperspherical geometry. Specifically, we define $D(x) \triangleq \|\boldsymbol{h}(x) - \boldsymbol{c}\|_2$, so that points with the same knowledge distance lie on a hypersphere centered at $\boldsymbol{c}$.

This formulation makes an isotropic assumption. Distance from the center is measured equally in all representation dimensions. Representations, however, may vary along different dimensions at different scales (Ethayarajh, 2019). We therefore relax this assumption and consider an anisotropic extension,

$$D(x) \triangleq \big(\boldsymbol{h}(x) - \boldsymbol{c}\big)^\top \boldsymbol{A} \big(\boldsymbol{h}(x) - \boldsymbol{c}\big), \tag{3}$$

where $\boldsymbol{A} \succeq 0$ is *dimension weights*, which is diagonal and controls the relative scaling of different representation dimensions. The resulting level sets are hyperellipsoids rather than hyperspheres, while retaining the assumption that accessibility is organized around a common center.

To test whether accessibility is organized around a center, rather than merely encoded in the representation, we compare against two references. Linear probing is widely used to assess whether a concept is encoded in representations. We therefore use the score $\boldsymbol{w}^\top \boldsymbol{h}(x) + b$ as a linear baseline. We additionally consider an MLP, with more parameters and nonlinear activations, which provides a more expressive reference. These models contrast a centered inductive bias with linear separability and flexible nonlinear predictability, respectively, as summarized in Table 1.

Hypothesis	Accessibility score	Geometry	Parameters
Linear	$\boldsymbol{w}^\top \boldsymbol{h} + b$	Hyperplane	$\mathcal{O}(d)$
Centered	$-\|\boldsymbol{h} - \boldsymbol{c}\|_2^2$	Hypersphere	$\mathcal{O}(d)$
Anisotropic centered	$-(\boldsymbol{h} - \boldsymbol{c})^\top \boldsymbol{A}(\boldsymbol{h} - \boldsymbol{c})$	Hyperellipsoid	$\mathcal{O}(d)$
Flexible nonlinear	$\mathrm{MLP}(\boldsymbol{h})$	MLP reference	$\mathcal{O}(dm)$

Table 1: Geometric hypotheses of knowledge accessibility in representation space. Here, d is the query representation dimension and m is the MLP hidden dimension.

Dataset	K	R	Task Type
TriviaQA	High	Low	Factual QA
SciQ	High	Medium	Science QA
MuSiQue	High	High	Multi-hop QA
GSM8K	Medium	High	Math reasoning

Table 2: Decomposing answerability across benchmarks. Correctness reflects knowledge accessibility (K) and reasoning capability (R).

3.2 The Knowledge Boundary

Under the accessibility-center hypothesis, queries farther from the accessibility center are less likely to support successful factual recall. This radial ordering naturally induces a transition between regions of higher and lower knowledge accessibility. We refer to this transition as the *knowledge boundary*.

Let τ denote a learned boundary radius and let $\epsilon > 0$ control the width of the transition region. Under this geometry, we define three regions:

$$D(x) < \tau - \epsilon \quad (\textit{interior}),\ |D(x) - \tau| \quad \leq \epsilon \quad (\textit{boundary}),\ D(x) > \tau + \epsilon \quad (\textit{exterior}). \tag{4}$$

For subsequent analysis, we define the signed distance to the boundary as $\delta(x) \triangleq D(x) - \tau$. Accordingly, the three regimes can equivalently be expressed as

$$\delta(x) < -\epsilon \quad (\textit{interior}),\ |\delta(x)| \quad \leq \epsilon \quad (\textit{boundary}),\ \delta(x) > \epsilon \quad (\textit{exterior}). \tag{5}$$

Queries in the interior are expected to have higher knowledge accessibility, while those in the exterior have lower accessibility. The boundary marks an unstable transition region between the two.

4 Evaluating the Accessibility Geometry

4.1 Experimental Setup

We evaluate two instruction-tuned language models, `Llama-3.1-8B-Instruct` (Grattafiori et al., 2024) and `Qwen-2.5-14B-Instruct` (Hui et al., 2024), across four QA benchmarks with different demands on knowledge accessibility and reasoning (Table 2).

TriviaQA (Joshi et al., 2017) and SciQ (Welbl et al., 2017) provide relatively clean settings for studying knowledge accessibility, as correctness mainly reflects factual recall. MuSiQue (Trivedi et al., 2022) and GSM8K (Cobbe et al., 2021) require more reasoning. They let us test whether the centered geometry still works when correctness depends less directly on knowledge access. In such settings, a more flexible nonlinear model may become advantageous. For each query, we obtain the observed answer using the greedy decoding policy π (Equation 1), and use `Mistral-Large` to judge its correctness.

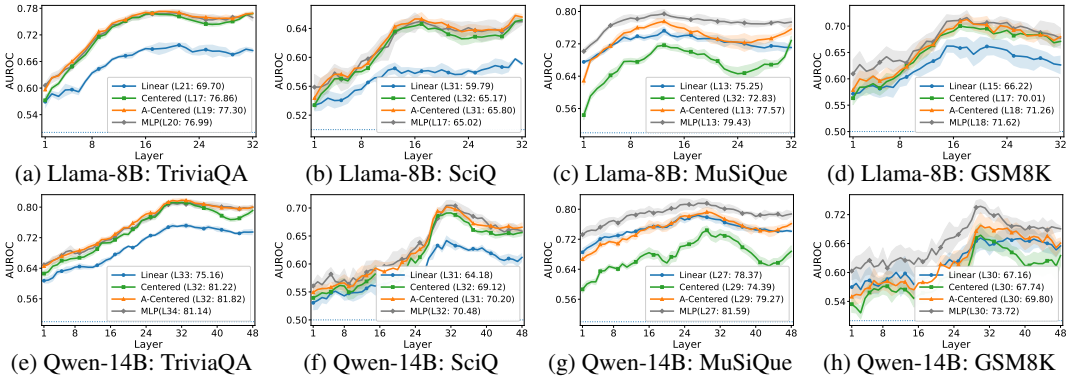
For each query, we extract the hidden state at the final query token from the query alone, without generating any answer tokens. We use a 40%/10%/50% train/validation/test split and use the same split for all layers within each random seed. All results are averaged over five random seeds.

For the layer-wise analysis, we fit each geometry independently at every layer. For all subsequent analyses, we use the layer with the highest mean validation AUROC and keep this layer fixed. We set τ on the validation set as the distance where the correctness probability is 0.5. We choose the boundary width ϵ such that the boundary region corresponds to predicted correctness probabilities between 0.45 and 0.55. Full training details and hyperparameters are provided in Appendix A.

4.2 Comparing Knowledge Accessibility Geometries

Table 1 compares four hypotheses for how answerability is organized in representation space: a linear geometry, an isotropic centered geometry, an anisotropic centered geometry (*A-Centered*), and a flexible nonlinear model (MLP). Figure 1 evaluates these hypotheses layer by layer across the four benchmarks and two language models.

On TriviaQA and SciQ, A-Centered is the strongest method on three of the four model–dataset pairs. Notably, even the simple Centered geometry achieves performance close to the flexible MLP baseline.



(a) Llama-8B: TriviaQA (b) Llama-8B: SciQ (c) Llama-8B: MuSiQue (d) Llama-8B: GSM8K

(e) Qwen-14B: TriviaQA (f) Qwen-14B: SciQ (g) Qwen-14B: MuSiQue (h) Qwen-14B: GSM8K

Figure 1: Layer-wise comparison of different geometry hypotheses across four datasets and two language models. Peak performance is reported in the legend.

In contrast, the Linear model performs substantially worse than the centered models. Thus, when correctness is more directly driven by knowledge accessibility, a structured centered geometry is sufficient to match or outperform a flexible nonlinear predictor. These results suggest that much of the accessibility signal can be captured by a simple centered geometry, rather than a complex nonlinear model.

On MuSiQue and GSM8K, where correctness depends more strongly on reasoning capability, the flexible MLP performs best overall. This result is expected because correctness on these tasks depends on both knowledge accessibility and reasoning capability. Our hypothesis focuses on the geometric structure of knowledge accessibility rather than correctness in general. Even so, A-Centered remains competitive while using only $\mathcal{O}(d)$ parameters, compared with $\mathcal{O}(dm)$ for the MLP, as shown in Table 1. This suggests that the centered geometry still captures a large part of the correctness signal with a much simpler model.

One may ask whether Centered performs better than Linear simply because it uses the squared norm of the representation, rather than because of the centered structure itself. The Centered score can be expanded as $-\|\boldsymbol{h}-\boldsymbol{c}\|_2^2 = 2\boldsymbol{c}^\top\boldsymbol{h} - \|\boldsymbol{h}\|_2^2 - \|\boldsymbol{c}\|_2^2$, which contains both a linear term and a squared norm term. We therefore introduce *Linear+Norm*, $s(\boldsymbol{h}) = \boldsymbol{w}^\top\boldsymbol{h} + \alpha\|\boldsymbol{h}\|_2^2 + b$, which learns these terms freely without enforcing a centered structure. For AUROC, Centered depends on d center parameters, while Linear+Norm depends on d linear weights and one norm coefficient α, giving it at least as much effective capacity.

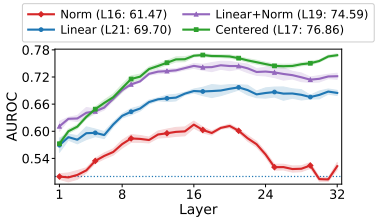


Figure 2: AUROC on TriviaQA with Llama-8B.

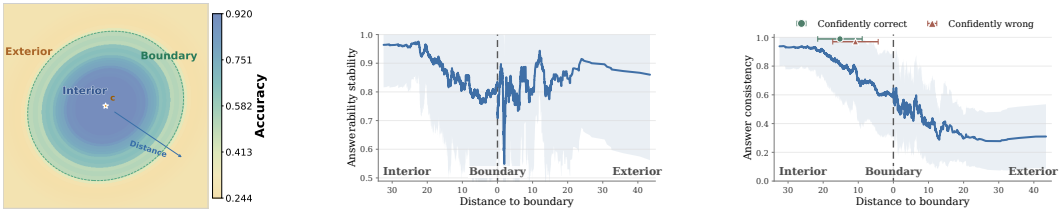
As shown in Figure 2, Linear+Norm improves over Linear but remains below Centered. This suggests that the benefit of Centered is not explained by the additional norm term or greater model capacity, but is consistent with a useful centered inductive bias. We further analyze the learned coefficient α in Appendix B.1, which shows that the Linear+Norm model also tends to learn a centered structure.

4.3 Separating Knowledge Accessibility from Reasoning

To further test more directly whether the centered geometry captures knowledge accessibility rather than reasoning capability, we construct a controlled evaluation on MQuAKE (Zhong et al., 2023). MQuAKE provides multi-hop questions together with their single-hop subquestions, which allows us to separate knowledge accessibility from reasoning capability. The single-hop questions test whether the individual facts needed for the question can be recalled, while the multi-hop question tests whether the model can use reasoning to combine these facts and produce the final answer. Thus, if the single-hop questions are answered correctly

Group	K / R	n	Dist.	Position
Success	✓ / ✓	312	7.85	Interior
Reasoning Failure	✓ / ✗	311	5.33	Interior
Knowledge Failure	✗ / ✗	281	1.05	Exterior
Recovered Success	✗ / ✓	96	5.29	Interior

Table 3: MQuAKE groups by absolute distance to the knowledge boundary, where K and R denote knowledge accessibility (*single-hop*) and reasoning (*multi-hop*) success, respectively.



(a) Accuracy vs. distance (b) Correctness stability (c) Answer consistency

Figure 3: Behavior of knowledge accessibility for Llama-8B on Trivia QA. (a) Accessibility decreases with distance from the center. (b) Answerability is least stable near the boundary. (c) Consistency decreases with distance from the center. Results for Qwen14B are provided in B.2 in Appendix.

but the multi-hop question is answered incorrectly, the failure is more likely due to reasoning rather than missing knowledge. In contrast, failure on the required single-hop facts provides evidence of limited knowledge accessibility. We then apply the A-Centered geometry learned on TriviaQA directly to MQuAKE without relearning, which makes MQuAKE an independent test of the learned geometry. (see Appendix B.2 for details).

As shown in Table 3, geometric position is more closely aligned with knowledge accessibility than with final correctness. Both *Success* and *Reasoning Failure* cases have accessible single-hop facts, and both remain inside the boundary, even though the latter fail on the final multi-hop question. In contrast, *Knowledge Failure* cases, where at least one required fact cannot be recalled, move into the exterior. This pattern suggests that queries with similar knowledge accessibility remain close in distance despite different reasoning outcomes. Interestingly, *Recovered Success* cases also remain inside the boundary. This suggests that failure on an isolated single-hop probe does not always mean that the required knowledge is inaccessible. Some of these accessibility failures may instead reflect sensitivity to prompting or decoding, which we study later.

4.4 Behavior of Knowledge Accessibility

We next test whether the learned geometry captures meaningful behavior beyond binary correctness. In particular, we study how knowledge accessibility changes with distance and how stable model behavior is around the learned boundary.

Accessibility Monotonicity. We next visualize how knowledge accessibility changes across the learned geometry. Figure 3a shows a low-dimensional projection of the hidden representations and the learned boundary on TriviaQA. Since TriviaQA mainly tests factual recall, accuracy provides a proxy for knowledge accessibility. Queries closer to the center tend to have higher accessibility, while accessibility decreases toward and beyond the boundary. This pattern is consistent with the learned accessibility structure.

Correctness Stability. We next study how stable correctness is around the learned boundary. For each query x, we sample N responses under a fixed stochastic decoding strategy π. Let $c_\pi^{(i)}(x)$ denote the correctness of the i-th sampled response. We define

$$\hat{p}_\pi(x) = \frac{1}{N}\sum_{i=1}^{N} c_\pi^{(i)}(x), \qquad S_\pi(x) \triangleq |2\hat{p}_\pi(x) - 1|\,. \tag{6}$$

Here, $S_\pi(x) \in [0, 1]$ is the correctness stability. A value close to 1 means that repeated generations are consistently correct or consistently incorrect, while a value close to 0 means that correctness varies across generations.

Figure 3b shows a U-shaped pattern of stability around the learned boundary. Queries deep in the interior are stably answered correctly. Stability decreases toward the boundary, where the same query can be answered correctly in some generations and incorrectly in others. Beyond the boundary, stability increases again, but now mainly reflects repeated failure. Together with the decrease in accuracy, this gives the following behavioral pattern: stable access $\longrightarrow$ unstable access $\longrightarrow$ stable failure. The boundary therefore marks a region where access to the required knowledge is least stable. Since repeated generations are not used to learn the boundary, this provides independent evidence for the learned geometry.

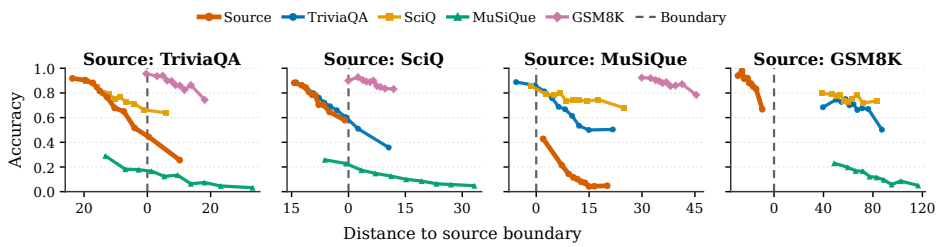


Figure 4: Transferability of the distance and accuracy relation for Llama-8B. Accuracy generally decreases with distance across target datasets. Results for Qwen-14B are provided in Figure B.3 in Appendix.

Answer Consistency. Figure 3c provides a complementary view through answer consistency. On average, repeated answers become less consistent as distance increases, which suggests that knowledge access becomes less reliable farther from the accessibility center. Some uncertainty methods rely on answer consistency across repeated generations. A limitation of such signals is that they cannot identify *confidently wrong* samples (Farquhar et al., 2024), while the centered geometry provides richer signals.

4.5 Transferability of Accessibility Distance

We next study whether distance to an accessibility center learned on one dataset can transfer to other datasets. We do not require the learned center to be the same across datasets. Instead, we ask whether distance under a geometry learned on one dataset can still preserve the accessibility ordering of queries from another dataset. Full experimental details are provided in Appendix B.4.

For each source dataset, we learn an accessibility center $\boldsymbol{c}$ and its dimension weights $\boldsymbol{A}$, and apply them directly to the target datasets without relearning. As shown in Figure 4, task performance generally decreases as the distance from the source accessibility center increases across different source-target pairs. This transfer is stronger on datasets where correctness mainly reflects knowledge accessibility. On reasoning datasets (e.g., GSM8K), the distance still preserves a useful ordering, but the relation is weaker. Overall, queries farther from the source center tend to have lower knowledge accessibility.

However, the accessibility centers $\boldsymbol{c}$ and the dimension weights $\boldsymbol{A}$ do not transfer across datasets, as shown in Figure B.4 in the Appendix. This suggests that different types of task have their own accessibility centers. Even so, the monotonic relation between distance and knowledge accessibility remains consistent across datasets. Queries farther from the source center still tend to have lower knowledge accessibility.

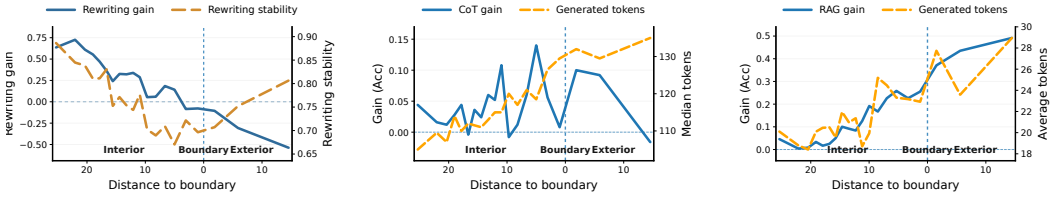
5 Knowledge Accessibility Interventions

So far, we have shown that distance is related to knowledge accessibility. Importantly, this distance is computed solely from the query representation, before any answer tokens are generated. For a query x, we extract its hidden representation $\boldsymbol{h}(x)$ and measure its distance to the learned accessibility center, $D(x) = d(\boldsymbol{h}(x), \boldsymbol{c})$.

We next ask whether effective knowledge accessibility can be improved at inference time, without retraining or modifying the model parameters. We consider three complementary interventions that operate through different mechanisms. (1) *Query rewriting* changes the wording of the query without changing or adding new information, (2) *chain-of-thought (CoT)* provides additional computation, and (3) *retrieval-augmented generation (RAG)* provides external information.

We study these interventions on TriviaQA, where factual recall provides a relatively clean setting for examining their interaction with knowledge accessibility. Figure 5 shows that these interventions are effective in different accessibility regions. Query rewriting yields its largest average gain in the interior, additional computation around the boundary, and external retrieval in the exterior. Experimental details are provided in Appendix C.

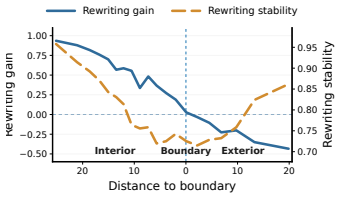
Query rewriting is most useful when knowledge is already accessible. Query rewriting does not add new information or additional reasoning computation. Instead, it changes how the query



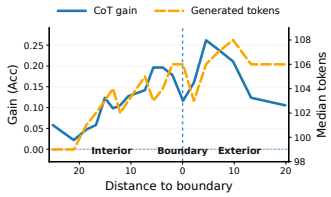
(a) Llama-8B: Rewriting

(b) Llama-8B: CoT

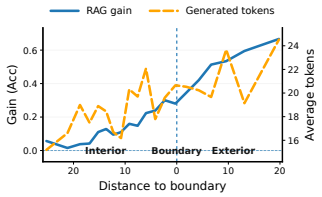
(c) Llama-8B: RAG



(d) Qwen-14B: Rewriting



(e) Qwen-14B: CoT



(f) Qwen-14B: RAG

Figure 5: Intervention effects across the knowledge boundary on TriviaQA.

accesses knowledge already stored in the model's parameters. On TriviaQA, Figure 5a shows that rewriting has its largest average gain in the interior, where the required knowledge is more likely to be parametrically accessible.

This suggests that some failures in the interior are from unsuccessful access rather than missing knowledge. Paraphrasing the same query can provide an alternative access path and recover the correct answer. As queries approach the boundary, access becomes more sensitive to wording. Correctness varies more across semantically equivalent paraphrases, and paraphrase stability decreases. Farther into the exterior, rewriting becomes less effective because changing the access path alone cannot compensate for increasingly unreliable parametric access.

Prior work has shown that failure under one prompt does not necessarily imply missing knowledge, since paraphrased prompts can recover correct responses (Jiang et al., 2020). Our results further relate this phenomenon to knowledge accessibility. Query rewriting is most effective when the required knowledge is already within the model's accessible region, but its benefit diminishes as parametric access weakens.

Additional computation is most useful near the boundary. Unlike query rewriting, CoT does not change the query provided to the model. Instead, it allocates additional computation to accessing and using knowledge already available from the model's parameters. On TriviaQA, Figure 5b shows that CoT yields its largest average accuracy gain near the knowledge boundary.

This pattern suggests that extra computation is most useful near the knowledge boundary, where the required knowledge is still accessible but access is unstable. Deep in the interior, direct access is already reliable, so there is little room for improvement. Near the boundary, the required knowledge is still available, but direct access is unstable. CoT gives the model more steps to retrieve and use this knowledge, so it has more room to improve than in the interior. In the exterior, parametric access becomes unreliable. Since CoT adds computation but no new information, it cannot recover knowledge that the model cannot reliably access, so the gain decreases even as more tokens are generated.

Existing work improves reasoning efficiency and reduces inference cost by adapting compute to the amount of reasoning or thinking required by a query (Wang et al., 2025; Yong et al., 2025b; Li et al., 2026). Our findings complement this line of work by showing that knowledge accessibility can serve as a pre-generation signal for when additional computation is likely to help.

External retrieval is most useful outside the boundary. RAG differs from the previous interventions because it does not rely only on knowledge already available from the model's parameters. Instead, it supplies external information when parametric access is weak. On TriviaQA, Figure 5c shows that the average benefit of RAG increases with knowledge distance and is largest in the exterior.

As queries move farther outside the accessible region, the required knowledge becomes less reliably available from the model's parameters, which makes external information increasingly useful.

This finding is closely related to adaptive RAG, which aims to decide when retrieval should be invoked at inference time. Existing methods make this decision using signals such as generation confidence, self-reflection, or query complexity (Gao et al., 2023; Jiang et al., 2023; Jeong et al., 2024). Our results complement this line of work by showing that, in factual QA, the geometry of knowledge accessibility provides a pre-generation signal for when external retrieval is likely to help.

A unified pattern of accessibility interventions. The three interventions affect knowledge access in different ways, and their benefits vary across the accessibility space. In the *interior*, where parametric access is reliable, query rewriting can recover failures caused by how the query is expressed. Near the *boundary*, where access becomes unstable, additional reasoning yields its largest gains. In the *exterior*, where parametric access is unreliable, external retrieval becomes more useful.

These results reveal a clear shift. As queries move from the *interior* (reliable access), through the *boundary* (unstable access), to the *exterior* (unreliable access), the useful strategy shifts from query rewriting $\rightarrow$ additional computation $\rightarrow$ external retrieval. Because knowledge distance is available before generation, this geometry may provide a simple signal for deciding how a model should intervene on its own inference process, which may help reduce inference cost (Wang et al., 2026).

6 Related Work

Prior work has studied the knowledge boundary of large language models, i.e., whether a model can distinguish queries it can answer from those it cannot (Kadavath et al., 2022; Yin et al., 2023; 2024; Li et al., 2025). Several studies have further used LLM internal states to probe knowledge boundaries (Slobodkin et al., 2023; Xiao et al., 2025). Some of them estimate the boundary from the query alone before generation, which provides a useful signal for deciding how to answer a query (Ni et al., 2025; Chen et al., 2026).

However, these studies predict observed correctness directly, without decoupling whether errors are from limited knowledge recall or from failures in reasoning steps (Press et al., 2023; Li & Goyal, 2025; Yang et al., 2026; Cheang et al., 2026; Wang et al., 2026). More importantly, they focus mainly on whether knowledge boundary information is encoded in hidden states, rather than how such information is organized geometrically in representation space. In this work, we separate knowledge accessibility from reasoning and study the geometry of knowledge accessibility in query representations.

7 Conclusion

This work shifts the study of the knowledge accessibility from whether accessibility is encoded in LLM representations to how it is geometrically organized. We find that accessible queries lie close to a common center. Accessibility drops as queries move farther from it. It also tracks what the model can access rather than whether it reasons correctly. Across datasets, the centers differ but the accessibility ordering transfers. The geometry also indicates what kind of help a query needs. Rewriting helps most in the interior, reasoning helps most near the boundary, and retrieval helps most beyond it. Because this signal is available before generation, it suggests that models could choose whether to answer directly, reason more, or retrieve information (Chen et al., 2026).

Our study has limitations that suggest several directions for future work. We focus on single-fact queries and use correctness as a behavioral proxy for accessibility. We do not separate knowledge that is never stored from knowledge that is stored but not reached (Allen-Zhu & Li, 2024). Our evidence is also correlational. Whether moving a representation toward the accessibility center can causally impact recall remains an open question. Other open questions are how the center emerges from pre-training (Papyan et al., 2020) and whether the geometry can be extended to reasoning capability.

AI USE STATEMENT

We used generative AI tools to assist with literature search, to polish the writing, and to check grammar. All references were verified by the authors. In addition, as part of the experimental methodology, `Mistral-Large` was used as a correctness judge and to generate query paraphrases (Section 4 and Appendix C).

REPRODUCIBILITY STATEMENT

All models (`Llama-3.1-8B-Instruct` and `Qwen-2.5-14B-Instruct`) and datasets (TriviaQA, SciQ, MuSiQue, GSM8K, and MQuAKE) used in this work are publicly available. The experimental setup is described in Section 4, and full training details, hyperparameters, data splits, and random seeds are provided in Appendix A. Details of the controlled MQuAKE evaluation, the repeated-sampling analysis, the cross-dataset transfer analysis, and the intervention experiments are given in Appendices B.2, B.3, B.4, and C, respectively. We will release our code upon publication.

REFERENCES

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=5x788rqbcj.

Chi Seng Cheang, Hou Pong Chan, Wenxuan Zhang, and Yang Deng. Do LLMs really know what they don't know? internal states mainly reflect knowledge recall rather than truthfulness. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 713–730, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.34. URL https://aclanthology.org/2026.findings-acl.34/.

Lihu Chen, Adam Dejl, and Francesca Toni. Identifying query-relevant neurons in large language models for long-form texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23595–23604, 2025.

Lihu Chen, Gerard de Melo, Fabian M. Suchanek, and Gaël Varoquaux. Query-level uncertainty in large language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=11QZITAMUO.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8493–8502, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.581. URL https://aclanthology.org/2022.acl-long.581/.

Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 55–65, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1006. URL https://aclanthology.org/D19-1006/.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

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

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7036–7050, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.389. URL https://aclanthology.org/2024.naacl-long.389/.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020. doi: 10.1162/tacl_a_00324. URL https://aclanthology.org/2020.tacl-1.28/.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.495. URL https://aclanthology.org/2023.emnlp-main.495/.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Aochong Oliver Li and Tanya Goyal. Memorization vs. reasoning: Updating LLMs with new knowledge. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25853–25874, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1326. URL https://aclanthology.org/2025.findings-acl.1326/.

Moxin Li, Yong Zhao, Wenxuan Zhang, Shuaiyi Li, Wenya Xie, See-Kiong Ng, Tat-Seng Chua, and Yang Deng. Knowledge boundary of large language models: A survey. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5131–5157, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.256. URL https://aclanthology.org/2025.acl-long.256/.

Zheng Li, Qingxiu Dong, Jingyuan Ma, Di Zhang, Kai Jia, and Zhifang Sui. SelfBudgeter: Adaptive token allocation for efficient LLM reasoning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 21135–21156, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1063. URL https://aclanthology.org/2026.findings-acl.1063/.

Shiyu Ni, Keping Bi, Jiafeng Guo, Lulu Yu, Baolong Bi, and Xueqi Cheng. Towards fully exploiting LLM internal states to enhance knowledge boundary perception. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24315–24329, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1184. URL https://aclanthology.org/2025.acl-long.1184/.

Vardan Papyan, Xiao Y Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2463–2473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250. URL https://aclanthology.org/D19-1250/.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.378. URL https://aclanthology.org/2023.findings-emnlp.378/.

Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5418–5426, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.437. URL https://aclanthology.org/2020.emnlp-main.437/.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pp. 4393–4402. PMLR, 2018.

Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. The curious case of hallucinatory (un)answerability: Finding truths in the hidden states of over-confident large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3607–3625, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.220. URL https://aclanthology.org/2023.emnlp-main.220/.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

Xinglin Wang, Shaoxiong Feng, Yiwei Li, Peiwen Yuan, Yueqi Zhang, Chuyi Tan, Boyuan Pan, Yao Hu, and Kan Li. Make every penny count: Difficulty-adaptive self-consistency for cost-efficient reasoning. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 6919–6932, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7.

doi: 10.18653/v1/2025.findings-naacl.383. URL https://aclanthology.org/2025.findings-naacl.383/.

Zixuan Wang, Yinze Ding, Zihan Wang, Jinyu Guo, Zhenhong Zhou, Junhao Dong, and Chaomeng Chen. X-router: Decoupling knowledge and reasoning for cost-effective LLM inference. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 19856–19874, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.994. URL https://aclanthology.org/2026.findings-acl.994/.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In Leon Derczynski, Wei Xu, Alan Ritter, and Tim Baldwin (eds.), *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4413. URL https://aclanthology.org/W17-4413/.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pp. 499–515. Springer, 2016.

Robert Wu and Vardan Papyan. Linguistic collapse: Neural collapse in (large) language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=G0LfcMiRkc.

Chenghao Xiao, Hou Pong Chan, Hao Zhang, Mahani Aljunied, Lidong Bing, Noura Al Moubayed, and Yu Rong. Analyzing LLMs' knowledge boundary cognition across languages through the lens of internal representations. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24099–24115, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1174. URL https://aclanthology.org/2025.acl-long.1174/.

Mutian Yang, Jiandong Gao, and Ji Wu. Decoupling knowledge and reasoning in llms: An exploration using cognitive dual-system theory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34268–34276, 2026.

Xunjian Yin, Xu Zhang, Jie Ruan, and Xiaojun Wan. Benchmarking knowledge boundary for large language models: A different perspective on model evaluation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2270–2286, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.124. URL https://aclanthology.org/2024.acl-long.124/.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8653–8665, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.551. URL https://aclanthology.org/2023.findings-acl.551/.

Xixian Yong, Xiao Zhou, Yingying Zhang, Jinlin Li, Yefeng Zheng, and Xian Wu. Think or not? exploring thinking efficiency in large reasoning models via an information-theoretic lens. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 2787–2827. Curran Associates, Inc., 2025a. doi: 10.52202/085713-0098. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/04185b5ae2d450ef39bd53c0ec4802cb-Paper-Conference.pdf.

Xixian Yong, Xiao Zhou, Yingying Zhang, Jinlin Li, Yefeng Zheng, and Xian Wu. Think or not? exploring thinking efficiency in large reasoning models via an information-theoretic lens. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=DpOSndSOZz.

Zexuan Zhong, Zhengxuan Wu, Christopher Manning, Christopher Potts, and Danqi Chen. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15686–15702, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.971. URL https://aclanthology.org/2023.emnlp-main.971/.

A Implementation Details

A.1 Representations and Correctness Labels

Query representations. For each query, the model receives the query alone, and we extract the hidden state $\boldsymbol{h}(x)$ at the final non-padding query token before any answer token is generated.

Standardization. Before fitting any geometry, each representation dimension is standardized with the mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$ computed on the training split. All geometries operate on standardized representations, so throughout the paper $\boldsymbol{h}(x)$ denotes the standardized representation.

Correctness labels. Correctness $c_\pi(x)$ (Eq. 1) is decided by `Mistral-Large`, used as a semantic-equivalence judge with temperature zero. Given the query x, a response, and the reference answer set $\mathbb{Y}(x)$ including its aliases, the judge decides whether the response expresses one of the reference answers. Unless stated otherwise, π is greedy decoding.

A.2 Geometry Models and Training

The four accessibility scores are listed in Table 1. They are trained as follows.

Linear. Logistic regression with inverse regularization strength $C = 1$ and the L-BFGS solver.

MLP. One hidden layer with $m = 128$ units, ReLU activation, and dropout (Srivastava et al., 2014) of 0.1. We use a batch size of 128 and early stopping with a patience of 10 epochs based on validation AUROC.

Centered. The center $\boldsymbol{c}$ is initialized to the mean representation of the correctly answered training queries and trained for 600 epochs with a soft-margin objective. The margin is selected from $\{5, 7, 10, 15, 20\}$ by validation AUROC, separately for each layer and random seed.

A-Centered. A-Centered uses the distance in Eq. 3 with $\boldsymbol{A}$. The dimension weights are positive and have geometric mean one, which removes their arbitrary global scale. Training has two stages. We first fit Centered and use its learned center to initialize A-Centered. We then train A-Centered for 300 epochs. The margin is inherited from the validation-selected Centered model.

Optimization. MLP, Centered, and A-Centered are trained with Adam (Kingma & Ba, 2014) using a learning rate of 10^{-2}.

Data splits and seeds. The 40%/10%/50% train/validation/test splits (Section 4) are stratified by the correctness label. For each seed, the same split is shared by all layers and all methods. The five seeds are $\{42, 43, 44, 45, 46\}$.

B Details of the Geometry Analyses

This section follows the order of Section 4.

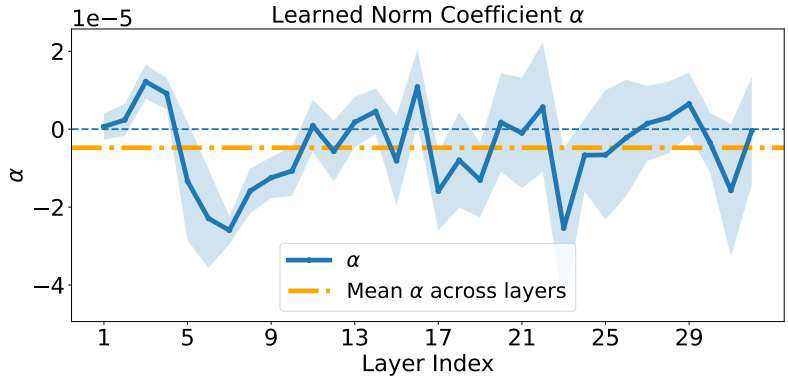


Figure B.1: Learned norm coefficient α in Linear+Norm across layers on TriviaQA with Llama-8B. The line shows the mean over five random seeds and the shaded region shows ± 1 standard deviation.

B.1 Testing the Centered Inductive Bias

This section complements the Linear+Norm comparison in Section 4.2. Linear+Norm uses the same splits as the other geometries, and all preprocessing statistics are fitted on the training split only. It is trained as a regularized logistic model, with the inverse regularization strength selected from $C \in \{0.01, 0.1, 1, 10\}$ by validation AUROC.

Figure B.1 shows the learned coefficient α across layers. Although α varies across layers, its average is negative. For $\alpha < 0$, completing the square gives

$$\boldsymbol{w}^\top \boldsymbol{h} + \alpha \|\boldsymbol{h}\|_2^2 + b = -|\alpha| \left\|\boldsymbol{h} - \tilde{\boldsymbol{c}}\right\|_2^2 + \text{const}, \qquad \tilde{\boldsymbol{c}} = -\frac{\boldsymbol{w}}{2\alpha}, \tag{B.1}$$

which is equivalent to the Centered score after positive rescaling, with an implied center $\tilde{\boldsymbol{c}}$. Thus, even without an explicit center, Linear+Norm tends to learn a centered structure. Together with the stronger performance of Centered, this supports the centered structure as a useful inductive bias rather than an effect of greater model capacity.

B.2 Controlled MQuAKE Evaluation

We use MQuAKE to separate failures in knowledge accessibility from failures in reasoning (Table 3). Each example contains a multi-hop query together with the single-hop questions that probe the facts required to answer it. We evaluate 1,000 examples with `Llama-3.1-8B-Instruct`.

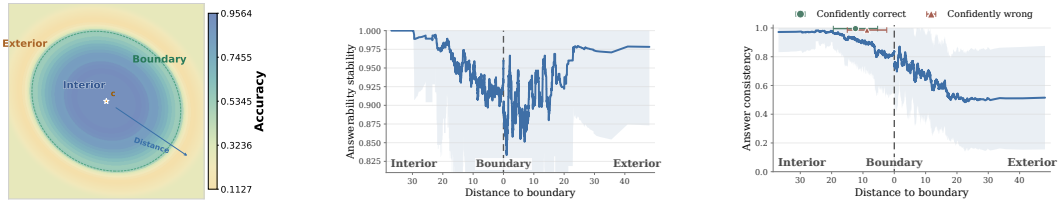
Behavioral groups. All single-hop questions and the multi-hop query are answered with the same prompt and greedy decoding. Correct recall of all single-hop questions is used as a behavioral proxy for accessible knowledge (K), and correctness of the multi-hop query as reasoning success (R). Combining the two outcomes gives four groups: *Success* (all single-hop questions and the multi-hop query correct), *Reasoning Failure* (all single-hop questions correct, multi-hop query incorrect), *Knowledge Failure* (at least one single-hop question and the multi-hop query incorrect), and *Recovered Success* (at least one single-hop question incorrect, multi-hop query correct).

Frozen TriviaQA geometry. No geometry is fitted on MQuAKE. We freeze the A-Centered geometry learned on TriviaQA. Each canonical MQuAKE query is represented exactly as in Appendix A.1 at the selected TriviaQA layer, and its distance is computed under the frozen geometry.

B.3 Behavior of Knowledge Accessibility

This section gives the details for Section 4.4; the results for Qwen-14B are shown in Figure B.2.

Frozen geometry on held-out queries. All analyses use test queries only. Their distances are computed under the geometry learned on the training split.



(a) Accuracy vs. distance (b) Correctness stability (c) Answer consistency

Figure B.2: Behavior of knowledge accessibility for Qwen-14B on TriviaQA (counterpart of Figure 3). (a) Accessibility decreases with distance from the center. (b) Correctness is least stable near the boundary. (c) Consistency decreases with distance from the center.

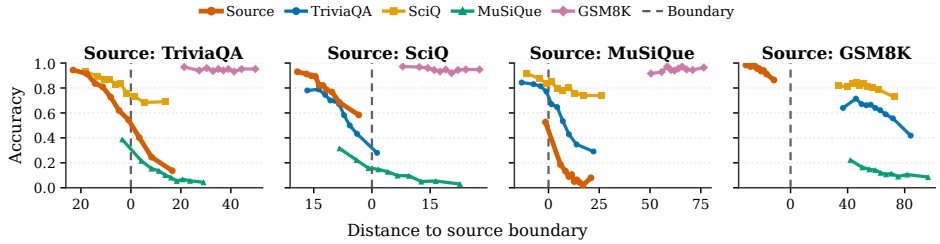


Figure B.3: Transferability of the distance and accuracy relation for Qwen-14B (counterpart of Figure 4). Accuracy generally decreases with distance across target datasets.

Repeated sampling and correctness stability. For each test query, we draw N independent responses under a fixed stochastic decoding strategy π. Each response is judged independently as in Appendix A.1, which gives $c_\pi^{(i)}(x)$. The empirical correctness probability $\hat{p}_\pi(x)$ and the correctness stability $S_\pi(x)$ then follow Eq. 6.

Answer consistency. Answer consistency measures whether repeated generations produce the same surface-form answer, regardless of whether that answer is correct. For visualization, queries with $\kappa_\pi(x) \geq 0.9$ are marked as *confidently correct* if their modal normalized answer is judged correct, and as *confidently wrong* otherwise. These categories are only used for visualization.

B.4 Cross-Dataset Transfer and Dataset-Specific Geometry

This section gives the details for Section 4.5. We separate two questions: whether the *relation* between distance and correctness transfers across datasets, and whether the *geometric parameters* themselves (center and dimension weights) are shared.

Transfer protocol. We use TriviaQA, SciQ, MuSiQue, and GSM8K, with the best-performing layer selected for Llama-8B. Each dataset is used in turn as the source s. Its A-Centered geometry is applied without relearning.

Distance–accuracy relation. Queries of a target dataset t are grouped into distance bins B_k under the source geometry. For each bin we compute the mean distance and the accuracy,

$$\bar{D}_k = \frac{1}{|B_k|} \sum_{x \in B_k} D_s(x), \qquad \mathrm{Acc}(B_k) = \frac{1}{|B_k|} \sum_{x \in B_k} c(x), \tag{B.2}$$

which gives the curves in Figures 4 and B.3. The relation is summarized by the Spearman correlation

$$\rho_{s \to t} = \rho_{\mathrm{Spearman}}\big(\{\bar{D}_k\}_k, \{\mathrm{Acc}(B_k)\}_k\big). \tag{B.3}$$

A negative value means that target queries farther from the source center are answered less accurately. This only requires the source geometry to preserve the relative ordering of target queries, not to share their center or shape.

Cross-dataset diagnostics. Figure B.4 reports four pairwise matrices. In (a) and (b), rows are source datasets and columns are target datasets. In (c) and (d), entries compare the geometries learned independently on each pair of datasets.

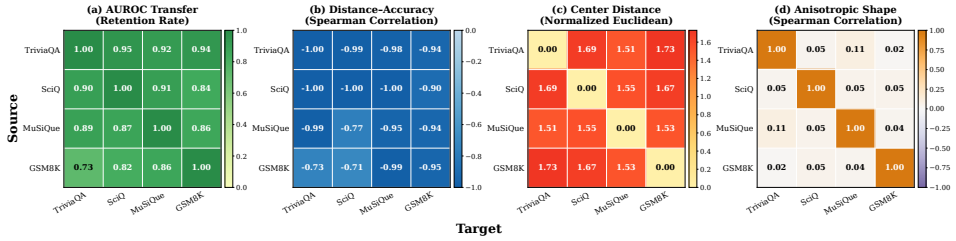


Figure B.4: Cross-dataset diagnostics for Llama-8B (layer 18). (a) AUROC retention, (b) distance–accuracy correlation, (c) distance between learned centers, and (d) correlation between learned dimension weights. The distance-based ordering transfers across datasets even though the centers and dimension weights are dataset-specific.

(a) **AUROC retention.** The transfer AUROC is normalized by the AUROC of the target's own geometry, $\text{Retention}_{s\to t} = \text{AUROC}_{s\to t}/\text{AUROC}_{t\to t}$.

(b) **Distance–accuracy monotonicity.** Entry (s, t) is $\rho_{s\to t}$ from Eq. B.3.

(c) **Center variation.** Since each geometry is fitted after dataset-specific standardization, we first map the learned centers back to the original representation space. We then measure the distance between centers and normalize it by the overall scale of the representations. Smaller values indicate more similar centers.

(d) **Dimension-weight variation.** We similarly map the learned dimension weights back to the original representation space and compare them using Spearman correlation. A high correlation means that the two geometries emphasize similar dimensions.

Interpretation. Panels (a) and (b) show that transferred geometries retain much of the target-specific AUROC and that distance–accuracy correlations remain negative for all source–target pairs. Panels (c) and (d) show that the learned centers and dimension weights differ across datasets.

C Intervention Experiments

This section provides the experimental details for Section 5.

C.1 Shared Setup

Geometry and queries. All interventions are evaluated on the TriviaQA test set. We use the same frozen A-Centered geometry and the same selected layer as in the main experiments.

Evaluation. For each query, we compare the correctness of the original greedy answer with the correctness after an intervention. Correctness is judged as in Appendix A.1. For CoT and RAG, only the final extracted answer is given to the judge, without the reasoning trace or retrieved context.

Binning. Queries are grouped into 20 equal-frequency bins according to their signed distance to the learned boundary. Negative distance corresponds to the interior side, while positive distance corresponds to the exterior side. All curves report raw bin-level statistics without smoothing or interpolation.

C.2 Query Rewriting

Paraphrase generation. We use `Mistral-Large` with temperature 0.8 to generate five semantically equivalent paraphrases for each query. The paraphrases are instructed to preserve the original meaning, while changing wording and sentence structure.

Answering and evaluation. The model answers each paraphrase independently using only its internal knowledge. We use greedy decoding with at most 128 new tokens and do not use chain-of-thought or retrieval. We keep queries for which all five paraphrases receive valid correctness judgments.

For each original query, we compute the fraction of paraphrases answered correctly. We also measure paraphrase stability from the consistency of the five correctness outcomes. The rewriting gain reported in Figure 5 is the change in correctness relative to the original query.

C.3 CHAIN-OF-THOUGHT

The model is prompted to reason step by step using only its internal knowledge. We allow at most five short reasoning steps before the final answer. Generation uses greedy decoding with at most 2,048 new tokens.

This intervention adds computation but no external information. We record both answer correctness and the number of generated tokens, which allows us to compare the benefit and cost of additional reasoning across different knowledge-boundary regions.

C.4 RETRIEVAL-AUGMENTED GENERATION

For retrieval, we use the evidence provided by TriviaQA. Each test question is matched to its corresponding example in the TriviaQA `rc` validation split. We use the highest-ranked non-empty `search_context` and truncate it to at most 4,096 tokens.

The retrieved context is prepended to the question, and the model is instructed to answer using it as external evidence. We do not request chain-of-thought so that the intervention mainly reflects the effect of additional information rather than additional reasoning. Generation uses greedy decoding with at most 256 new tokens.

Queries without a matching TriviaQA example or usable retrieval context are excluded from the RAG analysis.