Masked by Consensus: Disentangling Privileged Knowledge in LLM Correctness

Anonymous ACL submission

Abstract

Humans use introspection to evaluate their understanding through private internal states inaccessible to external observers. We investigate whether Large Language Models possess similar *privileged knowledge*—information unavailable through external observation. Specifically, we ask whether models have unique signals about answer correctness given only the question. We train correctness classifiers on question representations from both a model's own hidden states and external models, testing whether self-representations provide a performance advantage. Standard evaluations show no advantage: self-probes perform comparably to peer-model probes, which we attribute to high inter-model agreement. To isolate genuine privileged knowledge, we evaluate on *disagreement subsets* where models produce conflicting predictions. Here, self-representations consistently outperform peer representations in factual knowledge tasks, but mathematical reasoning shows no advantage. Our findings reveal domain-specific privileged knowledge: models possess genuine self-signals about factual correctness, while mathematical reasoning correctness appears universally observable. We explore potential mechanisms underlying this distinction, finding evidence that factual correctness relies on entity-driven memory retrieval while mathematical correctness may involve more universal computational patterns.

1 Introduction

In the philosophy of mind, *epistemic privilege* refers to the idea that an agent has special access to its own internal states—information that cannot be fully recovered from external observation alone (Alston, 1971; Gertler, 2010). Inspired by this notion, recent research suggests that LLMs encode meta-information about their own outputs, ranging from entity recognition (Ferrando et al., 2024) and temperature inference (Comsa and Shanahan, 2025) to the representation of cognitive-like states (Chen et al., 2025; Ji-An et al., 2025). A central aspect of this meta-information is *output correctness*: numerous studies have demonstrated that output correctness can be predicted with high accuracy (Kadavath et al., 2022), primarily via linear probes trained on internal hidden states (Cencerrado et al., 2025; Seo et al., 2025). This raises a fundamental question: do LLMs have internal correctness signals that are inaccessible to external models—in other words, *privileged knowledge* about whether their answer will be correct?

Recent findings cast doubt on the existence of privileged knowledge in the context of correctness prediction. Chi et al. (2025) argue that probes primarily detect retrieval activation patterns rather than correctness signals, while Seo et al. (2025) and Xiao et al. (2025) show that external models can achieve prediction performance comparable to methods that rely on a model's own internal representations, suggesting little to no privileged information exists.

In this paper, we argue that prior conclusions about the absence of privileged knowledge may be premature due to confounded evaluation. Specifically, when external models can exploit proxy signals from shared correctness patterns, genuine privileged knowledge—if it exists—may be masked. To test whether privileged knowledge exists, we measure the *premium gap*: the performance advantage of a correctness classifier trained on a model's *own* internal representations over one trained on *external* model representations (Figure 1). However, this gap may vanish on random samples due to high inter-model agreement. If models often succeed or fail on the same questions, probes trained on external representations can exploit the external model's own correctness patterns as a *proxy* for the target model's behavior, obscuring whether the target model possesses unique internal signals about its correctness.

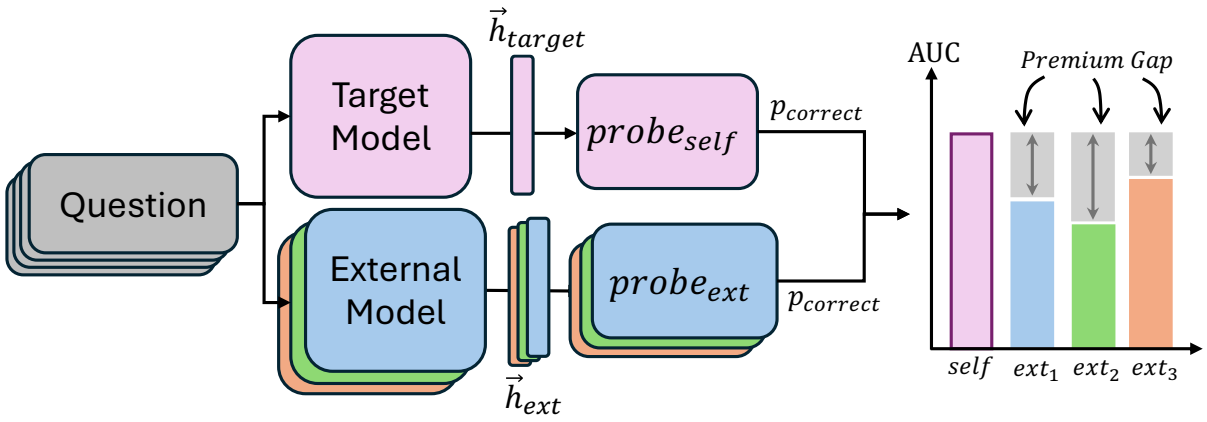

Figure 1: **Overview of the experimental framework.** Questions are input to a target model and to external models, yielding representations $\mathbf{h}_{\text{target}}$ and $\mathbf{h}_{\text{ext}}$. Probes trained on these representations predict answer correctness. We evaluate probe performance using mean AUC averaged over layers and define the *premium gap* as the performance advantage of self over external probes.

To address this challenge, we evaluate on *disagreement subsets*—questions where models produce conflicting correctness labels. By restricting evaluation to these subsets, we eliminate shared correctness patterns and isolate each model's unique behavior, enabling a direct test of whether self-representations contain privileged information unavailable from peer models.

We use this methodology to systematically investigate the existence of privileged knowledge in the context of correctness prediction across five datasets spanning factual knowledge (Mintaka, TriviaQA, HotPotQA) and mathematical reasoning (GSM1K, MATH), using three similar-sized models (Qwen-2.5-7B, Llama-3.1-8B, Gemma-2-9B).

Our results show that measuring the premium gap on a random sample is insufficient to establish privileged knowledge. While a premium gap persists for one model, its high success in predicting peer models suggests this reflects superior general representations rather than private information. However, when evaluated on a disagreement subset, a statistically significant premium gap re-emerges in factual knowledge domains ($\sim$5%) across all models, providing evidence for privileged knowledge. In contrast, mathematical reasoning shows no premium gap, as external models achieve performance comparable to self-probes even under disagreement (Figure 3).

We summarize our main findings as follows:

- We systematically evaluate correctness prediction across five datasets and three model families, demonstrating that the premium gap effectively vanishes when tested against strong external model baselines.
- We identify *inter-model agreement* as a critical confounder: probes leverage shared difficulty patterns to predict correctness without needing access to the target's internal state.
- We introduce evaluation on *disagreement subsets* to isolate internal signals, revealing that genuine privileged knowledge is domain-specific: it emerges in factual tasks but remains absent in mathematical reasoning.

2 Related Work

LLM correctness knowledge. Previous research has explored whether LLMs can evaluate their own knowledge and correctness. For instance, Kadavath et al. (2022) demonstrate that large models are well-calibrated on multiple-choice and true/false tasks, accurately estimating the probability that a given answer is correct. Similarly, Yin et al. (2023) introduce the "SelfAware" dataset of unanswerable questions and show that modern LLMs exhibit an intrinsic capacity for knowledge, successfully identifying when a question cannot be answered. Together, these findings suggest that LLMs possess a degree of self-evaluation ability, motivating direct investigation of their hidden states for signals of correctness.

LLM introspection. Recent work investigates whether models possess privileged access to their internal processes, often termed introspection. Li et al. (2025a) find that models fine-tuned to explain their own internal computations—such as feature encoding and causal structure—outperform external explainers, suggesting a unique capacity for

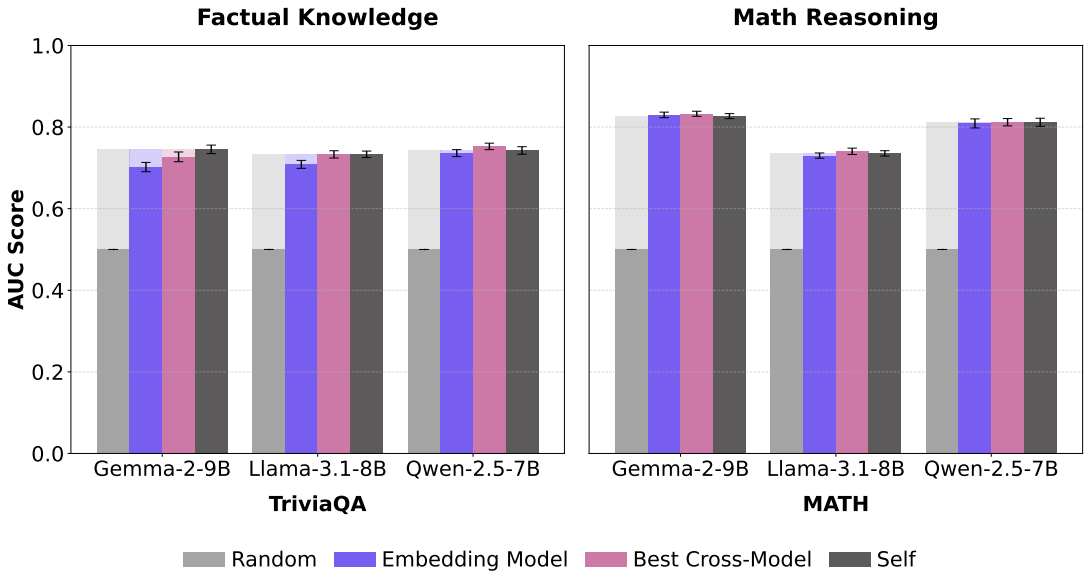

Figure 2: **Premium Gap.** Mean AUC for correctness prediction, averaged over layers, on two task types: factual knowledge (TriviaQA) and mathematical reasoning (MATH). Bars compare Random, Embedding, and Best Cross-Model baselines to the Self-Probe (*Self*) across three target models. Semi-transparent overlays indicate the performance gain (or lack thereof) of *Self* relative to each baseline. Error bars denote 95% confidence intervals.

self-explanation. Similarly, Binder et al. (2025) define introspection as knowledge derived from internal states, showing that models trained on their own behavior predict their hypothetical choices more accurately than third-party models. However, the reliability of this access is debated. Li et al. (2025b) argue that verbalized explanations often reflect the model's parametric knowledge rather than a faithful decoding of internal states, succeeding on benchmarks even without access to internals. Furthermore, Binder et al. (2025) note that the observed self-prediction advantage is often limited to simple settings and does not consistently generalize to out-of-distribution tasks.

Using probes to predict correctness. Several recent studies use probing of hidden states to predict answer accuracy. Linear probes trained on the hidden states of reasoning models have been used to verify intermediate reasoning steps and even predict answer correctness prior to generation (Zhang et al., 2025). Similarly, Tamoyan et al. (2025) demonstrate that residual-stream features encode a "factual self-awareness" signal: simple linear projections can predict whether a model will recall a fact correctly. Orgad et al. (2025) also report that internal representations carry rich truthfulness information, concentrated in specific tokens; notably, they show that a model may internally encode the correct answer even when its generated output is incorrect.

However, other studies argue that predictive probes often exploit artifacts from the question or answer rather than reflecting genuine introspection. Seo et al. (2025) show that much of the reported probe accuracy arises from superficial question patterns. Similarly, Cencerrado et al. (2025) train probes on representation after reading the question—before generating an answer—and demonstrate that these probes outperform strong black-box baselines in predicting answer accuracy. Extending this idea, Xiao et al. (2025) propose a 'generalized correctness model' across multiple LLMs, finding that predictors trained on historical answer patterns perform comparably to model-specific probes, suggesting that LLMs have little privileged knowledge of their own correctness. This view is further supported by mechanistic analyses from Chi et al. (2025), who show that when LLMs hallucinate due to retrieving incorrect knowledge, their internal states are indistinguishable from those of correct answers, suggesting that LLMs do not explicitly encode correctness. Our objective is to reconcile these two often conflicting lines of work. To this end, we employ a rigorous experimental framework designed to uncover whether LLMs genuinely possess privileged knowledge of their own correctness.

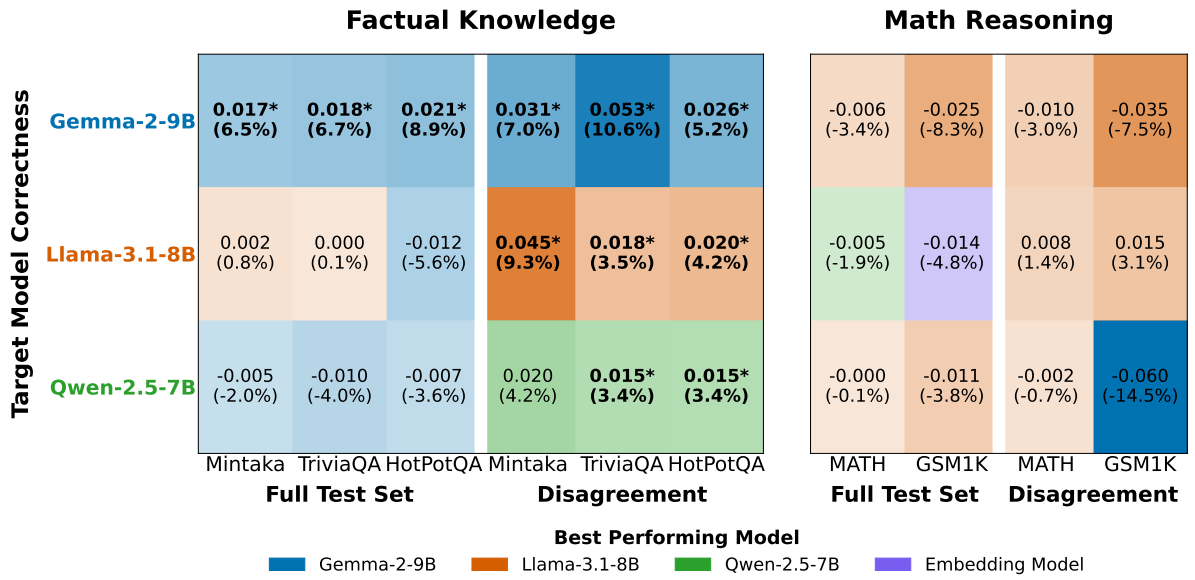

Figure 3: **Target Model Correctness Prediction.** Heatmap of correctness prediction differences across target models, datasets, and test subsets. Each cell reports the AUC difference ($\Delta\text{AUC} = \text{Self} - \text{Best External}$), with the percentage of the gap closed shown in parentheses, computed as $\frac{\text{Self}-\text{Best External}}{1-\text{Best External}} \times 100$. The y-axis lists target models, and cell colors indicate the best-performing external source model for each setting. Asterisks (*) denote statistically significant differences (paired t-test, $p < 0.05$, Bonferroni–Holm correction).

3 Methodology

3.1 Problem Formulation

Let M_{target} be the model whose correctness we wish to predict. For a given question q, M_{target} generates an answer with a binary correctness label $y \in \{0, 1\}$. We predict y using a classifier (probe) f trained on the internal hidden states $\mathbf{h}$ of a source model M_{source} processing the same question q:

$$\hat{y} = f(\mathbf{h}(q; M_{source})) \tag{1}$$

Defining Privileged Knowledge. To understand the information captured by the probe f, we posit that the hidden state $\mathbf{h}(q; M_{source})$ encodes two distinct latent components:

$$\mathbf{h} \approx \mathbf{z}_{public} \oplus \mathbf{z}_{private} \tag{2}$$

Here, $\mathbf{z}_{public}$ captures inherent features of the input question q (e.g., domain, entity types) which are universally accessible to any model processing the question. In contrast, $\mathbf{z}_{private}$ captures internal states specific to M_{source} (e.g., memory retrieval success, reasoning confidence). We define $\mathbf{z}_{private}$ as *privileged knowledge*, henceforth referring exclusively to internal states predictive of correctness.

3.2 Probing Configurations

To isolate privileged knowledge, we vary the source model M_{source} to create distinct configurations. In all cases, the probes are trained and tested on the correctness labels y of the target model (M_{target}).

1. Self-Probe. We set $M_{source} = M_{target}$. The probe is trained on the model's own representations to predict y.

2. External-Probe. We set $M_{source} \neq M_{target}$. The probe is trained on the external model's representations to predict y.

We evaluate two types of external probes:

- **Cross-Model:** M_{source} is a peer LLM of comparable size (e.g., predicting Qwen's correctness using Llama's hidden states).
- **Embedding-Model:** M_{source} is an embedding model of comparable size.

3.3 Analysis Framework

The Premium Gap. We refer to the advantage in correctness prediction performance of a self-probe over an external-probe as the *premium gap*. If privileged knowledge ($\mathbf{z}_{\text{private}}$) contains no correctness signal, then correctness prediction relies solely on public features ($\mathbf{z}_{\text{public}}$). In this case, external models with more informative representations of public features should outperform self-probes. Conversely, if a premium gap persists—where self-probes outperform all external probes—this provides evidence that the model possesses unique internal signals inaccessible to external observers.

Disagreement Subsets. To eliminate the confound of *inter-model agreement*, we evaluate performance on the disagreement subset, defined as the set of examples where M_{target} and M_{source} produce opposite correctness labels ($y_{\text{target}} \neq y_{\text{source}}$).

Crucially, we do *not* retrain probes on this subset. Training exclusively on disagreement subsets would introduce a perfect negative correlation between self and external labels. This would allow the probe to trivially exploit the external model's inverted correctness signal. Instead, we train probes on the full dataset to learn the full correctness pattern of the source model, and filter predictions during inference to strictly evaluate on the disagreement subset.

3.4 Experimental Setup

Models. We evaluate three instruction-tuned model families of comparable size: `Llama-3.1-8B` (Grattafiori et al., 2024), `Qwen2.5-7B` (Qwen et al., 2025), and `Gemma-2-9B` (Team et al., 2024), alongside the embedding model `Qwen3-Embedding-8B` (Yang et al., 2025).

Datasets. Our evaluation spans five datasets. Three focus on factual knowledge recall: Mintaka (Sen et al., 2022), TriviaQA (Joshi et al., 2017), and HotpotQA (Yang et al., 2018). Two focus on mathematical reasoning: MATH (Hendrycks et al., 2021) and GSM1K (Zhang et al., 2024). Note that while HotpotQA is often considered a multi-hop reasoning dataset, we use question-only evaluation without supporting documents, making it a test of parametric memory retrieval. See Appendix B for dataset sizes and Appendix D for evaluation protocols.

Probing method. We extract hidden states of the question from the final token of every 5^{th} layer. Our primary analysis uses a **Linear Probe** (Logistic Regression with L_2 regularization). To ensure findings are not artifacts of linearity, we replicated all experiments using non-linear **MLP probes**, yielding qualitatively similar results (see Appendix A.1). All probes are evaluated via Nested Stratified K-Fold Cross-Validation ($k = 10$), reporting AUC on the aggregated out-of-fold probabilities.

3.5 Evaluation Metrics

We evaluate performance using the Area Under the ROC Curve (AUC). AUC is threshold-independent and robust to class imbalance, ensuring we measure genuine separation ability given the varying correctness rates across datasets.

Statistical Significance. We assess the significance of the premium gap using paired t-tests across validation folds. To control for family-wise error rates in multiple comparisons, we apply the Bonferroni-Holm correction (Holm, 1979) ($p < 0.05$).

4 Results

We present our empirical findings in three parts. First, we demonstrate that cross-probes match self-probe performance in 2 out of 3 models in factual tasks and in all models in mathematical reasoning tasks (Section 4.1). Second, we identify high inter-model agreement on correctness labels as a critical confound: when models frequently agree on which questions they answer correctly or incorrectly, external probes can exploit the external model's own correctness patterns to predict the target model's behavior (Section 4.2). Third, by isolating performance on disagreement subsets, we reveal that a statistically significant yet modest premium gap re-emerges in factual tasks but remains absent in mathematical reasoning (Section 4.3).

4.1 The Vanishing Premium Gap

We first evaluate correctness prediction on the standard full test sets. As shown in Figures 2 and 6, self-probes successfully predict correctness across both factual knowledge and mathematical reasoning. However, this performance is not unique to the model's internal states. Embedding model's probes significantly reduce the premium gap in factual tasks, and cross-model probes eliminate it entirely in 2 out of 3 models. In mathematical reasoning, *both* embedding model and cross-model probes match self-probe performance, yielding a non-existent premium gap. This initial finding suggests that correctness prediction does not benefit from access to a model's unique internal states. The observed performance parity aligns with recent work by Xiao et al. (2025), challenging the existence of privileged knowledge regarding a model's own correctness.

4.2 The Agreement Confound

We hypothesize that the absence of a premium gap in the cross-model scenario stems from a fundamental confound: *inter-model agreement*. As shown in

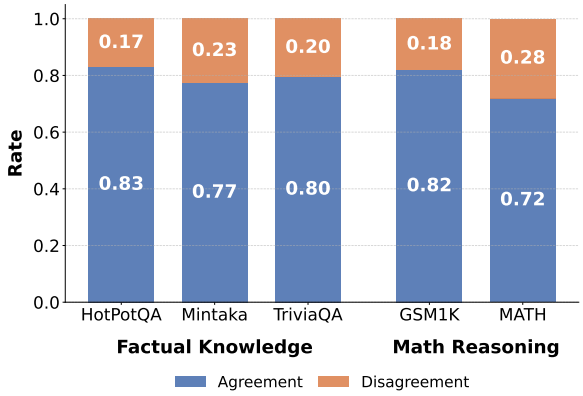

Figure 4: **Agreement vs. Disagreement Rates.** Stacked bar chart showing the proportion of questions on which models agree (blue) or disagree (orange) on correctness across datasets.

Figure 4, models agree on correctness for approximately 80% of questions in factual tasks and 75% in mathematical reasoning. This high agreement rate creates a critical problem for interpreting our results.

When models agree on the majority of examples, the external model's correctness becomes highly correlated with the target's correctness. This means that any signal in the external representation that predicts the external model's success—whether from input question features ($\mathbf{z}^{\text{ext}}_{\text{public}}$) or the external model's own privileged information about correctness ($\mathbf{z}^{\text{ext}}_{\text{private}}$)—will also predict the target's success on most cases. Consequently, an external probe can achieve high predictive performance compared to the self-probe without accessing the target model's privileged information.

This confound makes interpreting our full-set results particularly challenging. Concretely, in our experiments we consistently observe that Gemma representations dominate cross-model prediction: in linear probes, Gemma achieves the best performance as an external representation in 7 out of 9 factual cases (Figure 3), while in MLP probes, it is universally dominant across all 9 cases (Figure 7). However, this dominance is ambiguous and could reflect two very different underlying mechanisms:

1. **No Privileged Knowledge:** Models truly lack internal correctness signals. Gemma simply encodes superior $\mathbf{z}_{\text{public}}$ features regarding question difficulty.

2. **Masked Privileged Knowledge:** Privileged knowledge exists in both models, but Gemma's representation masks the target's signal. Because Gemma provides both robust public features *and* a private signal that serves as a high-fidelity proxy (due to agreement), the superior summed contribution ($\mathbf{z}^{\text{Gemma}}_{\text{public}} \oplus \mathbf{z}^{\text{Gemma}}_{\text{private}}$) dominates the probe's learned weights, effectively obscuring the target's own internal signal.

To distinguish between these scenarios, we evaluate on the *Disagreement Subset*, where the external model's private signal cannot be leveraged to predict target correctness.

4.3 Re-emergence of Domain-Specific Privileged Knowledge

Evaluation on the disagreement subset reveals a sharp contrast between factual and mathematical domains across both linear and MLP probes (Figure 3, Figure 7).

Factual Knowledge: The Gap Re-emerges. In factual tasks, stripping away the agreement confound reveals genuine privileged knowledge. Self-probes consistently outperform external probes, exhibiting a statistically significant premium gap in 8 out of 9 experimental configurations using linear probes (Figure 3) and in all 9 configurations using MLP probes (Figure 7). This indicates that when the external proxy fails, external representations cannot fully account for the target model's correctness. The target model retains unique internal signals that remain inaccessible to observers. Detailed AUC scores for each factual dataset and self-cross model pairing are presented in Figures 8 and 10. While self-probes consistently outperform cross-probes across all factual datasets and model pairs, AUC values on the disagreement subset are substantially lower than on the full test set. This is expected, as inter-model disagreement indicates boundary regions where models exhibit higher uncertainty (Lakshminarayanan et al., 2017), resulting in less stable correctness patterns that are harder to predict.

Mathematical Reasoning: No Evidence for Privileged Knowledge. In sharp contrast, mathematical reasoning tasks show no premium gap. Even on the disagreement subset, external model probes closely match or even outperform self-probes across all targets. Detailed AUC scores for each mathematical reasoning dataset and model pairing are presented in Figures 9 and 11.

5 Mechanisms of Latent Correctness

Our findings in Section 4.3 establish a fundamental difference between domains: factual knowledge retrieval tasks present a consistent premium gap indicating privileged knowledge in disagreement subsets, whereas mathematical reasoning tasks do not. This raises a critical question: *Why does privileged knowledge emerge for factual recall, while no such signal is found for mathematical reasoning?*

5.1 Retrieval vs. Reasoning Mechanisms

We hypothesize that this distinction stems from the nature of the internal computations involved. Recent mechanistic analyses by Chi et al. (2025) demonstrate that LLM internal states primarily encode subject-driven "knowledge recall" patterns rather than objective correctness. Crucially, their work shows that when models hallucinate on factual queries, they employ the same internal recall process as for correct responses, producing indistinguishable hidden-state geometries. This suggests that LLMs encode *whether retrieval occurred*, not whether the retrieved information is correct. Building on this insight, we propose that factual correctness depends on the presence of model-specific memory traces—internal representations that vary idiosyncratically across models based on their specific knowledge. In contrast, mathematical reasoning relies on multi-step computations over universal logical structures that do not depend on memorized associations, making correctness signals universally observable through computational difficulty patterns. This observation aligns with prior research emphasizing the distinction between memorization and reasoning as separate computational modes (Hong et al., 2025)

5.2 Experimental Design

To test this hypothesis, we introduce two control experiments: (1) **External-Labels**, which measures whether difficulty is universal by directly aggregating peer model correctness labels without training, and (2) **Lexical-Only**, which isolates entity-driven signals by training probes on bag-of-words representations containing only named entities and nouns, stripped of all syntax and function words. To test whether factual correctness relies on model-specific memory traces while mathematical correctness reflects universal difficulty patterns, we introduce two control experiments that isolate different sources of predictive signal.

5.3 Control Experiments

External-Labels. We train a logistic regression probe using *only* the binary correctness labels of external peer models as features, aggregating the labels of the two peer models by their mean. This tests whether question difficulty is universal: high performance would indicate that questions that challenge peer LLMs tend to challenge the target model as well, even without access to its internal states or question representations.

Lexical-Only. To test whether correctness signals are entity-driven, we retain only a bag-of-words representation of named entities and nouns, discarding all syntax and function words. We train probes on the target model's own hidden states when processing these *Lexical-Only* inputs and compare against the full *Original Question* baseline. This reveals whether the model's internal representations of specific tokens—entities in factual tasks or mathematical concepts in reasoning tasks—carry the correctness signal independently of syntactic or contextual processing.

Full implementation details of the analysis can be found in Appendix E.

5.4 Decomposing Correctness Signals

Results are presented in Figure 5 (Qwen), with consistent patterns replicated for Llama and Gemma in Figures 13 and 14.

Universal Difficulty Dominates. The *External-Labels* predictor—derived solely from peer correctness without training—achieves remarkably high AUC across all datasets, surpassing probes trained on the target's own hidden states. This demonstrates that peer labels alone, without any question representation, provide stronger signals than internal states. This reinforces that disagreement subsets are essential for isolating privileged knowledge.

Factual Correctness is Entity-Driven. In factual domains (Mintaka, TriviaQA, HotpotQA), lexical features recover 53.7%, 75.0%, and 73.5% correspondingly of original probe performance relative to the random baseline (0.5 AUC). That high predictive success is achieved using only input named entities and nouns suggests that much of the correctness signal exploited by the probe comes from concept familiarity, supporting the findings of Chi et al. (2025). The entities themselves serve as proxies for whether the model possesses a memory

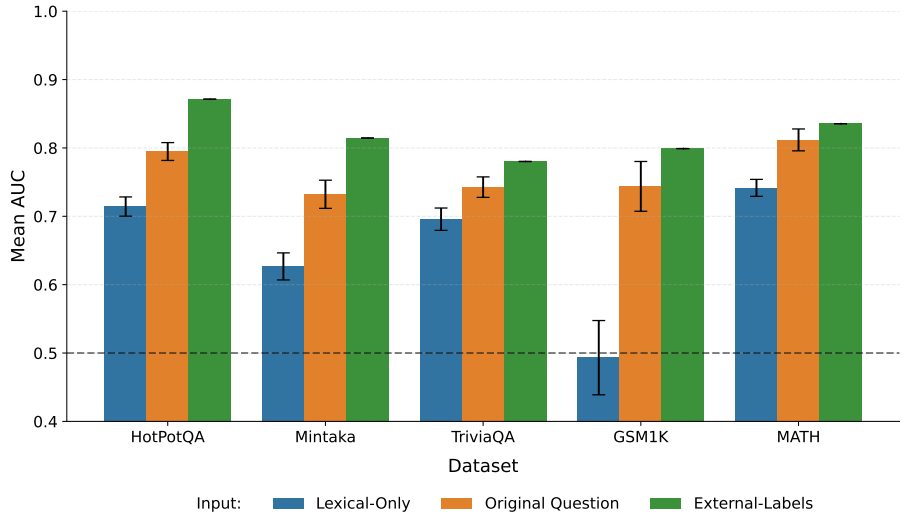

Figure 5: **Control Experiments (`Qwen-2.5-7B`).** Mean AUC for correctness prediction, averaged over layers, of probes trained on the *Original Question* compared to two controls. *External-Labels* uses the direct average of peer-model correctness as a predictor (no training), while *Lexical-Only* uses a trained probe on a bag-of-words representation of entities and nouns. Strong External-Labels performance reflects high inter-model agreement, while Lexical-Only results suggest that much of the signal in factual tasks arises from entity familiarity.

trace. Critically, the presence or absence of such traces are unique to each model. One model may have consolidated a memory for "Édith Piaf" while another has not. This model-specific retrieval capability may be the source of privileged knowledge in factual tasks.

Math: Topic Signals Without Privilege. In mathematical reasoning, we observe contrasting behavior. For GSM1K, the Lexical-Only probe fails completely (AUC $\approx$ 0.49). When we examine the stripped input, this is intuitive: entities like "savings account" or "$50" carry no information about reasoning complexity. However, MATH presents an unexpected result: lexical probes perform surprisingly well, recovering 75.6% of performance—comparable to factual tasks. Lexical tokens in MATH encode mathematical *topics* (e.g., "eigenvalue", "asymptote"). We speculate these topic signals operate at a coarser granularity than entity-specific memory traces: while factual recall depends on model-specific memories of input entities, mathematical topics reflect broader difficulty patterns universally observable across models. This high-level structure may explain the absence of privileged knowledge in mathematical reasoning, though further investigation is needed.

6 Discussion

We investigate whether LLMs possess privileged knowledge about the correctness of their forthcoming answer by comparing probes trained on self-representations versus external model representations. Our key methodological contribution is identifying inter-model agreement as a critical confound: when models share correctness patterns, external probes exploit peer correctness as a proxy, masking genuine privileged signals. On disagreement subsets, self-probes consistently outperform external probes in factual tasks, suggesting models possess unique signals about correctness stemming from model-specific memory traces. In contrast, mathematical reasoning tasks show no evidence for privileged knowledge, suggesting that correctness is determined by universal difficulty patterns accessible to any capable model. These findings challenge the view that LLMs lack privileged knowledge about their output correctness using question-only features. Models demonstrably encode private signals in retrieval-based domains, while reasoning tasks exhibit publicly observable correctness patterns. Our disagreement-based methodology can be extended to study privileged knowledge in hybrid domains (coding, commonsense reasoning) and other forms of model introspection beyond correctness prediction. Furthermore, from a practical perspective, investigating whether the correctness signal can be leveraged to improve model accuracy through methods such as activation steering presents an important avenue for future work.

7 Limitations

Our analysis has several limitations: (1) We limit our evaluation to models with 7B–9B parameters–larger models may display different patterns of privileged knowledge; (2) our scope is limited to factual knowledge and mathematical reasoning, while hybrid domains such as coding and commonsense reasoning remain outside the scope of this study; (3) we rely on linear and MLP probes which, although standard in prior work, may have limited capacity to fully extract privileged signals; and (4) our study reveals systematic patterns linking representational structure to correctness; complementary intervention experiments could further establish the causal mechanisms by which memory traces contribute to factual privileged knowledge.

References

William Alston. 1971. Varieties of privileged access. *American Philosophical Quarterly*, 8(3):223–241.

Felix Jedidja Binder, James Chua, Tomek Korbak, Henry Sleight, John Hughes, Robert Long, Ethan Perez, Miles Turpin, and Owain Evans. 2025. Looking inward: Language models can learn about themselves by introspection. In *The Thirteenth International Conference on Learning Representations*.

Iván Vicente Moreno Cencerrado, Arnau Padrés Masdemont, Anton Gonzalvez Hawthorne, David Demitri Africa, and Lorenzo Pacchiardi. 2025. No answer needed: Predicting LLM answer accuracy from question-only linear probes. *CoRR*, abs/2509.10625.

Sirui Chen, Shu Yu, Shengjie Zhao, and Chaochao Lu. 2025. From imitation to introspection: Probing self-consciousness in language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7553–7583, Vienna, Austria. Association for Computational Linguistics.

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

Iulia M Comsa and Murray Shanahan. 2025. Does it make sense to speak of introspection in large language models? *arXiv preprint arXiv:2506.05068*.

Javier Ferrando, Oscar Balcells Obeso, Senthooran Rajamanoharan, and Neel Nanda. 2024. Do i know this entity? knowledge awareness and hallucinations in language models. In *The Thirteenth International Conference on Learning Representations*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2023. A framework for few-shot language model evaluation.

Brie Gertler. 2010. *Self-knowledge*. Routledge.

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. *Preprint*, arXiv:2407.21783.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Yihuai Hong, Dian Zhou, Meng Cao, Lei Yu, and Zhijing Jin. 2025. The reasoning-memorization interplay in language models is mediated by a single direction. *Preprint*, arXiv:2503.23084.

Li Ji-An, Marcelo G Mattar, Hua-Dong Xiong, Marcus K Benna, and Robert C Wilson. 2025. Language models are capable of metacognitive monitoring and control of their internal activations. *ArXiv*, pages arXiv–2505.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *CoRR*.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Belinda Z Li, Zifan Carl Guo, Vincent Huang, Jacob Steinhardt, and Jacob Andreas. 2025a. Training language models to explain their own computations. *arXiv e-prints*, pages arXiv–2511.

Millicent Li, Alberto Mario Ceballos Arroyo, Giordano Rogers, Naomi Saphra, and Byron C Wallace. 2025b.

Do natural language descriptions of model activations convey privileged information? *arXiv preprint arXiv:2509.13316*.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2025. Llms know more than they show: On the intrinsic representation of llm hallucinations. In *The Thirteenth International Conference on Learning Representations*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. Pytorch: an imperative style, high-performance deep learning library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 8026–8037.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Priyanka Sen, Alham Fikri Aji, and Amir Saffari. 2022. Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1604–1619.

Yeongbin Seo, Dongha Lee, and Jinyoung Yeo. 2025. Quantifying self-awareness of knowledge in large language models. *CoRR*, abs/2509.15339.

Hovhannes Tamoyan, Subhabrata Dutta, and Iryna Gurevych. 2025. Factual self-awareness in language models: Representation, robustness, and scaling. *arXiv e-prints*, pages arXiv–2505.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Hanqi Xiao, Vaidehi Patil, Hyunji Lee, Elias Stengel-Eskin, and Mohit Bansal. 2025. Generalized correctness models: Learning calibrated and model-agnostic correctness predictors from historical patterns. *arXiv preprint arXiv:2509.24988*.

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. *Preprint*, arXiv:2505.09388.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2369–2380.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8653–8665. Association for Computational Linguistics.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. 2025. Reasoning models know when they're right: Probing hidden states for self-verification. *ArXiv*, abs/2504.05419.

Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Charlotte Zhuang, Dylan Slack, and 1 others. 2024. A careful examination of large language model performance on grade school arithmetic. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 46819–46836.

A Additional Experimental Results

A.1 MLP Probe Results

To ensure that the vanishing premium gap is not an artifact of the limited expressivity of linear classifiers, we replicated our primary analysis using non-linear Multi-Layer Perceptron (MLP) probes (implementation details in Appendix C).

The results, visualized in Figures 7 and 12, align closely with the linear probe findings, demonstrating that our conclusions are robust to probe architecture:

Full Test Set. On the full test sets, the premium gap diminishes or vanishes, particularly in mathematical reasoning. Similar to the linear setting, we observe strong external dominance. Notably, Gemma representations are even more dominant in the non-linear setting, achieving the best cross-model performance in all 9 experimental configurations (Figure 7). This reinforces the hypothesis that external representations often capture public

correctness features more effectively than the target's own features.

Disagreement Subset. Crucially, the re-emergence of privileged knowledge in factual tasks is even more consistent under non-linear probing. While linear probes detected a significant premium gap in 8 out of 9 factual configurations, MLP probes detect a significant gap in all 9 out of 9 configurations. Conversely, in mathematical reasoning (GSM1K, MATH), the premium gap remains absent, further validating that mathematical correctness signals are publicly accessible even to non-linear observers.

B Dataset and Disagreement Statistics

We utilize five datasets with varying total sample sizes (Total). A critical component of our methodology involves analyzing the *disagreement subset*—instances where the source and target models predict different correctness labels ($y_{\text{ext}} \neq y_{\text{target}}$).

Since the size of this subset varies depending on the specific pair of models being compared, we report the exact counts ($N_{\text{disagreement}}$) for each unique model pair in Table 1. Across all configurations, the disagreement subsets retain sufficient scale ($\approx 20\%$ of the original data).

		Subset Size ($N_{\text{disagreement}}$)		
Dataset	**Total**	**G ↔ L**	**G ↔ Q**	**L ↔ Q**
Mathematical Reasoning				
GSM1K	1k	186	142	216
MATH	10k	2,932	2,967	2,519
Factual Knowledge				
HotpotQA	10k	1,592	1,730	1,802
Mintaka	4k	805	973	946
TriviaQA	10k	1,588	2,238	2,320

Table 1: Dataset statistics and disagreement subset sizes. Total denotes the full test set size. The subsequent columns show the size of the disagreement subset for each unique model pair. (G=Gemma-2-9B, L=Llama-3.1-8B, Q=Qwen2.5-7B).

C Implementation Details

C.1 Probe Training and Hyperparameters

All probes were trained using a Stratified K-Fold Cross-Validation scheme with $k = 10$ outer folds to estimate generalization performance. AUC metric is computed from pooled out-of-fold (OOF) predictions across these splits.

Linear Probe: We used Logistic Regression with L_2 regularization and the `liblinear` solver. Hyperparameters were selected using an inner 3-fold cross-validation on the training split of each outer fold, tuning the regularization strength $C \in \{0.01, 0.1\}$. The model was trained with balanced class weights, standardized inputs, and a maximum of 500 iterations.

MLP Probe: We used a Multi-Layer Perceptron (MLP) classifier with a single hidden layer of size $(100,)$ and ReLU activation. Hyperparameters were fixed across folds (i.e., no inner cross-validation), with an L_2 penalty of $\alpha = 0.1$. The model was trained with early stopping enabled (using a 10% validation split), standardized inputs, and a maximum of 500 training iterations.

C.2 Significance Testing

To assess statistical uncertainty, we computed 95% confidence intervals (CIs) using bootstrap resampling over the pooled out-of-fold (OOF) predictions. Specifically, we resampled the OOF predictions with replacement for 1000 iterations and computed the empirical 2.5 and 97.5 percentiles of the resulting AUC distribution.

D Dataset Generation and Evaluation Details

We standardized our generation and evaluation protocols using official model pipelines, aligning our methodology with the Language Model Evaluation Harness (Gao et al., 2023) to ensure reproducibility.

D.1 Response Generation

Models were loaded using standard Hugging Face integration. We utilized greedy decoding (`do_sample=False`) across all experiments to ensure deterministic outputs. Input prompts followed standard dataset-specific templates. Generation lengths were strictly controlled based on the domain: we set `max_new_tokens=32` for factual knowledge tasks to enforce concise entity generation, and `max_new_tokens=2048` for mathematical reasoning to accommodate full Chain-of-Thought derivations.

D.2 Correctness Evaluation

Factual Knowledge. For Mintaka, TriviaQA, and HotpotQA, we evaluated correctness using standard Exact Match criteria. A response was labeled correct if any valid alias from the ground

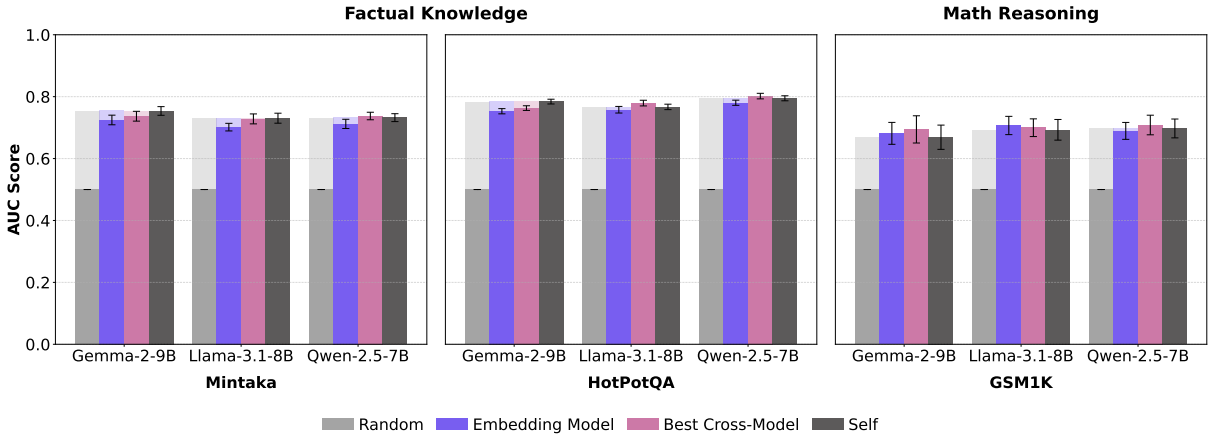

Figure 6: **Premium Gap (Remaining Datasets).** Mean AUC for correctness prediction, averaged over layers, on two task types: factual knowledge (TriviaQA) and mathematical reasoning (MATH). Bars compare Random, Embedding, and Best Cross-Model baselines to the Self-Probe (*Self*) across three target models. Semi-transparent overlays indicate the performance gain (or lack thereof) of *Self* relative to each baseline. Error bars denote 95% confidence intervals.

truth appeared as a case-insensitive substring within the generated text.

Mathematical Reasoning. For MATH and GSM1K, correctness was evaluated using the official evaluation scripts provided with each dataset. These scripts perform robust answer extraction by parsing the generated text to identify the final answer (e.g., via LaTeX `\boxed{}` markers when present) and verify symbolic equivalence with the ground-truth solution, accounting for algebraic and notational variations.

E Implementation Details for Analysis Control Experiments

Unless otherwise stated, all experimental configurations—including model architectures, layer selection (final token of every 5^{th} layer), cross-validation strategy ($k = 10$), and classifier hyperparameters—follow the methodology described in Section 3.4.

E.1 Lexical-Only Experiment

This experiment tests whether predictive performance stems from deep reasoning or simple concept familiarity. We construct a "bag-of-words" representation to isolate lexical features.

Concept Extraction. We extract concepts from the original question using a two-stage pipeline:

1. **Named Entities:** We use `GLiNER` (specifically `urchade/gliner_medium-v2.1`) to identify entities across 20 label types (e.g., person, organization, date, quantity).
2. **Noun Chunks:** We use `spaCy` (`en_core_web_sm`) to capture remaining non-entity concepts.

The union of these extractions is deduplicated, removing substrings and stopwords to form the final concept set.

Input Construction. We format the extracted concepts into a natural language description using the template:

> "This text discusses [Concept A], [Concept B], and [Concept C]."

This synthesized text replaces the original question as the input to the model.

Probe Training. We extract hidden states from the target model while it processes this *concept description*. A Logistic Regression classifier is then trained on these states to predict the model's correctness on the *original* question.

E.2 External-Labels (Peer Labels) Experiment

This experiment evaluates the predictive power of objective difficulty without accessing internal states.

Peer Label Score. For a target model M and question q, we define the "Peer Label" score as

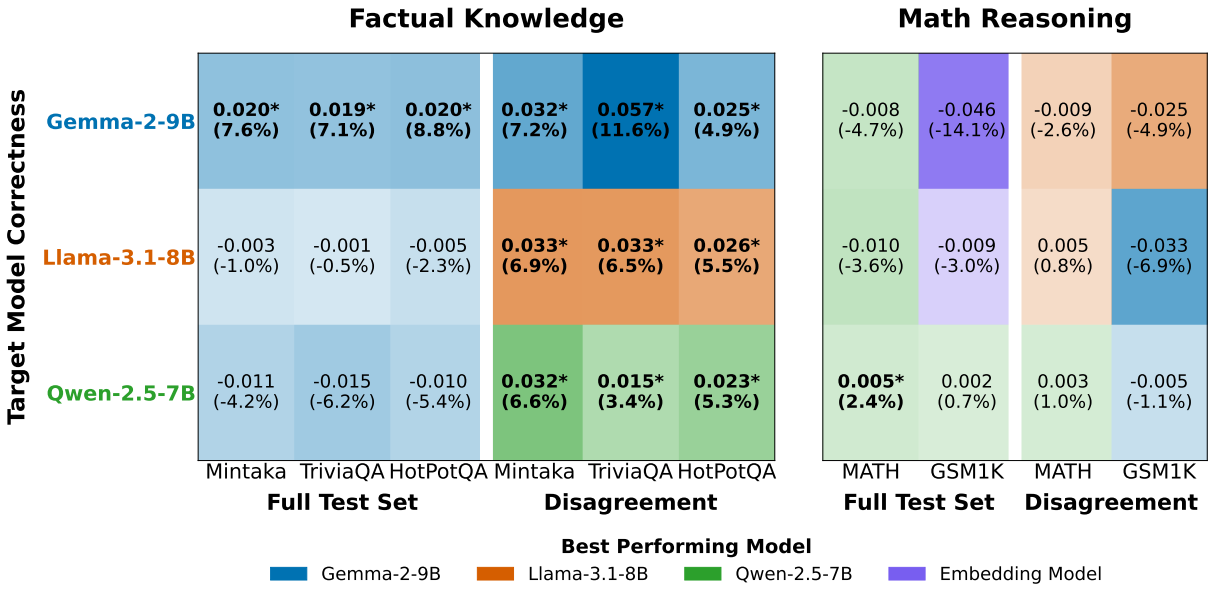

Figure 7: **Target Model Correctness Prediction (MLP Probes).** Heatmap of correctness prediction differences across target models, datasets, and test subsets. Each cell reports the AUC difference ($\Delta\text{AUC} = \text{Self} - \text{Best External}$), with the percentage of the gap closed shown in parentheses, computed as $\frac{\text{Self}-\text{Best External}}{1-\text{Best External}} \times 100$. The y-axis lists target models, and cell colors indicate the best-performing external source model for each setting. Asterisks (*) denote statistically significant differences (paired t-test, $p < 0.05$, Bonferroni–Holm correction).

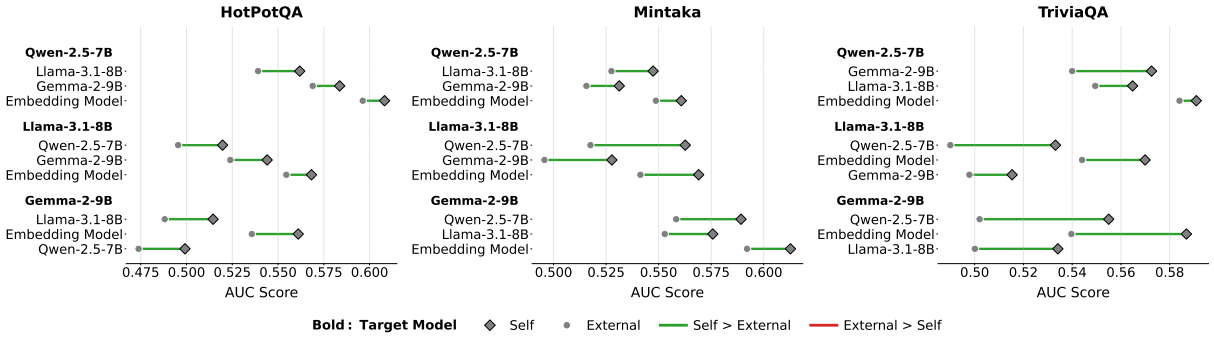

Figure 8: **Disagreement Gap: Factual Knowledge Breakdown (Linear Probes).** Detailed performance on the disagreement subset across Mintaka, TriviaQA, and HotPotQA. Self-probes consistently outperform external probes across all factual datasets, reinforcing the existence of idiosyncratic memory traces (privileged knowledge).

the mean correctness of all other available models (peers) in our evaluation suite:

$$\text{Score}(q) = \frac{1}{|\text{Peers}|} \sum_{P \in \text{Peers}} \mathbb{I}(P \text{ is correct on } q) \tag{3}$$

Evaluation. Unlike the main experiments, no training is involved. We use the calculated Score(q) directly as a soft probability label to predict the target model's correctness. The AUC is calculated by thresholding this score.

F Hardware Details

All experiments were conducted on a system with 32 Intel(R) Xeon(R) Gold 6430 CPUs and 1.0 TB of RAM. The system was equipped with three NVIDIA RTX 6000 Ada Generation GPUs, each with 49 GB of VRAM.

G Licenses and Third-Party Usage

This work is implemented using **PyTorch** (Paszke et al., 2019), an open-source deep learning framework licensed under the BSD license, and the **Hugging Face Transformers** library (Wolf et al., 2019), licensed under Apache 2.0. We also employ **spaCy** (MIT License) and **GLiNER** (Apache 2.0) for the lexical analysis described in the control experiments. All software usage complies with their respective license terms.

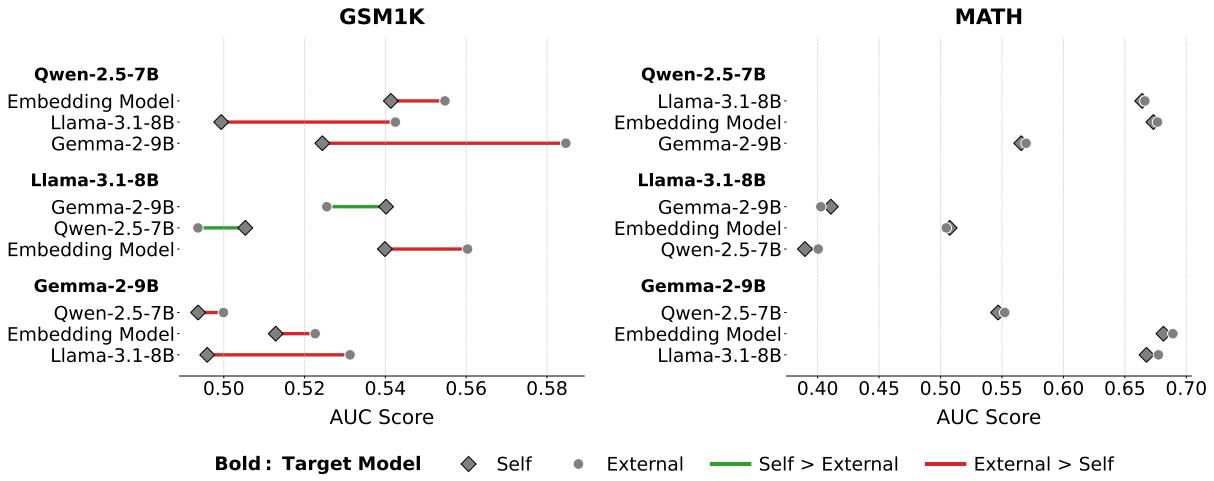

Figure 9: **Disagreement Gap: Mathematical Reasoning Breakdown (Linear Probes).** Detailed performance on the disagreement subset across GSM1K and MATH. Unlike factual tasks, mathematical correctness shows no consistent premium gap, indicating that reasoning difficulty is a public feature accessible to external models.

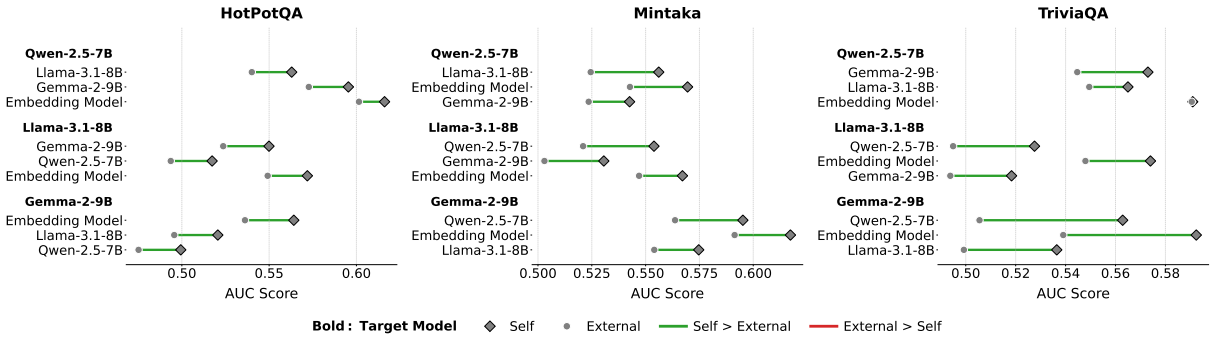

Figure 10: **Disagreement Gap: Factual Knowledge Breakdown (MLP Probes).** Detailed disagreement subset performance using MLP probes. The premium gap is even more pronounced with non-linear probes, with Self-representations outperforming Best External probes in 9 out of 9 configurations.

Datasets. We utilize several open-source datasets for evaluation:

- **Mintaka** (Sen et al., 2022) is licensed under CC-BY 4.0.
- **HotpotQA** (Yang et al., 2018) is licensed under CC-BY-SA 4.0.
- **TriviaQA** (Joshi et al., 2017) is licensed under Apache 2.0.
- **GSM1K** (Zhang et al., 2024) and **MATH** (Hendrycks et al., 2021) are licensed under the MIT License.

H Use of AI Assistants

We utilized AI assistants for refining text clarity and coding assistance. All scientific claims, experimental results, and final text were written by the authors.

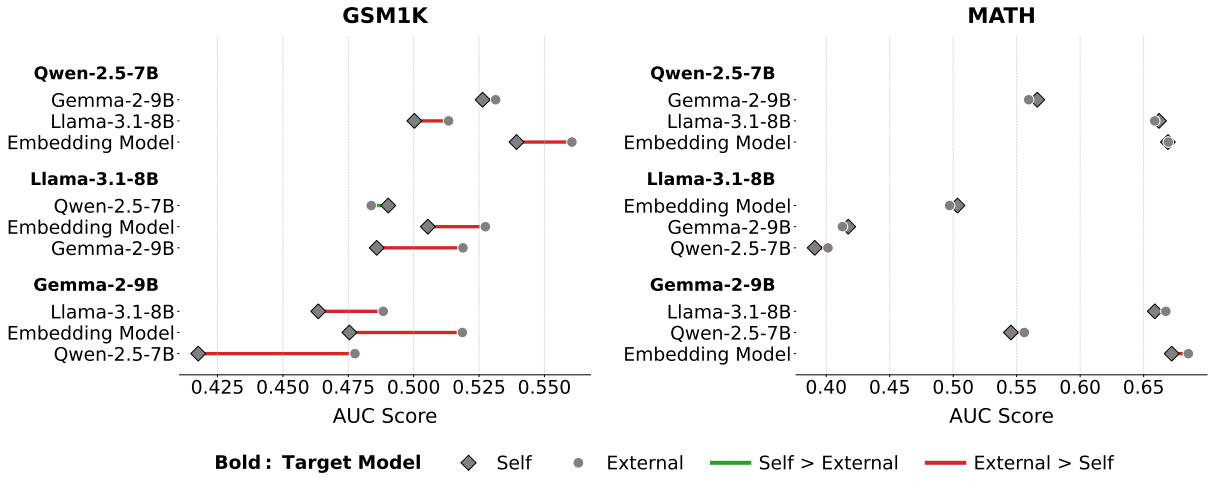

Figure 11: **Disagreement Gap: Mathematical Reasoning Breakdown (MLP Probes).** Detailed disagreement subset performance using MLP probes across GSM1K and MATH. Consistent with linear results, increased probe expressivity does not uncover hidden privileged info in math tasks; external models remain effective predictors.

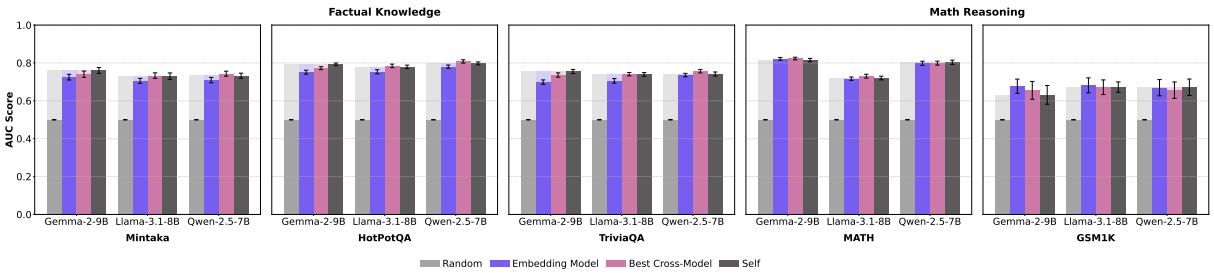

Figure 12: **Premium Gap (MLP Probes).** Mean AUC for correctness prediction, averaged over layers, on two task types: factual knowledge (TriviaQA) and mathematical reasoning (MATH). Bars compare Random, Embedding, and Best Cross-Model baselines to the Self-Probe (*Self*) across three target models. Semi-transparent overlays indicate the performance gain (or lack thereof) of *Self* relative to each baseline. Error bars denote 95% confidence intervals.

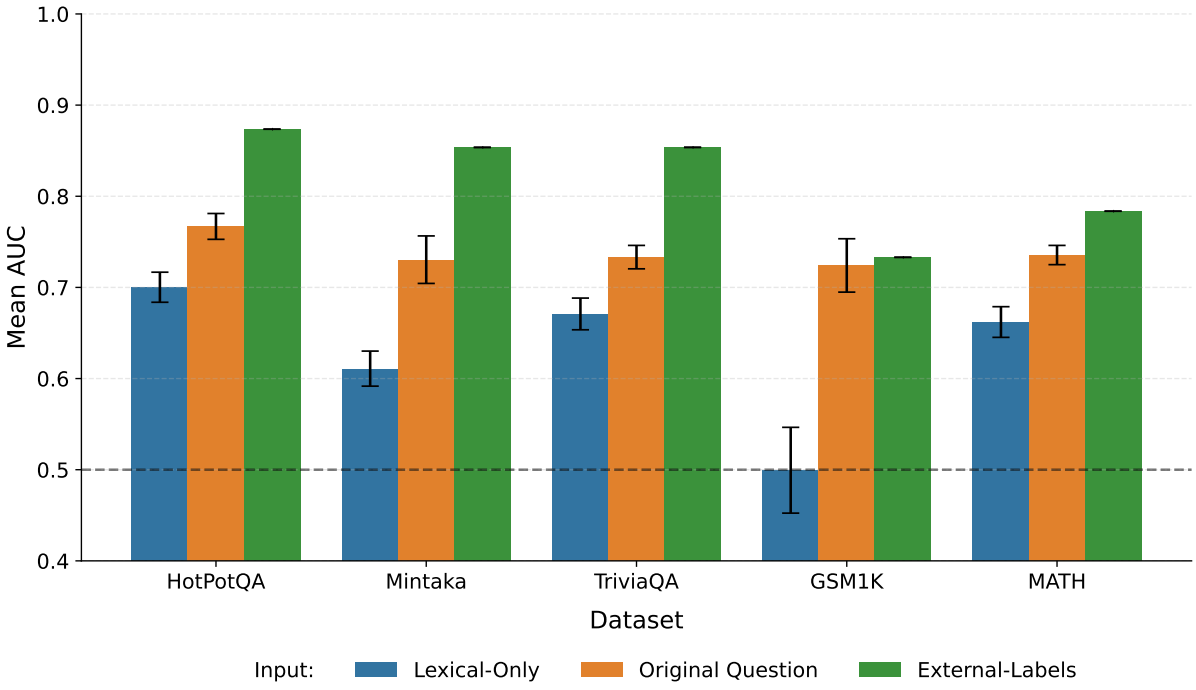

Figure 13: **Control Experiments (Llama-3.1-8B).** AUC of probes trained on the *Original Question* compared to two controls. *External-Labels* uses the direct average of peer-model correctness as a predictor (no training), while *Lexical-Only* uses a trained probe on a bag-of-words representation of entities and nouns. Strong External-Labels performance reflects high inter-model agreement, while Lexical-Only results suggest that much of the signal in factual tasks arises from entity familiarity.

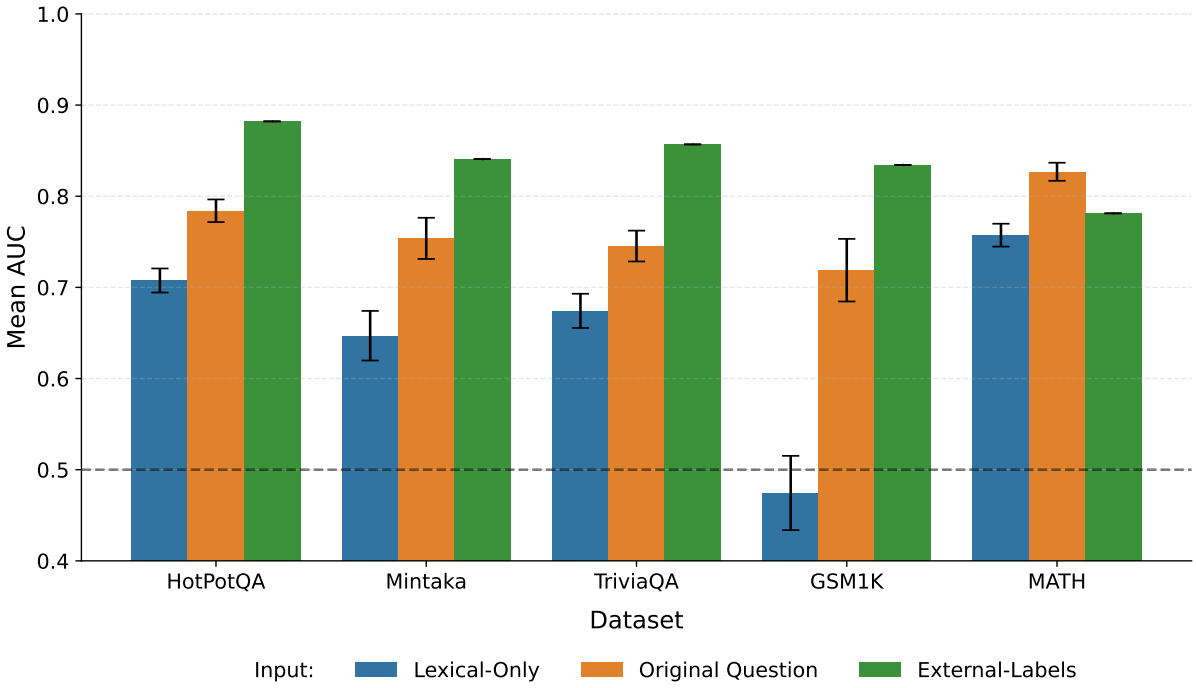

Figure 14: **Control Experiments (Gemma-2-9B).** AUC of probes trained on the *Original Question* compared to two controls. *External-Labels* uses the direct average of peer-model correctness as a predictor (no training), while *Lexical-Only* uses a trained probe on a bag-of-words representation of entities and nouns. Strong External-Labels performance reflects high inter-model agreement, while Lexical-Only results suggest that much of the signal in factual tasks arises from entity familiarity.