Masked by Consensus: Disentangling Privileged Knowledge in LLM Correctness

Tomer Ashuach[1] **Liat Ein-Dor**[2] **Shai Gretz**[2] **Yoav Katz**[2] **Yonatan Belinkov**[1,3]

[1]Technion – Israel Institute of Technology [2]IBM Research
[3]Kempner Institute, Harvard University

tomerashuach@campus.technion.ac.il, belinkov@technion.ac.il
{liate,avishaig,katz}@il.ibm.com

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* about answer correctness, information unavailable through external observation. 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. On standard evaluation, we find no advantage: self-probes perform comparably to peer-model probes. We hypothesize this is due to high inter-model agreement of answer correctness. To isolate genuine privileged knowledge, we evaluate on *disagreement subsets*, where models produce conflicting predictions. Here, we discover domain-specific privileged knowledge: self-representations consistently outperform peer representations in factual knowledge tasks, but show no advantage in math reasoning. We further localize this domain asymmetry across model layers, finding that the factual advantage emerges progressively from early-to-mid layers onward, consistent with model-specific memory retrieval, while math reasoning shows no consistent advantage at any depth.

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 large language models (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, do they possess *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.

To address this challenge, we construct *disagreement subsets*: questions where models produce

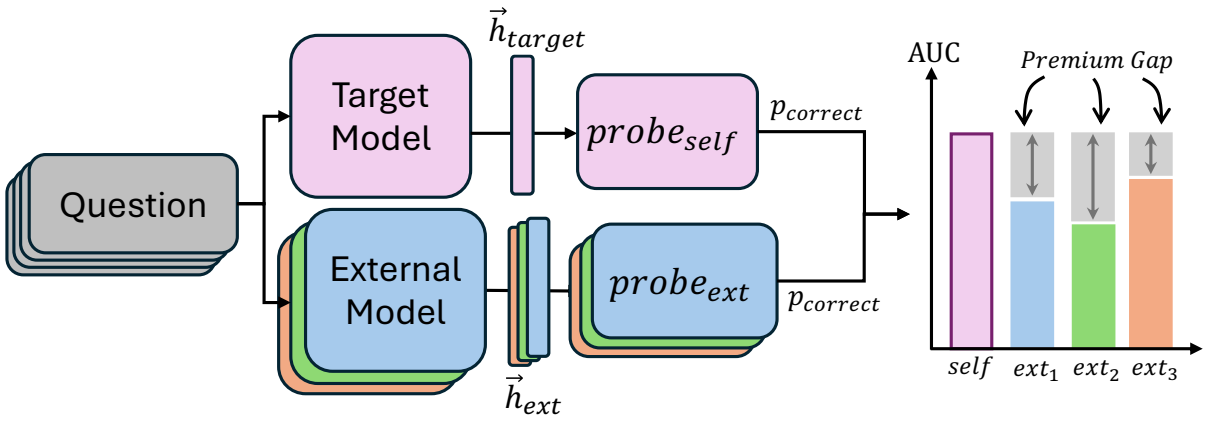


Figure 1: **Overview of the experimental framework.** Questions are input to a target model and to external models, yielding representations $\mathbf{h}_{\text{target}}$ and $\mathbf{h}_{\text{ext}}$. Probes trained on these representations predict answer correctness. We evaluate probe performance using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

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. When a gap does appear for a particular model, that same model also excels at predicting peer correctness—suggesting its stronger representations reflect a general advantage rather than privileged self-knowledge. However, when evaluated on disagreement subsets, a statistically significant premium gap emerges in factual knowledge domains ($\sim$5%) across all models, providing genuine evidence for privileged knowledge. In contrast, mathematical reasoning shows no such advantage: probes trained on an external model's representations of the same question perform comparably to those trained on the model's own internal representations, even under disagreement (Figure 3). We further localize this domain asymmetry across model layers, finding that the factual advantage emerges progressively from early-to-mid layers onward and strengthens with depth, consistent with model-specific memory retrieval that accumulates through the forward pass, while mathematical reasoning shows no consistent advantage across layers.

We summarize our main findings as follows:

- We systematically evaluate correctness prediction across five datasets and three models, 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.
- We localize this domain asymmetry across network depth, showing that the factual advantage emerges progressively from early-to-mid layers onward, while mathematical reasoning shows no consistent advantage at any depth.

2 Related Work

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 self-explanation. Similarly, Binder et al. (2025) define introspection as knowledge derived from in-

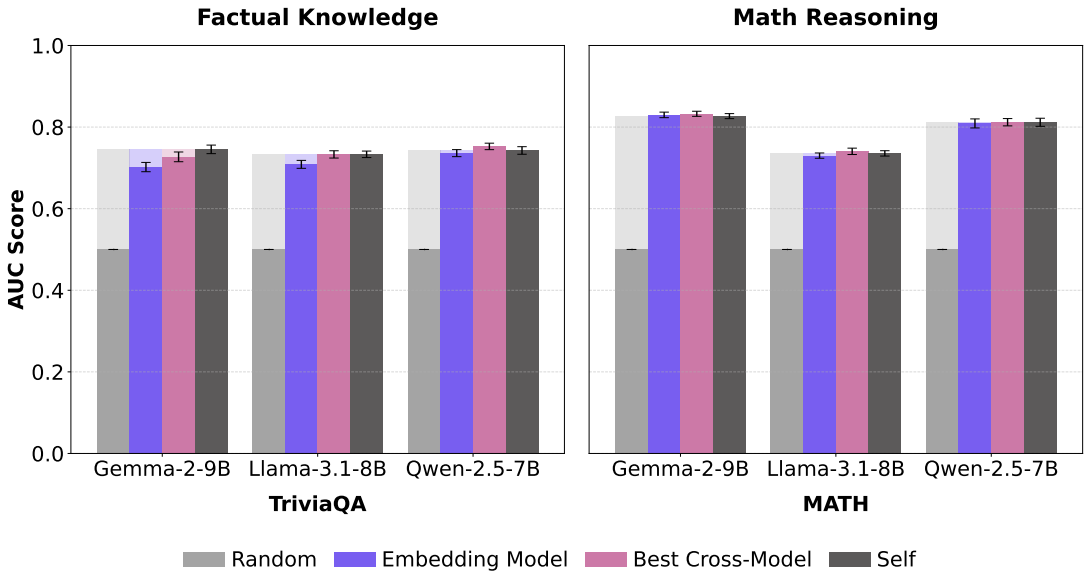


Figure 2: **Premium Gap.** Mean AUC for correctness prediction, averaged over layers, on two task types: factual knowledge (TriviaQA) and mathematical reasoning (MATH). Bars compare Random, Embedding, and Best Cross-Model baselines to the Self-Probe (*Self*) across three target models. Semi-transparent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

ternal 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. LLMs have been shown to possess a degree of self-evaluation ability—accurately estimating answer correctness on multiple-choice tasks (Kadavath et al., 2022) and identifying unanswerable questions (Yin et al., 2023)—motivating direct investigation of their hidden states for correctness signals. 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 encode the correct answer internally 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. 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.

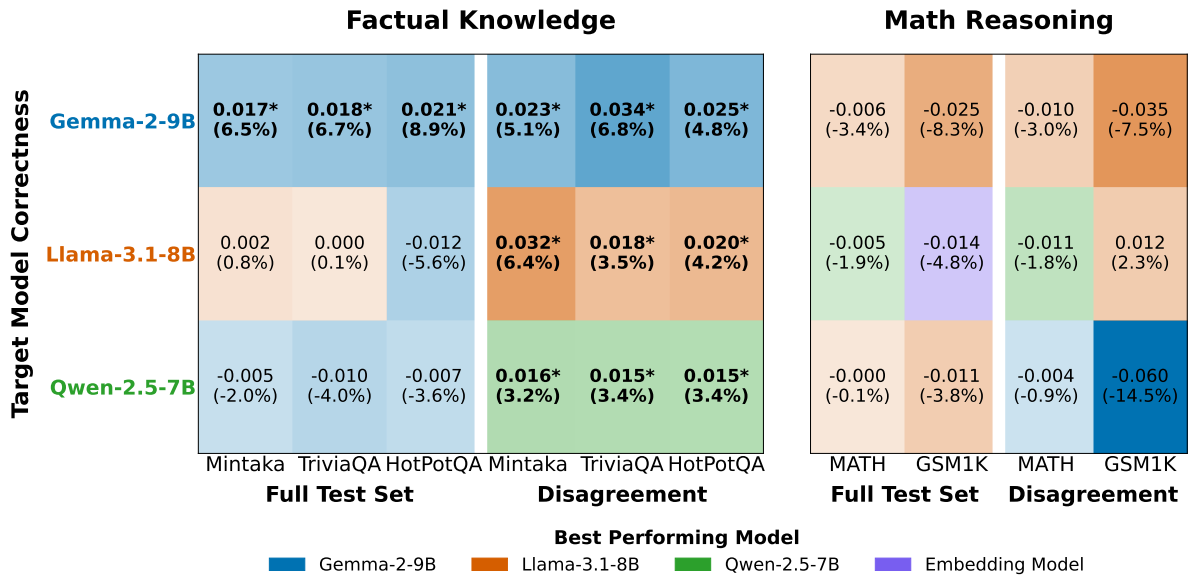


Figure 3: **Target Model Correctness Prediction.** Heatmap of correctness prediction differences across target models, datasets, and test subsets. Each cell reports the AUC difference ($\Delta\text{AUC} = \text{Self} - \text{Best External}$), with the percentage of the gap closed shown in parentheses, 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 externally 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 an 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. On random samples, high inter-model agreement allows external probes to exploit peer correctness patterns as a proxy for the target model's behavior, masking any privileged knowledge signal. To eliminate this confound, 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, allowing the probe to trivially exploit the external model's inverted correctness signal. Instead, we train probes on the full training dataset to learn the full correctness pattern of the source model, and filter predictions during inference to strictly evaluate on the disagreement test subset.

3.4 Experimental Setup

Models. We evaluate three instruction-tuned decoder LMs of comparable size: `Llama-3.1-8B` (Grattafiori et al., 2024), `Qwen2.5-7B` (Yang et al., 2025c), and `Gemma-2-9B` (Riviere et al., 2024), alongside the embedding model `Qwen3-Embedding-8B` (Yang et al., 2025a). The three decoder LMs serve as both target and source models, while the embedding model is used only as a source. To assess scalability, we additionally evaluate `Qwen-3-32B` (Yang et al., 2025b) as both a target model and an external probe candidate; results are reported in Appendix A.1.

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 E 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.2). 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 external-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 emerges in factual tasks but remains absent in mathematical reasoning (Section 4.3). We additionally evaluate on a larger model (`Qwen-3-32B`), verifying that the same overall trends hold despite its richer representations strengthening the external probe baseline (Appendix A.1).

4.1 Full Test Sets Reveal No Premium Gap

We first evaluate correctness prediction on the standard full test sets. As shown in Figure 2 (remaining datasets in Figure 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. In factual tasks, self-probes show only a small advantage over embedding model probes, and are comparable to cross-model probes in 2 out of 3

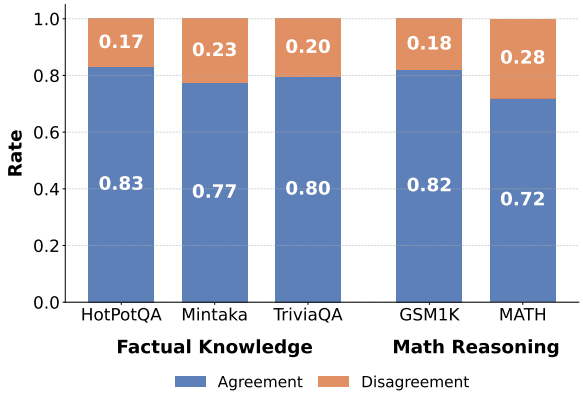


Figure 4: **Agreement vs. Disagreement Rates.** Stacked bar chart showing the proportion of questions on which models agree (blue) or disagree (orange) on correctness across datasets, averaged over all model pairs.

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, 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 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 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 across all 9 experimental configurations using both linear and MLP probes (Figures 3 and 7, Left). 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 (Appendix C). While self-probes consistently outperform external-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. We discuss and argue against a potential alternative explanation based on distributional shifts in Appendix E.3.

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 outperform self-probes across all targets (Figures 3 and 7, Right). Detailed AUC scores for each mathematical reasoning dataset and model pairing are presented in Figures 9 and 11 (Appendix C).

5 Where Does Privileged Knowledge Emerge?

Our findings in Section 4.3 establish that privileged knowledge emerges in factual tasks but remains absent in mathematical reasoning, based on layer-averaged premium gaps. A natural follow-up question is *where* in the network this signal originates: is the premium gap uniformly distributed across layers, or does it emerge at specific depths?

To investigate this, we compute the premium gap (self-probe AUC − best external-probe AUC) at each individually probed layer on the disagreement subset, rather than averaging across layers as in Section 4.3. We probe every 5^{th} layer plus the final layer, yielding 6–7 measurement points per model, and plot the premium gap against normalized layer depth (0 = first probed layer, 1 = last). Per-model bar figures showing absolute self and external AUC values at each layer are provided in Appendix F.

5.1 Factual Tasks: Progressive Emergence

Figure 5a presents the per-layer premium gap for factual datasets. Across all three models and datasets, a consistent pattern emerges: the premium gap is near zero or slightly negative in early layers and grows progressively toward deeper layers. Early layers, which primarily encode surface-level and syntactic features (Tenney et al., 2019; Jawahar et al., 2019), show no self-probe advantage, consistent with these representations encoding primarily public information ($\mathbf{z}_{\text{public}}$). The premium gap becomes reliably positive from approximately layer 10–15 onward (normalized depth ∼0.25–0.40), suggesting that privileged knowledge is encoded in deeper representations where models consolidate factual knowledge (Orgad et al., 2025). This pattern is consistent with the view that the privileged signal reflects idiosyncratic memory retrieval states that build up through the forward pass. In particular, Chi et al. (2025) show that knowledge recall in LLMs is dominated by a mid-layer information-flow signal from subject to answer tokens, aligning with our finding that the privileged advantage first appears in early-to-mid layers and strengthens toward later layers.

5.2 Mathematical Reasoning: No Consistent Advantage

The per-layer analysis for mathematical reasoning (Figure 5b) confirms that the absence of privileged knowledge in math is not an artifact of layer averaging. For MATH, the premium gap fluctuates near zero across all layers and models, with no systematic trend. For GSM1K, the gap is predominantly *negative*—external probes outperform self-probes at most depths, inverting the factual pattern.

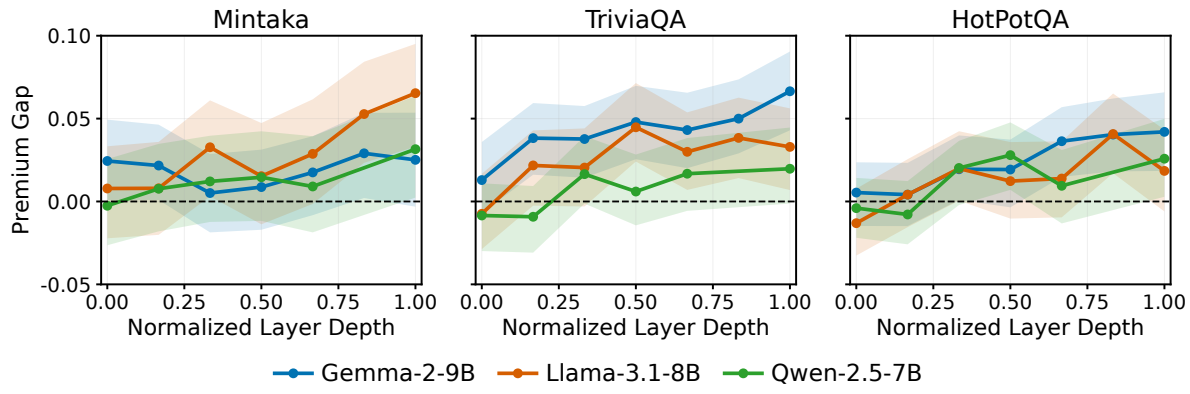
This layer-level analysis strengthens the domain-specificity conclusion from Section 4.3. Mathematical correctness signals appear to be publicly accessible at every depth of the network, suggesting that reasoning difficulty is governed by problem structure rather than model-specific knowledge.

5.3 What Drives the Correctness Signal?

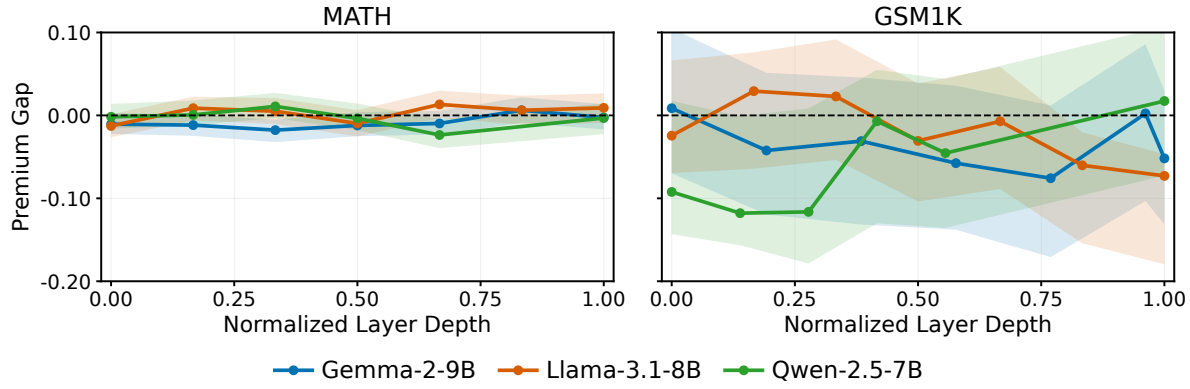
The layer-level analysis above localizes where privileged knowledge emerges, but does not address what linguistic features drive correctness prediction more broadly. In Appendix G, we strip questions of their syntax to isolate named entities and nouns, finding that concept-level familiarity accounts for a substantial portion of the predictive performance in factual tasks and MATH, whereas GSM1K correctness relies on structural problem features that lexical stripping destroys.

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. Evaluating on



(a) Factual Knowledge.



(b) Mathematical Reasoning.

Figure 5: **Per-Layer Premium Gap.** Premium gap (self-probe AUC − best external-probe AUC) on the disagreement subset as a function of normalized layer depth. Shaded regions denote 95% confidence intervals. **(a)** For factual tasks, the gap is near zero in early layers and grows progressively toward deeper layers across all models, indicating that privileged knowledge emerges in early-to-mid representations. **(b)** For mathematical reasoning, MATH fluctuates near zero at all depths and GSM1K is predominantly negative (external probes outperform self-probes). No layer exhibits a consistent self-probe advantage across the full network depth.

disagreement subsets reveals that privileged knowledge is domain-specific — it emerges consistently in factual tasks, while mathematical reasoning correctness remains externally observable.

These findings reconcile prior conflicting results: privileged knowledge does exist, but is domain-specific and was previously masked by inter-model agreement. Beyond correctness prediction, our disagreement-based methodology can be extended to study privileged knowledge in hybrid domains (coding, commonsense reasoning) and other forms of model introspection. Practically, our results suggest that model-specific activations carry signals that black-box tools miss, with potential applications in hallucination detection and monitoring.

Our probe-based analysis is correlational in nature, and the causal mechanisms underlying the privileged knowledge signal remain an open question. A natural avenue for future work is activation steering: if the factual correctness signal is genuinely tied to subject-specific retrieval, intervening on the identified correctness direction in the residual stream should predictably modulate output correctness.

7 Limitations

Our analysis has several limitations: (1) Although we additionally evaluate `Qwen-3-32B` (Appendix A.1), our main analysis is limited 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. (4) Our study reveals systematic patterns linking representational structure to correctness; complementary intervention experiments could further establish the causal mechanisms underlying 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.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

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.

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, Anton Tsitsulin, and 178 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

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.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025b. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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, Junyang Lin, Kai Dang, and 23 others. 2025c. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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 Larger Model: Qwen-3-32B

To assess scalability, we evaluate `Qwen-3-32B` using the same pipeline as the main experiments, incorporating it both as a target model and as an external probe candidate for the three main models. As a significantly larger model, `Qwen-3-32B` likely produces richer public feature representations, which would be expected to benefit external probes—making it a stronger test of our claims, while also strengthening the best cross-model baseline for the three main models; results therefore differ slightly from Figure 3.

Layer-averaged and best-layer results are presented in Table 1 respectively, and are consistent with our main findings. On the full test set, no model exhibits a reliable premium gap, consistent with our inter-model agreement analysis (Section 4.2). On the disagreement subset, self-probe advantages emerge in factual tasks across all models, most consistently on TriviaQA and HotPotQA, while Mintaka shows weaker and less consistent gains. `Qwen-3-32B` follows the same pattern, with statistically significant advantages on TriviaQA and HotPotQA in both evaluation settings.

A.2 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 D).

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 emergence of privileged knowledge in factual tasks is fully consistent across both probe types. Both linear and MLP probes detect a significant premium gap in all 9 factual configurations, confirming that the target model retains unique internal signals inaccessible to external observers. Conversely, in mathematical reasoning (GSM1K, MATH), the premium gap remains absent under both probe types, 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 2. Across all configurations, the disagreement subsets retain sufficient scale ($\approx 20\%$ of the original data).

C Detailed Disagreement Performance

This section provides the detailed performance breakdowns on the disagreement subsets across all datasets, comparing self-probes against external probes for both linear and MLP configurations.

Detailed results for factual knowledge tasks are shown in Figure 8 (linear) and Figure 10 (MLP). Corresponding results for mathematical reasoning tasks are presented in Figure 9 (linear) and Figure 11 (MLP).

D Implementation Details

D.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

	Target Model	Mintaka	TriviaQA	HotPotQA	MATH	GSM1K
Full	Gemma-2-9B	**+0.017*** (6.5%)	**+0.018*** (6.7%)	**+0.021*** (8.9%)	−0.006 (−3.4%)	−0.025 (−8.3%)
	Llama-3.1-8B	**+0.002** (0.8%)	**+0.000** (0.1%)	−0.012 (−5.6%)	−0.005 (−1.9%)	−0.014 (−4.8%)
	Qwen-2.5-7B	−0.005 (−2.0%)	−0.010 (−4.0%)	−0.007 (−3.6%)	−0.000 (−0.1%)	−0.011 (−3.8%)
	Qwen-3-32B	−0.014 (−5.6%)	−0.010 (−3.6%)	−0.006 (−3.0%)	−0.013 (−5.2%)	−0.020 (−4.5%)
Disagree	Gemma-2-9B	**+0.023** (5.1%)	**+0.034*** (6.8%)	**+0.025*** (4.8%)	−0.010 (−3.0%)	−0.133 (−36.6%)
	Llama-3.1-8B	**+0.032*** (6.4%)	**+0.018*** (3.5%)	**+0.020*** (4.2%)	−0.011 (−1.8%)	−0.090 (−24.2%)
	Qwen-2.5-7B	**+0.016** (3.2%)	**+0.015*** (3.4%)	**+0.015*** (3.4%)	−0.004 (−0.9%)	−0.173 (−51.0%)
	Qwen-3-32B	−0.000 (−0.1%)	**+0.022*** (4.1%)	**+0.015*** (3.2%)	−0.009 (−2.0%)	−0.024 (−4.4%)

(a) Averaged over all layers

	Target Model	Mintaka	TriviaQA	HotPotQA	MATH	GSM1K
Full	Gemma-2-9B	**+0.011** (4.5%)	**+0.031*** (12.5%)	**+0.024*** (11.1%)	−0.004 (−2.9%)	−0.028 (−10.3%)
	Llama-3.1-8B	−0.011 (−4.5%)	−0.007 (−2.9%)	−0.012 (−5.9%)	−0.010 (−4.3%)	−0.009 (−3.3%)
	Qwen-2.5-7B	**+0.009** (3.7%)	−0.005 (−2.1%)	**+0.000** (0.2%)	**+0.004** (2.3%)	−0.029 (−11.8%)
	Qwen-3-32B	−0.005 (−2.0%)	**+0.001** (0.5%)	−0.003 (−1.6%)	−0.016 (−6.6%)	−0.018 (−4.5%)
Disagree	Gemma-2-9B	**+0.056*** (11.7%)	**+0.081*** (15.5%)	**+0.035*** (6.9%)	**+0.005** (1.8%)	−0.145 (−45.2%)
	Llama-3.1-8B	**+0.065*** (12.6%)	**+0.030*** (6.0%)	**+0.042*** (8.3%)	−0.007 (−1.3%)	−0.090 (−26.8%)
	Qwen-2.5-7B	**+0.033** (7.0%)	**+0.039*** (8.6%)	**+0.031*** (6.9%)	**+0.011** (3.4%)	−0.096 (−24.4%)
	Qwen-3-32B	**+0.005** (1.2%)	**+0.012** (3.0%)	**+0.045*** (9.1%)	**+0.001** (0.1%)	**+0.015** (2.7%)

(b) Averaged over all layers

Table 1: Correctness prediction results with Qwen-3-32B included. Each cell reports ΔAUC (Self − Best Cross) with gap closed (%) in parentheses. **Bold** indicates self-probe advantage (positive ΔAUC). *: $p < 0.05$, Holm–Bonferroni corrected. Results for Gemma-2-9B, Llama-3.1-8B, and Qwen-2.5-7B differ slightly from Figure 3 as Qwen-3-32B representations are included as an additional external probe candidate, strengthening the best cross-model baseline.

		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 2: 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).

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.

D.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.

E 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.

E.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 genera-

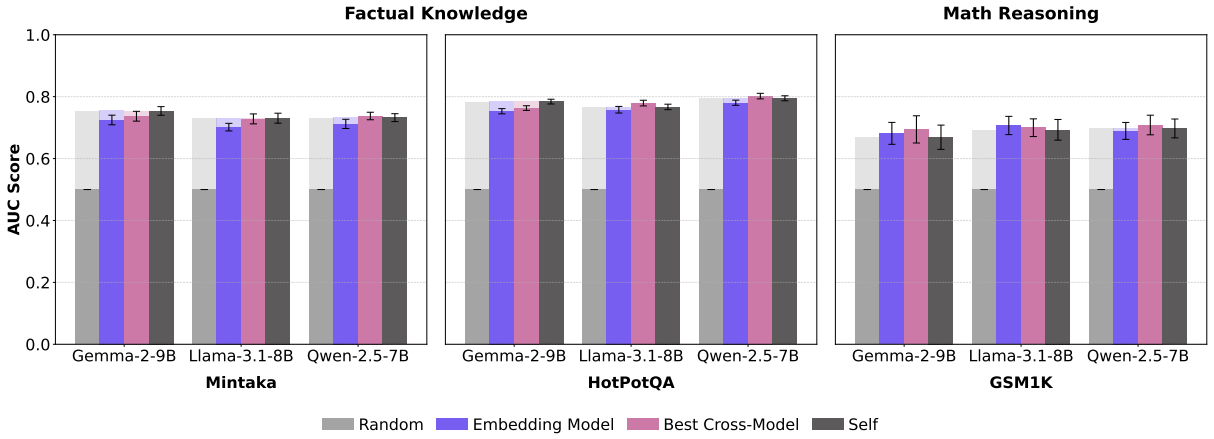


Figure 6: **Premium Gap (Remaining Datasets).** Mean AUC for correctness prediction, averaged over layers, on two task types: factual knowledge (Mintaka, HotPotQA) and mathematical reasoning (GSM1K). Bars compare Random, Embedding, and Best Cross-Model baselines to the Self-Probe (*Self*) across three target models. Semi-transparent overlays indicate the performance gain (or lack thereof) of *Self* relative to each baseline. Error bars denote 95% confidence intervals.

tion, and `max_new_tokens=2048` for mathematical reasoning to accommodate full Chain-of-Thought derivations.

E.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 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.3 Distributional Shifts vs. Privileged Knowledge

A potential alternative explanation for the premium gap observed on disagreement subsets is a general distributional shift near the decision boundary: harder examples may induce representational differences that benefit self-probes independently of any privileged knowledge. However, this explanation predicts that both self and peer probes should degrade similarly when restricted to disagreement subsets, as the increased difficulty would affect all probes equally. Instead, peer probes degrade substantially on disagreement subsets while self-probes retain a consistent and statistically significant advantage, suggesting the premium gap reflects genuine privileged signals specific to the target model's internal states rather than a general distributional artifact.

F Per-Layer Analysis: Additional Figures

Figures 5a and 5b in the main text shows the per-layer premium gap for both factual and mathematical reasoning datasets. Here we provide per-model bar figures showing absolute self and best-external AUC values at each probed layer on the disagreement subset (Figures 13 to 15). In each bar figure, the lighter bar represents the best external-probe AUC and the darker bar represents the self-probe AUC; the premium gap is directly visible as the height difference.

G Probing the Sources of Correctness Signals

Our main analysis establishes that privileged knowledge emerges in factual tasks but not in mathematical reasoning. A related question is what information drives correctness prediction in each domain. To test whether entity identity alone can partially account for the correctness signal, we introduce a **Lexical-Only** control: we retain only named entities and nouns from the question (discarding syntax and function words) and train probes on the target model's hidden states when processing this

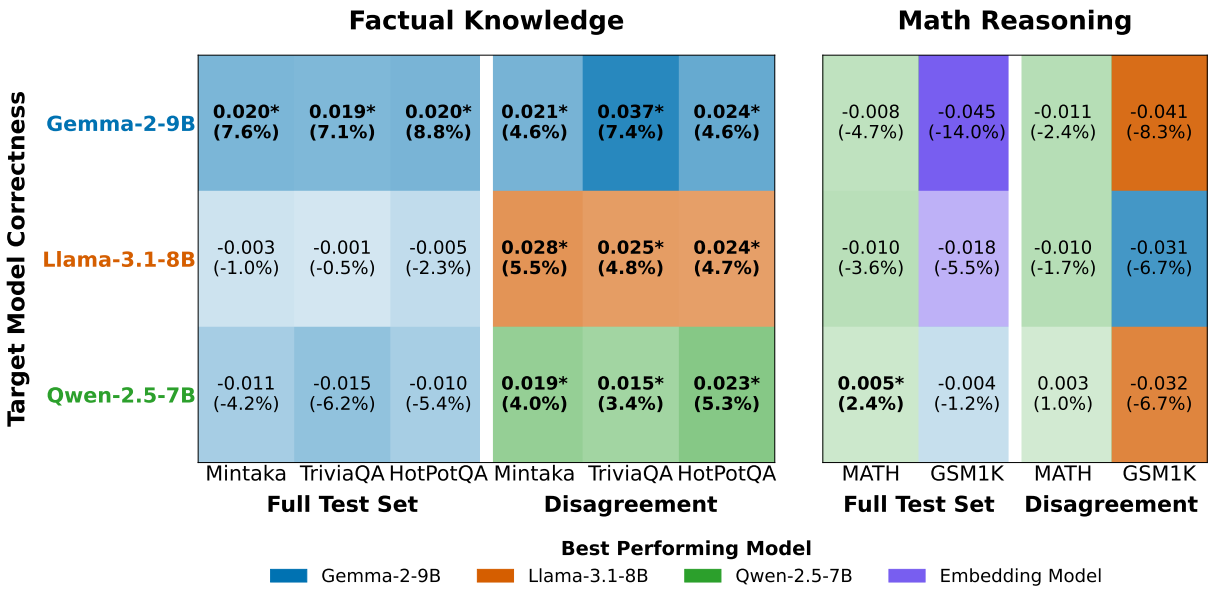


Figure 7: **Target Model Correctness Prediction (MLP Probes).** Heatmap of correctness prediction differences across target models, datasets, and test subsets. Each cell reports the AUC difference ($\Delta\text{AUC} = \text{Self} - \text{Best External}$), with the percentage of the gap closed shown in 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).

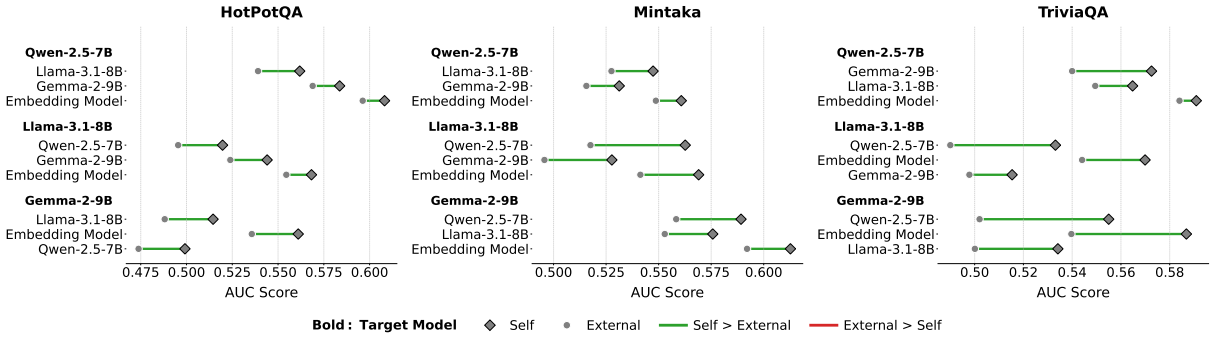


Figure 8: **Disagreement Gap: Factual Knowledge Breakdown (Linear Probes).** Detailed performance on the disagreement subset across Mintaka, TriviaQA, and HotPotQA. Self-probes consistently outperform external probes across all factual datasets, reinforcing the existence of privileged knowledge in factual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

stripped input. If correctness prediction is driven by entity-level familiarity, entity tokens alone should recover substantial performance even without syntactic context.

Implementation. We extract concepts from the original question using a two-stage pipeline: (1) named entities via `GLiNER` (`urchade/gliner_medium-v2.1`) across 20 label types, and (2) noun chunks via `spaCy` (`en_core_web_sm`). The union is deduplicated, removing substrings and stopwords. The resulting concepts are formatted as: *"This text discusses [Concept A], [Concept B], and [Concept C]."* This synthesized text replaces the original question as input to the target model, and probes are trained on the extracted hidden states to predict correctness on the *original* question. All other experimental configurations follow Section 3.4.

Results. As shown in Figure 16, Lexical-Only probes recover a non-trivial portion of the original probe's performance across all datasets except GSM1K. In factual domains (Mintaka, TriviaQA, HotpotQA), lexical features recover 53.7%, 75.0%, and 73.5% of original probe performance relative to the random baseline (0.5 AUC), suggesting that much of the correctness signal stems from concept-level familiarity. MATH shows a similar pattern, recovering 75.6% of performance, likely because

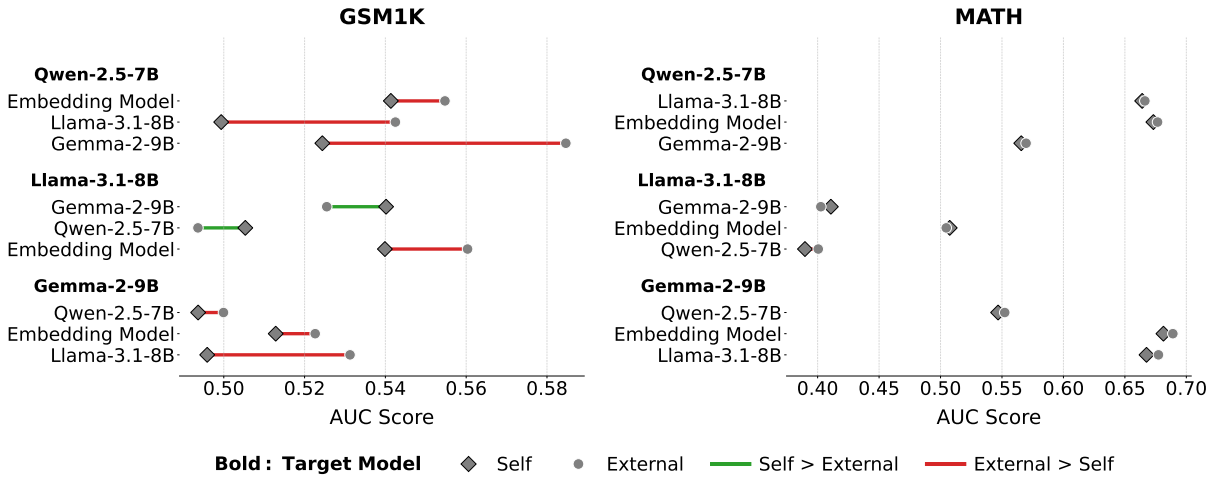


Figure 9: **Disagreement Gap: Mathematical Reasoning Breakdown (Linear Probes).** Detailed performance on the disagreement subset across GSM1K and MATH. Unlike factual tasks, mathematical correctness shows no consistent premium gap, indicating that reasoning difficulty is a public feature accessible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

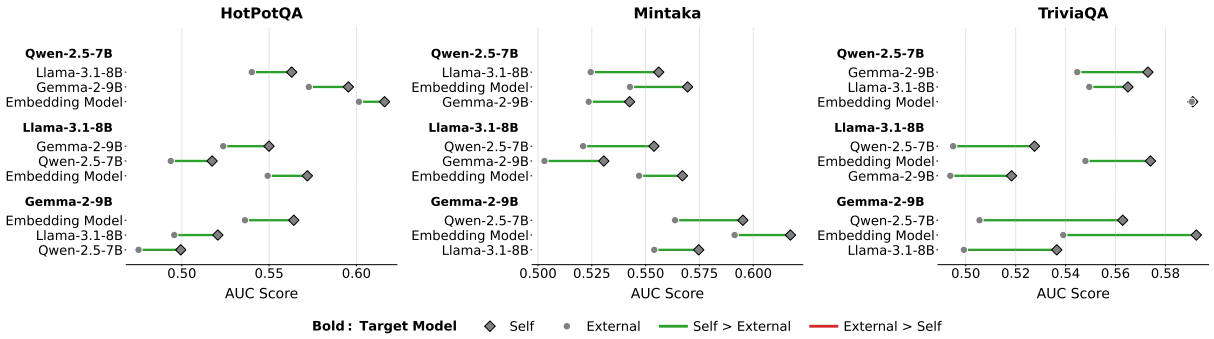


Figure 10: **Disagreement Gap: Factual Knowledge Breakdown (MLP Probes).** Detailed disagreement subset performance using MLP probes. The premium gap is even more pronounced with non-linear probes, with Self-representations outperforming Best External probes in 9 out of 9 configurations.

lexical tokens encode mathematical topic indicators (e.g., "eigenvalue", "asymptote") that correlate with difficulty. In contrast, GSM1K drops to near chance (AUC ≈ 0.49), consistent with its correctness being governed by computational structure rather than surface tokens such as "savings account" or "$50". These results indicate that correctness prediction in factual tasks, and to some extent in MATH, relies substantially on concept-level signals, whereas GSM1K correctness depends on structural problem features that lexical stripping destroys.

H 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.

I 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., 2020), 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.

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

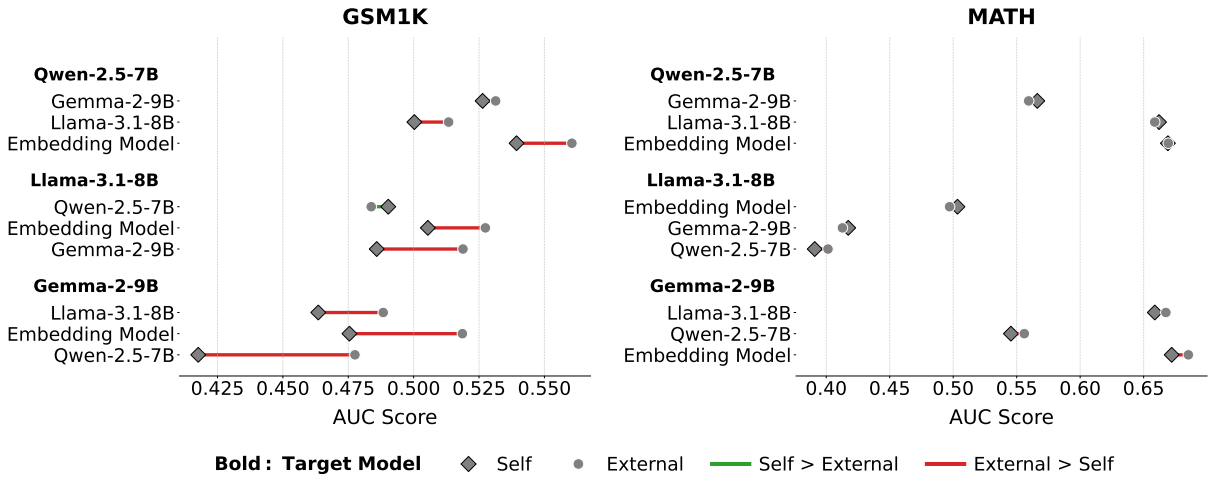


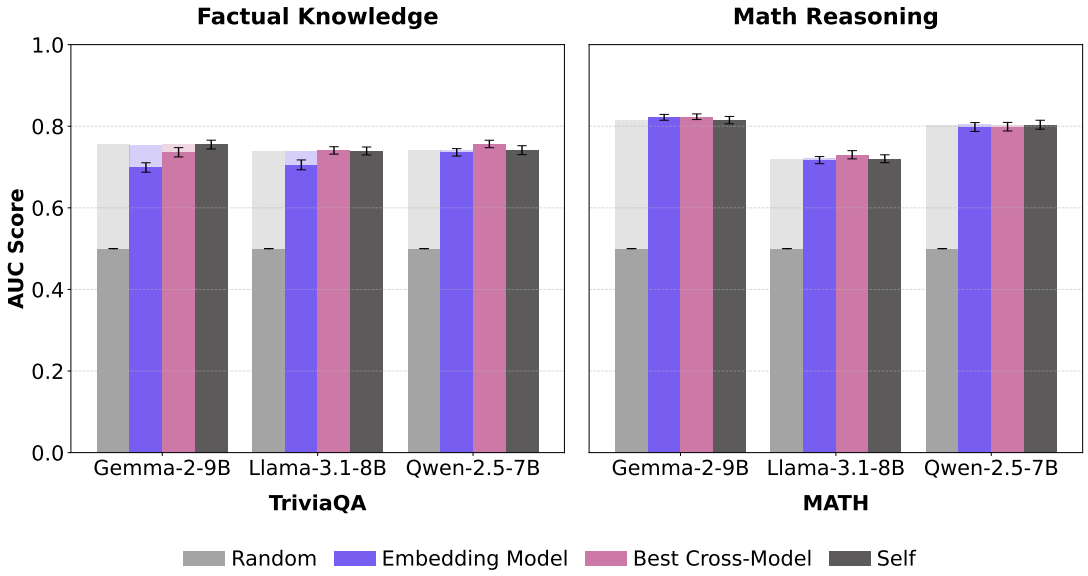
Figure 11: **Disagreement Gap: Mathematical Reasoning Breakdown (MLP Probes).** Detailed disagreement subset performance using MLP probes across GSM1K and MATH. Consistent with linear results, increased probe expressivity does not uncover hidden privileged info in math tasks; external models remain effective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

Apache 2.0.

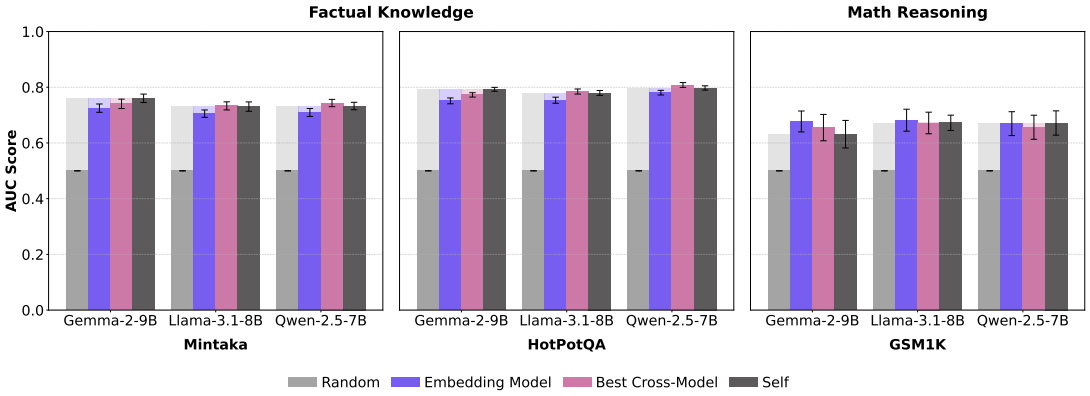
- **GSM1K** (Zhang et al., 2024) and **MATH** (Hendrycks et al., 2021) are licensed under the MIT License.

J 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.

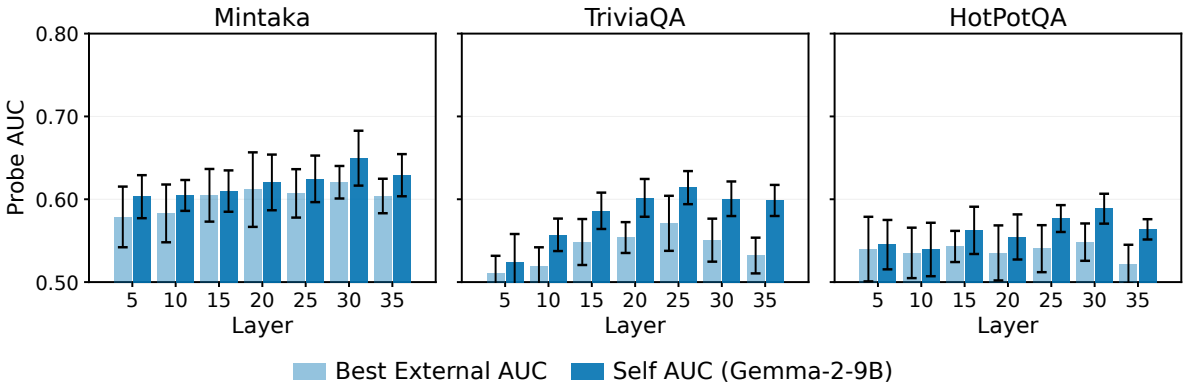


(a) TriviaQA (Factual Knowledge) and MATH (Mathematical Reasoning).

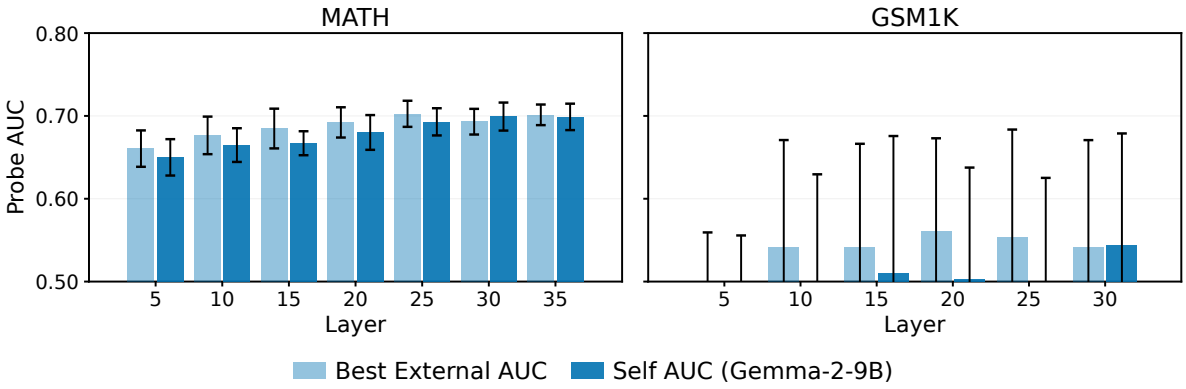


(b) Remaining datasets: Mintaka, HotPotQA (Factual) and GSM1K (Math).

Figure 12: **Premium Gap (MLP Probes).** Mean AUC for correctness prediction, averaged over layers. 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. **(a)** Core comparison on TriviaQA and MATH, mirroring the logistic-regression analysis in Figure 2. **(b)** Extended results on the remaining datasets.

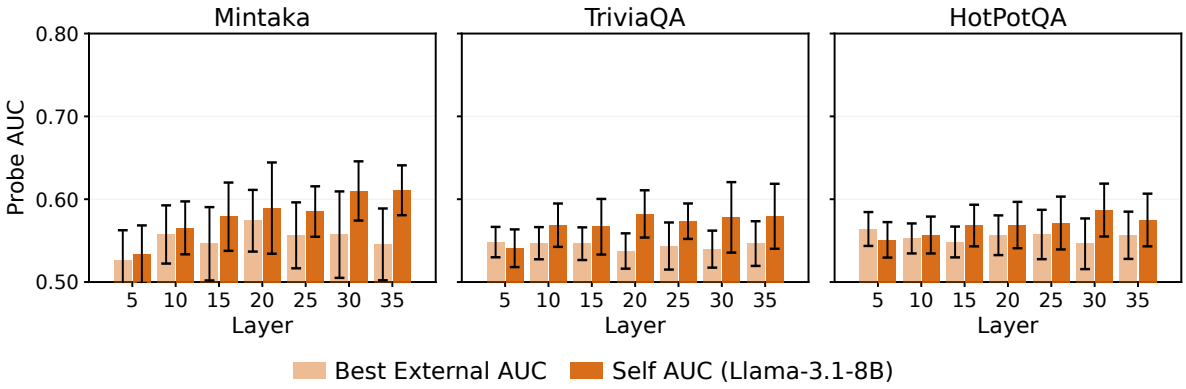


(a) Factual Knowledge (Mintaka, TriviaQA, HotPotQA).

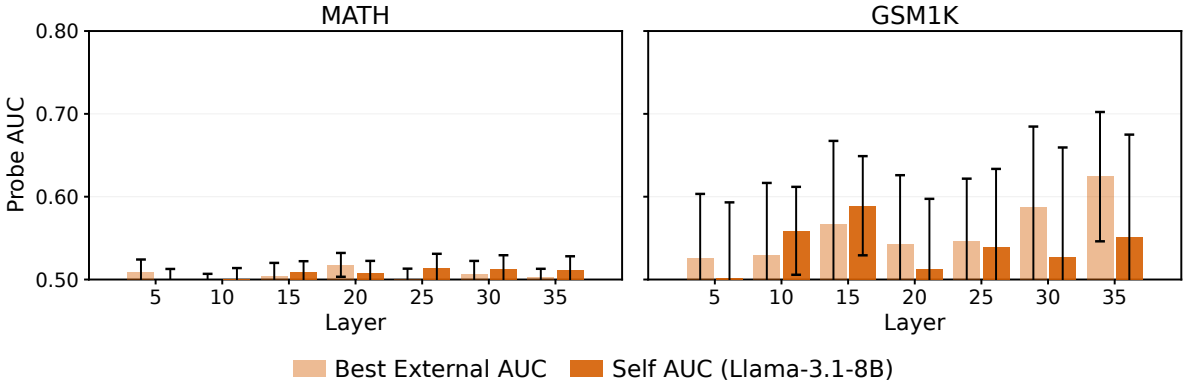


(b) Mathematical Reasoning (MATH, GSM1K).

Figure 13: **Per-Layer AUC: Gemma-2-9B.** Absolute AUC at each probed layer on the disagreement subset. Lighter bars: best external probe; darker bars: self-probe. **(a)** For factual datasets, the self-probe advantage grows visibly from mid-layers onward, particularly in TriviaQA. **(b)** For mathematical reasoning, bars are of similar or reversed height, consistent with the absence of a premium gap. Error bars denote 95% confidence intervals from cross-validation fold scores.

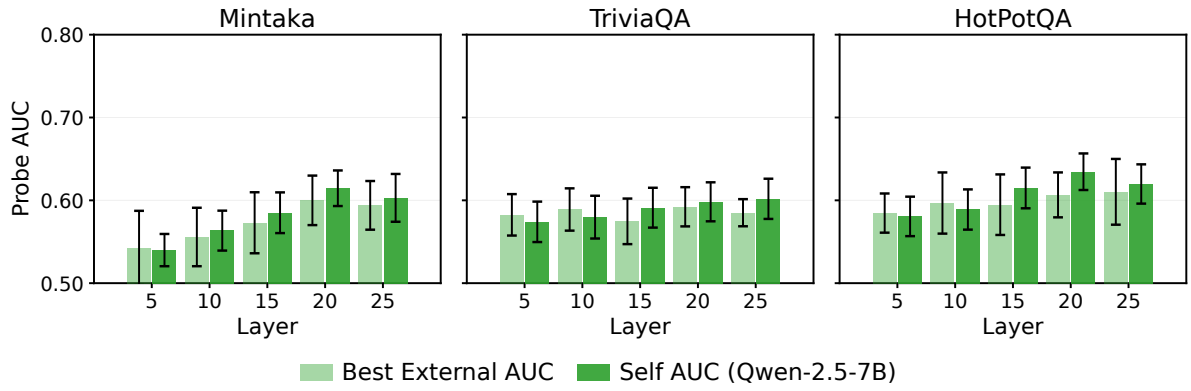


(a) Factual Knowledge (Mintaka, TriviaQA, HotPotQA).

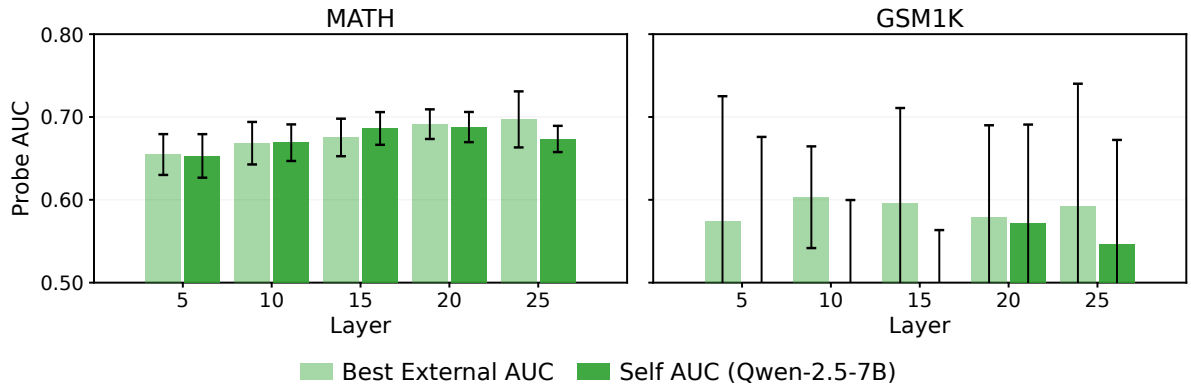


(b) Mathematical Reasoning (MATH, GSM1K).

Figure 14: **Per-Layer AUC: Llama-3.1-8B.** Absolute AUC at each probed layer on the disagreement subset. Lighter bars: best external probe; darker bars: self-probe. **(a)** For factual datasets, the self-probe advantage emerges in early-to-mid layers. **(b)** For mathematical reasoning, bars are of similar or reversed height, consistent with the absence of a premium gap. Error bars denote 95% confidence intervals from cross-validation fold scores.



(a) Factual Knowledge (Mintaka, TriviaQA, HotPotQA).



(b) Mathematical Reasoning (MATH, GSM1K).

Figure 15: **Per-Layer AUC: Qwen-2.5-7B.** Absolute AUC at each probed layer on the disagreement subset. Lighter bars: best external probe; darker bars: self-probe. **(a)** For factual datasets, the self-probe advantage emerges in early-to-mid layers. **(b)** For mathematical reasoning, bars are of similar or reversed height, consistent with the absence of a premium gap. Error bars denote 95% confidence intervals from cross-validation fold scores.

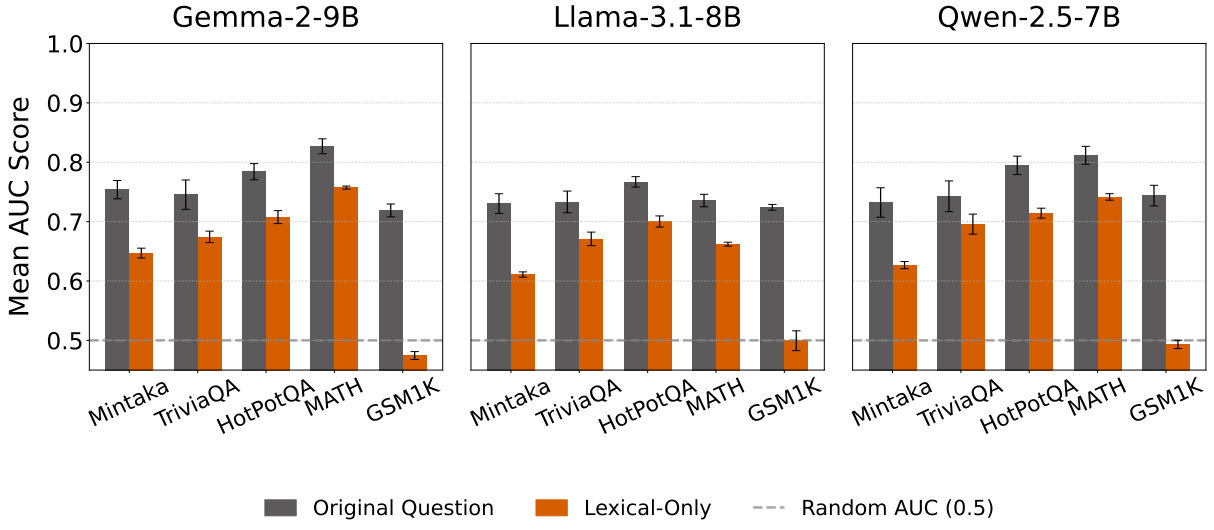


Figure 16: **Lexical-Only vs. Original Question.** Mean AUC for correctness prediction, averaged over layers, of probes trained on the *Original Question* versus the *Lexical-Only* input (named entities and nouns only), aggregated across all models (`Gemma-2-9B`, `Llama-3.1-8B`, `Qwen-2.5-7B`). The gap between conditions reflects the contribution of syntactic and contextual processing beyond entity identity. Error bars denote 95% confidence intervals from cross-validation fold scores.