How Reliable are Confidence Estimators for Large Reasoning Models? A Systematic Benchmark on High-Stakes Domains

Reza Khanmohammadi[1*], **Erfan Miahi**[2], **Simerjot Kaur**[3],
Charese H. Smiley[3], **Ivan Brugere**[3], **Kundan Thind**[4,†], **Mohammad M. Ghassemi**[1,†]

[1]Michigan State University [2]Independent AI Researcher
[3]JPMorgan AI Research [4]Henry Ford Health
[1] `{khanreza, ghassem3}@msu.edu` [2] `mhi.erfan1@gmail.com`
[3] `{simerjot.kaur, ivan.brugere, charese.h.smiley}@jpmchase.com` [4] `kthind1@hfhs.org`

Abstract

The miscalibration of Large Reasoning Models (LRMs) undermines their reliability in high-stakes domains, necessitating methods to accurately estimate the confidence of their long-form, multi-step outputs. To address this gap, we introduce the Reasoning Model Confidence estimation Benchmark (RMCB), a public resource of 347,496 reasoning traces from six popular LRMs across different architectural families. The benchmark is constructed from a diverse suite of datasets spanning high-stakes domains, including clinical, financial, legal, and mathematical reasoning, alongside complex general reasoning benchmarks, with correctness annotations provided for all samples. Using RMCB, we conduct a large-scale empirical evaluation of over ten distinct representation-based methods, spanning sequential, graph-based, and text-based architectures. Our central finding is a persistent trade-off between discrimination (AUROC) and calibration (ECE): text-based encoders achieve the best AUROC (0.672), while structurally-aware models yield the best ECE (0.148), with no single method dominating both. Furthermore, we find that increased architectural complexity does not reliably outperform simpler sequential baselines, suggesting a performance ceiling for methods relying solely on chunk-level hidden states. This work provides the most comprehensive benchmark for this task to date, establishing rigorous baselines and demonstrating the limitations of current representation-based paradigms.

Code — https://github.com/ledengary/RMCB
Data — https://huggingface.co/datasets/ledengary/RMCB

1 Introduction

Despite impressive performance, large language models (LLMs) often struggle with confidence calibration. The model's estimated probability that an answer is correct frequently misaligns with the actual outcome, leading to high confidence in wrong answers and low confidence in right ones. This unreliability forces costly manual review of every output, which undermines the primary benefits of automation. The problem is especially critical in high-stakes domains like medicine, finance, and law, where a single miscalibrated output can have significant consequences. Therefore, accurate confidence scores are essential to build trustworthy systems, allowing users to efficiently determine which outputs require human verification. A reliable confidence score must be both well-calibrated, ensuring the predicted probability aligns with the true likelihood of correctness, and discriminative, effectively distinguishing correct answers from incorrect ones.

This challenge is particularly pronounced for Large Reasoning Models (LRMs), which articulate their problem-solving process through a sequence of intermediate thoughts before providing a final answer. While these models often achieve superior performance, their long and detailed outputs substantially increase the cost and complexity of manual verification. Current confidence estimation techniques, which primarily analyze logits or probe hidden states, were developed for short-form, token-level prediction and are not designed for these long-form outputs. The confidence in a multi-step argument is an emergent property of the entire reasoning trajectory, not a static feature of any single component. Our comprehensive benchmark confirms this mismatch, revealing that existing methods consistently fail to achieve both high discriminative power and good calibration for LRMs. Consequently, a significant gap remains for a generalizable framework that can reliably assess the confidence of multi-step reasoning across diverse model architectures and high-stakes domains.

In this work, we make two primary contributions. First, we introduce the **R**easoning **M**odel

---

[*] Corresponding author: `khanreza@msu.edu`
[†] Shared senior authorship

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1669–1754
March 24-29, 2026 ©2026 Association for Computational Linguistics

Confidence estimation **B**enchmark (**RMCB**), the first large-scale, publicly available resource dedicated to this task. The benchmark comprises 347,496 reasoning traces generated by six popular LRMs. Whereas prior studies have often focused on a narrow set of domains like programming or mathematics, or drawn from specific tasks within broader evaluations, RMCB is constructed from a diverse suite of full datasets spanning high-stakes domains, including clinical, financial, and legal reasoning, alongside general knowledge understanding. A core component of RMCB is its detailed annotation; we release all model inferences with corresponding correctness labels covering every evaluated response.

Building on this resource, we conduct a comprehensive empirical study evaluating over ten distinct confidence estimation methods. Our investigation covers the full spectrum of representation-based techniques, from baseline approaches like verbalized confidence and text classification with specialized encoders, to more complex models that operate on sequences of hidden states, including stacked sequential models and various graph-based architectures. This large-scale evaluation reveals our second primary contribution: the empirical finding that methods relying on hidden-state representations, regardless of their architectural complexity, hit a consistent performance ceiling and face a persistent trade-off between discriminative power (AUROC) and calibration (ECE). We show that even sophisticated models designed to capture the relational structure of the reasoning trace fail to consistently outperform simpler baselines. Ultimately, this work rigorously quantifies the limitations of the current paradigm and establishes a clear set of baselines that highlight the need for future research to explore alternative signal sources beyond static hidden states.

2 Related Work

The estimation of an LLM's confidence is a well-established field of study, with methods generally falling into several distinct categories. One of the most direct approaches is *verbalized confidence*, where a model is explicitly prompted to state its own certainty level in natural language, which is then parsed into a numerical score (Lin et al., 2022). Another common black-box technique is *self-consistency*, which assesses the robustness of an answer by generating multiple responses with sampling and measuring the consensus among them (Wang et al., 2023). Moving to white-box methods, a significant body of work focuses on *generative signals* derived from the model's output logits. These techniques analyze the properties of the token-level probability distributions, using features like the log-probability of the generated sequence (Zhou et al., 2025) or applying post-hoc calibration methods such as temperature scaling (Jiang et al., 2021). Other white-box approaches operate on the model's *internal hidden states*, training lightweight classifiers on hidden-state representations to predict correctness. This includes methods like P(IK), which uses the hidden state of the input prompt (Kadavath et al., 2022), SAPLMA, which identifies particular hidden layers whose activations best capture correctness signals (Azaria and Mitchell, 2023), and InternalInspector, which leverages the full set of hidden states across all layers for confidence estimation (Beigi et al., 2024). A more recent and distinct category involves probing the model's *internal stability*. For instance, CCPS (Khanmohammadi et al., 2025) introduces targeted perturbations to a model's final hidden states and measures the resulting representational shift, using the magnitude of this shift as a proxy for the model's underlying confidence.

However, applying these established methods to LRMs presents a series of fundamental challenges. The long-form, multi-step nature of LRM outputs makes many of these techniques computationally intractable or conceptually misaligned. Self-consistency, for example, becomes prohibitively expensive due to the high cost of generating multiple, thousand-token reasoning traces for a single query. Verbalized confidence may be unreliable, as the step-by-step reasoning process can lead a model to become overconfident in its final conclusion, even when the underlying logic is flawed. Methods relying on token-level signals, such as those analyzing generative log-probabilities or probing internal stability like CCPS, struggle with extreme representational scale—storing and processing full logit vectors or computing perturbations across thousands of tokens per sample is not a feasible strategy for producing a single, coherent representation.

This infeasibility of token-level analysis for long-form reasoning has naturally led to a focus on higher-level representations, with two notable approaches shaping the initial exploration of this problem. The first, from Yoon et al. (2025) (YVCE), adapts the verbalized confidence approach for

LRMs. The second, and more foundational for representation-based methods, is Probing Hidden States for Self-Verification (PHSV) by Zhang et al. (2025). PHSV trains a lightweight MLP on the hidden state of each intermediate reasoning "chunk," defined as a contiguous segment of the reasoning trace that ends in an intermediate answer. At inference time, the classifier is applied sequentially over chunks, and the confidence score is determined by the first chunk exceeding a threshold, or by the final chunk if none do.

While these methods provide a crucial starting point, the field of LRM confidence estimation remains underexplored. The initial works were largely evaluated in isolation, without a comprehensive comparison against a broader set of architectural alternatives. Furthermore, their evaluations have often focused on a limited set of domains, such as mathematics or general reasoning, leaving their generalizability to diverse, high-stakes applications an open question. This work is motivated by three critical gaps in the literature: (1) the lack of a large-scale, public benchmark for LRM confidence estimation that spans multiple high-stakes domains; (2) the absence of a systematic comparison between the foundational chunk-level methods and more complex sequential, graph-based, and text-based architectures; and (3) an incomplete understanding of the trade-offs between discrimination and calibration for these representation-based methods. We aim to fill these gaps by providing the first comprehensive benchmark that directly addresses these challenges.

3 The RMCB Benchmark

To address the gap in LRM confidence estimation, we constructed the RMCB benchmark, a large-scale, publicly available resource designed to facilitate the systematic evaluation of confidence estimation methods. This section details the data sources, model suite, and annotation methodology used to build the benchmark.

3.1 Data Sources & LRM Suite

The foundation of RMCB is a diverse collection of reasoning problems sourced from well-established public benchmarks. Whereas prior studies have often focused on a narrow set of domains or specific tasks within broader evaluations, our benchmark is constructed from full datasets spanning multiple high-stakes domains—specifically clinical, financial, and legal reasoning—alongside challenging mathematical and general commonsense reasoning tasks. As shown in Table 1, the benchmark is split into two entirely disjoint sets: a balanced training set of 10,000 samples and a challenging evaluation set of over 50,000 samples. A full breakdown of these datasets, including our standardized curation process, is provided in Appendix B. A complete list of dataset sources, access methods, and exact versions used in this benchmark is provided in Appendix B.2.

Table 1: Distribution of datasets used for the RMCB training and evaluation splits. The test set is entirely disjoint from the training set. Domains indicate the primary reasoning type assessed.

Dataset	**Domain**	**Split**	**Samples**
Training Datasets (Total = 10,000)			
GSM8K	Mathematical	Train	1,000
TAT-QA	Financial	Train	1,000
MedQA	Medical	Train	1,000
LEXam	Legal	Train	1,000
ARC	General	Train	1,000
CommonsenseQA2	General	Train	1,000
LogiQA	General	Train	1,000
OpenBookQA	General	Train	1,000
QuaRTz	General	Train	1,000
ReClor	General	Train	1,000
Evaluation Datasets (Total = 51,951)			
MATH	Mathematical	Test	5,000
FinQA	Financial	Test	1,138
MedMCQA	Medical	Test	6,150
LegalBench	Legal	Test	21,167
MMLU-Pro	General	Test	11,987
BBH	General	Test	6,509
Grand Total = 61,951			

To generate the reasoning traces for our benchmark, we selected a suite of six popular and powerful open-weight LRMs. These models represent diverse architectural families (Phi, Qwen, Mistral, EXAONE) and a wide range of parameter counts (from 4B to 33B), establishing a broad and representative sample of modern reasoning capabilities. Specifically, our evaluation includes `Phi-4-mini-flash-reasoning` (3.85B parameters), `Qwen3-8B` (8.19B parameters), `Qwen3-14B` (14.8B parameters), `Magistral-Small-2506` (23.6B parameters), `EXAONE-Deep-32B` (32.0B parameters), and `QwQ-32B` (32.8B parameters). Detailed architectural parameters for each LRM are provided in Appendix A.

3.2 Data Generation & Annotation

We generated responses from each of the six LRMs for every prompt, resulting in a total corpus of 347,496 unique reasoning traces. This count reflects the valid, high-quality generations remaining after filtering. During large-scale inference, a small fraction of outputs (approximately 6.5%) were excluded to ensure benchmark integrity. These exclusions were primarily due to generation failures such as empty outputs or repetition loops. As a result, the number of successfully annotated samples varies slightly across LRMs, as detailed in Table 28 (e.g., 44,409 test samples for `Phi-4-mini-flash-reasoning` versus 50,792 for `QwQ-32B`). We prioritize the validity and quality of reasoning traces over enforcing a uniform sample count across architectures. All remaining outputs were generated using deterministic decoding (temperature of 0.0) with a maximum length of 4096 tokens to ensure reproducibility; further details on the generation setup are available in Appendix B.5. To enable a granular, step-by-step analysis, each generated response was first segmented into coherent units of thought, or "chunks." Our process is an enhancement of the methodology first introduced by Zhang et al. (2025), which groups paragraphs based on a set of linguistic keywords that signal a shift in the reasoning process (e.g., self-correction, verification). Further details on our segmentation methodology and the expanded keyword list can be found in Appendix B.6. An example reasoning trace illustrating the chunking process is provided in Appendix B.8.

A core contribution of RMCB is its detailed annotation of correctness. Each of the 347,496 reasoning traces was systematically graded to ensure comprehensive evaluation across all responses. For overall correctness, multiple-choice answers were checked via string matching, while open-ended responses were evaluated for semantic equivalence by a state-of-the-art LLM judge (`GPT-5-nano`), a practice validated by recent literature (Kapoor et al., 2024; Zhang et al., 2025; Chaudhry et al., 2024; Khanmohammadi et al., 2025). To provide a more granular supervisory signal, we extended this grading to the per-chunk level, applying the same LLM-based judging process to each individual reasoning chunk. This allows us to obtain a correctness label for each intermediate step in the reasoning process. Full details on the prompts and methodology for this automated grading are provided in Appendix B.7. The final sample- and chunk-level statistics after this comprehensive annotation process are available in Appendix Tables 27 and 28.

3.3 Benchmark Components

We publicly release all components of the RMCB to accelerate research on reasoning confidence estimation. The released artifacts include: (a) a comprehensive dataset of 347,496 reasoning traces, including input prompts and model-generated responses from six LRMs; (b) complete correctness annotations provided for every trace; and (c) the full implementation of all evaluated confidence estimation methods, enabling reproducible and extensible experimentation.

4 Experimental Methodology

To systematically evaluate confidence estimation methods on our RMCB benchmark, we designed a rigorous experimental framework. All LRMs are kept frozen throughout our experiments; no fine-tuning is performed on the underlying reasoning models. All learning occurs exclusively in the downstream, post-hoc confidence estimators. This section details the task formulation, the metrics used for evaluation, the suite of methods investigated, and the consistent protocol for training and hyperparameter optimization.

4.1 Task Formulation

We formulate LRM confidence estimation as a binary classification problem. For a given reasoning trace generated by an LRM, the task is to produce a single scalar probability score $p \in [0, 1]$ that predicts whether the final answer is correct (label 1) or incorrect (label 0). This unified formulation enables direct comparison across all methods.

4.2 Evaluation Metrics

A reliable confidence score must satisfy two distinct criteria: it must be discriminative (able to distinguish correct from incorrect answers) and well-calibrated (its predicted probability must align with the true likelihood of correctness). To capture this fundamental trade-off, we focus on two primary metrics:

Area Under the ROC Curve (AUROC): Measures the model's discriminative power by evaluating its ability to rank correct answers higher than incorrect ones across all thresholds. A score of 1.0 represents perfect discrimination.

Expected Calibration Error (ECE): Measures the alignment between a model's predicted confidence and its empirical accuracy. A lower ECE indicates better calibration, with 0 being a perfectly calibrated model.

Our analysis centers on the inherent tension between maximizing AUROC and minimizing ECE, a core challenge in confidence estimation. We also report a suite of supplementary metrics, including the Brier Score and F1 Score, to provide a more complete picture of performance. Detailed definitions for all metrics are provided in Appendix D.

4.3 Confidence Estimation Methods

Our investigation covers a wide spectrum of over fifteen distinct representation-based techniques, which we group into logical families to systematically explore different architectural hypotheses. These methods operate on one of two primary signal sources: the LRM's raw chunk hidden states, or correctness-focused features derived from a pre-trained probe (`PHSV-half`). We began by implementing established **baseline methods**, including verbalized confidence (`YVCE`), probing the initial prompt state (`P(IK)`), and the state-of-the-art chunk-level probing method (`PHSV`). To move beyond single-chunk analysis, we developed **Stacked Final Hidden State (`SFHS`)** models that process the entire stack of chunk hidden states using various backbones. In addition, we introduce **Token-Level Correctness Classification (`TLCC`)**, a sequential model that replaces chunk hidden states with low-dimensional statistics derived solely from token-level logit distributions, enabling a controlled evaluation of generative uncertainty signals without access to internal representations. To explicitly model the non-linear structure of reasoning, we benchmarked several **graph-based architectures**, including models that learn from the simple chronological flow (`GNN-SB`), the global logical and semantic coherence using rich edge features (`GNN-SR`), and the meta-level dynamics of the confidence trajectory itself (`GNN-CD`). Finally, we evaluated **hybrid and text-based models** that fuse different signal types (`LateFusion` and `CE`) or operate on the raw text of the entire reasoning trace (`ETTIN`). A comprehensive description of each method's architecture and implementation details is provided in Appendix C.

4.4 Training & Hyperparameter Optimization

To ensure a fair and rigorous comparison, all trainable models were developed using a consistent training and optimization protocol. Hyperparameter tuning was performed for each method using Optuna (Akiba et al., 2019), with each study running for up to 100 trials. For each LLM, we allocated 20% of the training portion as a validation subset, stratified by dataset to preserve the domain distribution within each LLM's training pool (see Tables 27 and 28 for the full train and test sizes).

The selection of an appropriate objective function for optimization was critical. Given the class imbalance inherent in our data—where a majority of LRM responses are correct—simply optimizing for accuracy would be a misleading metric. Our preliminary experiments also revealed that optimizing solely for discriminative power (AUROC) often produced models with very poor calibration, sacrificing ECE to maximize ranking performance. As our goal is to evaluate methods on their ability to perform well on *both* axes, we designed a composite score to guide the optimization process. The objective function was a composite score designed to jointly address discrimination and calibration:

$$\text{CompositeScore} = \alpha \cdot \text{AUROC} + (1 - \alpha) \cdot (1 - \text{ECE})$$

We set $\alpha = 0.6$ to place a slight emphasis on AUROC, prioritizing models with strong discriminative power while still imposing a significant penalty for poor calibration.

For final model selection, we imposed an additional practical constraint: a trial was only considered "feasible" if its best-performing epoch also achieved a minimum sensitivity and specificity of 0.50 at its Youden's J optimal threshold. This ensures our final reported models demonstrate a tangible predictive ability better than random chance on both positive and negative classes. Among all feasible trials, the one with the highest composite score was selected. To ensure a fair comparison of *architectural efficiency*, all model configurations were constrained to a maximum of 3.2 million trainable parameters. Each trial was trained for up to 200 epochs with an early-stopping patience of 20 epochs. The full details of our optimization strategy and the hyperparameter search space for each method are provided in Appendix E.

5 Results

We evaluated over ten distinct confidence estimation methods across six LRM families and six challenging test datasets. The overall performance, aggregated across all models and datasets, is presented in Table 2. While this table provides the

Table 2: Overall performance metrics for each method aggregated across all LLMs and test datasets. Each metric value represents a double-averaged result with standard deviation: first, each method's performance is averaged across all datasets for each LLM (unweighted mean ± std), then these LLM-specific means are averaged across all LLMs (unweighted mean ± std). **Bold** values indicate the overall best-performing method for each metric. The full table with detailed values is provided in Appendix Table 46.

Method	**ECE** ↓	**Brier** ↓	**Acc** ↑	**F1** ↑	**Spec** ↑	**AUCPR** ↑	**AUROC** ↑
YVCE	0.279 ±0.14	0.307 ±0.12	0.586 ±0.14	0.671 ±0.15	0.143 ±0.15	**0.696** ±0.10	0.603 ±0.06
TLCC-CONV	0.178 ±0.03	0.222 ±0.02	0.665 ±0.03	0.669 ±0.09	0.424 ±0.14	0.655 ±0.10	0.639 ±0.04
PHSV	0.197 ±0.05	0.251 ±0.03	0.609 ±0.05	0.600 ±0.12	0.389 ±0.20	0.577 ±0.11	0.598 ±0.04
SFHS-Conv	0.165 ±0.02	0.222 ±0.01	0.676 ±0.02	0.656 ±0.10	0.453 ±0.14	0.659 ±0.10	0.653 ±0.02
GNN-SB-GCN	0.150 ±0.02	0.216 ±0.01	0.678 ±0.02	0.645 ±0.10	0.432 ±0.14	0.657 ±0.10	0.653 ±0.03
GNN-SB-GraphSAGE	0.154 ±0.02	0.217 ±0.01	0.675 ±0.02	0.643 ±0.09	0.446 ±0.15	0.664 ±0.09	0.659 ±0.03
GNN-SR-Transformer	0.175 ±0.02	**0.208** ±0.01	0.677 ±0.02	0.648 ±0.12	0.432 ±0.19	0.664 ±0.09	0.656 ±0.03
GNN-CD-noft-GCN2Conv-dual	0.174 ±0.04	0.223 ±0.02	0.663 ±0.03	0.611 ±0.10	**0.523** ±0.14	0.657 ±0.11	0.651 ±0.02
ETTIN	0.160 ±0.01	0.217 ±0.01	0.677 ±0.02	**0.698** ±0.10	0.290 ±0.22	0.680 ±0.09	**0.672** ±0.04
ETTIN-HGA	**0.148** ±0.02	0.211 ±0.01	**0.689** ±0.01	0.693 ±0.12	0.282 ±0.23	0.678 ±0.10	0.670 ±0.03

highest-level summary of our findings, a comprehensive set of detailed results—including performance breakdowns per-LRM, per-dataset, and combinations thereof—is provided in Appendix F to facilitate a deeper, more granular analysis. This section reports the key empirical findings from our benchmark.

Discrimination Performance. The primary metrics for discrimination are AUROC, AUCPR, and F1 score. Across these, the text-based encoder methods, `ETTIN` and `ETTIN-HGA`, emerge as the top performers. `ETTIN` achieves the highest overall AUROC (0.672) and F1 Score (0.698), with `ETTIN-HGA` following closely (AUROC 0.670, F1 0.693). The best-performing hidden-state-based methods, such as `GNN-SB-GraphSAGE` (AUROC 0.659) and `GNN-SR-Transformer` (AUROC 0.656), form a competitive second tier but do not surpass the text-based approaches. In contrast, foundational baselines like `PHSV` (AUROC 0.598) and prompting-based methods like `YVCE` (AUROC 0.603) show significantly weaker discriminative ability. A surprising result is the strong performance of `YVCE` on AUCPR (0.696), suggesting that while poorly calibrated, it is effective at identifying some high-confidence correct answers.

Calibration Performance. The primary metrics for calibration are ECE and Brier score. Here, a different set of models excel. The best overall calibration is achieved by `ETTIN-HGA`, with an ECE of 0.148. This is closely followed by the simple graph models, specifically `GNN-SB-GCN` (0.150) and `GNN-SB-GraphSAGE` (0.154). Notably, the baseline `ETTIN` model, despite its top discriminative performance, has a comparatively weaker ECE of 0.160. The best Brier score, which balances discrimination and calibration, is achieved by `GNN-SR-Transformer` (0.208). The poorest calibration is observed in the `YVCE` baseline (ECE 0.279). A key observation across all methods is that no single model is dominant, with top performers in discrimination often being distinct from top performers in calibration.

Visualizing the Performance Landscape. Figure 1 provides a visual summary of these results, plotting each method's calibration (1-ECE) against its discrimination (AUROC). The center of each ellipse marks the method's mean performance, while its height and width represent the standard deviation of AUROC and ECE, respectively, indicating consistency across the benchmark. The plot shows that no single method occupies the ideal top-right corner. The text-based methods `ETTIN` and `ETTIN-HGA` are positioned furthest to the right, indicating the highest mean AUROC. The `GNN-SB` family models are clustered towards the top of the plot, indicating strong mean calibration. Many methods, such as the `SFHS` family and `GNN-SR-Transformer`, occupy a dense cluster in the center, representing a balance between the two metrics. Supplementary plots for other metric pairs are provided in Appendix G.

Visualizing Calibration Across LRMs. To examine how calibration varies across reasoning domains, Figure 2 presents reliability diagrams for each test dataset, with methods aggregated across all six LRM families. The ECE values in the legend represent dataset-specific averages across LRMs. Each subplot plots predicted confidence against empirical accuracy, with the dashed diagonal representing perfect calibration. The plots reveal substantial variation in calibration quality across both methods and datasets. Methods like `ETTIN-HGA` (avg. ECE=0.091) and `GNN-SR-Transformer` (avg.

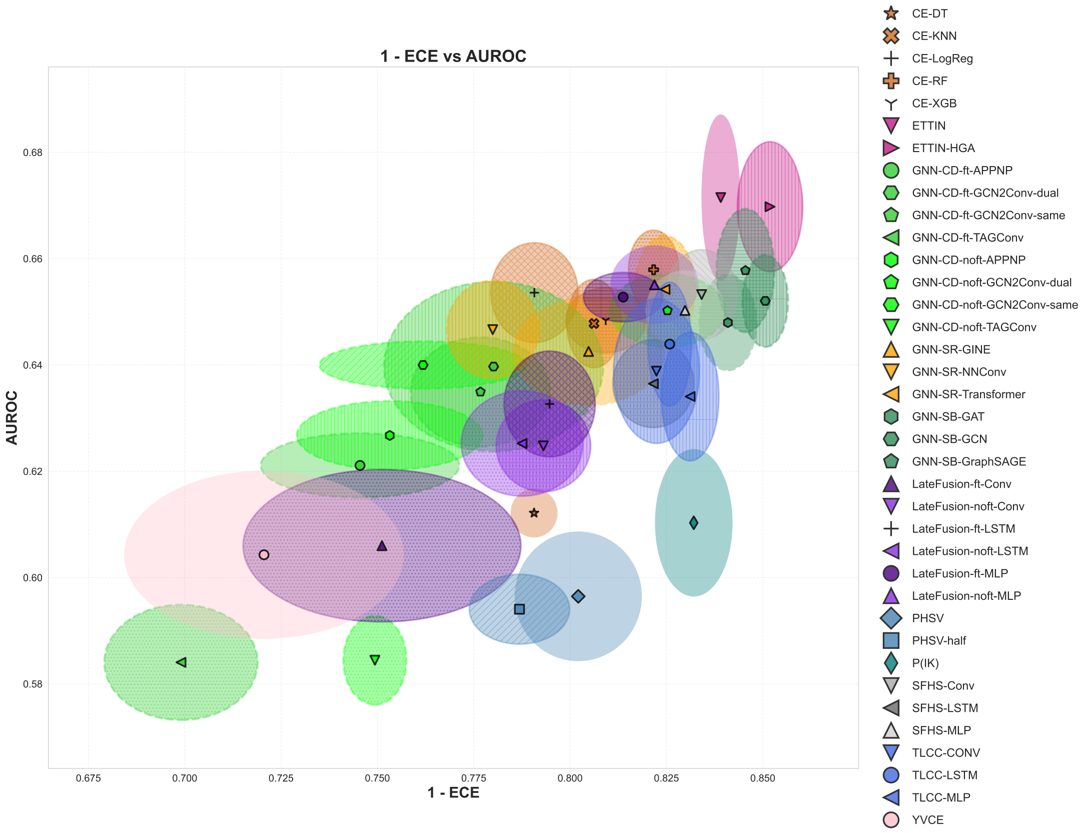

Figure 1: Performance trade-off between calibration (1-ECE) and discrimination (AUROC). Each ellipse denotes a method, with its center showing the unweighted doubly-averaged mean performance (first across datasets per LRM, then across LRMs). Ellipse width and height represent the standard deviation of these LRM-specific means, reflecting consistency across model architectures. The top-right corner marks ideal performance.

ECE=0.089) maintain consistently strong calibration with curves close to the diagonal across most datasets. Conversely, `YVCE` (avg. ECE=0.202) exhibits the poorest overall calibration. Notably, MedMCQA proves exceptionally challenging, with nearly all methods producing flat, poorly-aligned curves indicating severe miscalibration, while MMLU-Pro and BBH demonstrate better calibration potential with curves more closely following the diagonal. FinQA and MATH show moderate calibration with substantial method-dependent variation. For per-LRM calibration analysis and further interpretation details, see Appendix H.

6 Discussion

Our comprehensive benchmark reveals several key insights into the challenges of representation-based confidence estimation for LRMs. We discuss the main findings below.

A persistent trade-off exists between discrimination and calibration. The central finding of our study is a consistent trade-off between a model's ability to distinguish correct from incorrect answers (AUROC) and its ability to produce calibrated scores (ECE). This is evident in Table 2, where the top-performing model for AUROC, `ETTIN`, is not the best for ECE. This trade-off is visually confirmed in Figure 1, where no single method occupies the ideal top-right corner. Instead, top-performing methods are scattered along a performance frontier, suggesting a fundamental tension: architectures that capture holistic features for discrimination may be less effective at modeling the structural nuances crucial for calibration.

Text-based encoders excel at discrimination but require structural awareness for calibration. The strong performance of the `ETTIN` model on AUROC demonstrates that treating the entire reasoning trace as a single text input is a powerful strategy for discrimination. This approach likely captures global semantic and stylistic patterns that are indicative of a correct answer. However, the performance of `ETTIN-HGA` provides a crucial insight. The only difference between these models is the HGA layer (described in C.2.6), which explicitly models the chunk-level structure of the rea-

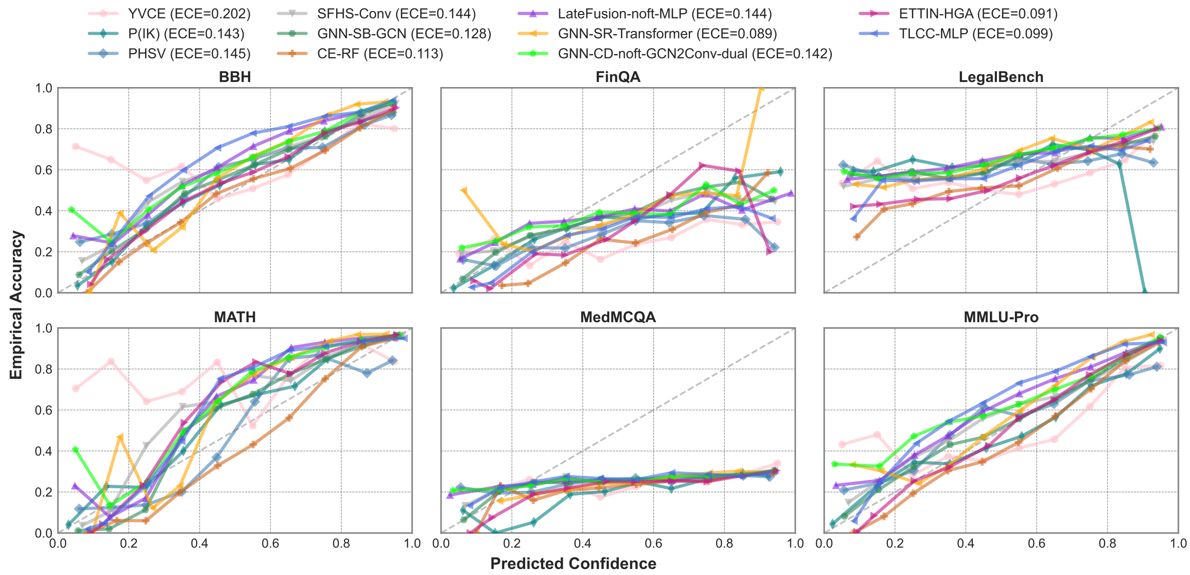

Figure 2: Calibration curves for each test dataset, aggregated across all LRMs. Each subplot shows one dataset's reliability diagram with methods averaged across all six LRM families. The ECE values in the legend represent each method's average performance across LRMs for that specific dataset. Points closer to the diagonal (dashed line) indicate better calibration.

soning. This single architectural change improves the ECE from 0.160 to 0.148 (a 7.5% relative improvement), establishing it as the best-calibrated model in our benchmark. The reliability diagrams (Figure 2) further illustrate its strong calibration, with `ETTIN-HGA` closely following the ideal diagonal line across multiple LRMs. This demonstrates that while holistic text representations are sufficient for ranking, achieving reliable calibration requires explicit awareness of the model's internal, step-by-step reasoning structure. However, even `ETTIN-HGA`'s strong calibration degrades substantially on certain datasets (particularly MedMCQA), suggesting that dataset characteristics significantly impact calibration quality beyond architectural choices.

Architectural complexity does not guarantee improved performance. A surprising result from our benchmark is that increased architectural complexity does not reliably lead to better performance. For example, the sophisticated `GNN-SR-Transformer` model, which processes a densely forward-connected graph with rich, 5-dimensional edge features, achieves a similar AUROC (0.656 vs. 0.653) and a worse ECE (0.175 vs. 0.165) compared to the much simpler `SFHS-Conv`, which applies a 1D convolution over a flat sequence of hidden states. This finding is reflected in Figure 1, where the ellipses for the more complex `GNN-SR` family are largely overlapping with, or even outperformed by, the simpler `SFHS` models. This pattern suggests that the primary limitation may not be the sophistication of the architecture, but rather the information content available in the chunk-level hidden states themselves.

Generative uncertainty signals are strong but insufficient for discrimination. The inclusion of `TLCC` provides a critical additional perspective on the observed performance ceiling. By relying exclusively on token-level logit statistics aggregated at the chunk level, `TLCC` removes hidden state representations entirely while retaining access to the model's intrinsic uncertainty signals. Across LRMs, `TLCC` achieves competitive calibration and, in several cases, improved specificity relative to hidden-state-based baselines, confirming that logit-derived statistics are effective indicators of uncertainty. However, `TLCC` consistently underperforms text-based encoders such as `ETTIN` in AUROC, indicating a reduced ability to separate correct from incorrect reasoning trajectories. This result reinforces a key conclusion of our benchmark: uncertainty signals alone, even when aggregated over full reasoning traces, are insufficient to capture the semantic and logical distinctions required for maximal discriminative power. High-performing confidence estimation for LRMs therefore appears to require both uncertainty-aware signals and access to global semantic context.

Two-stage feature extraction is a viable but inconsistent strategy. Several of our methods (`GNN-CD`, `LateFusion`, `CE`) use a two-stage setup where a `PHSV-half` model is trained first and then used as a feature extractor. This approach is reasonable but not always effective. The best two-stage model, `GNN-CD-noft-GCN2Conv-dual` (AUROC 0.651), performs comparably to the best single-stage hidden state models, yet many other `GNN-CD` variants perform poorly, as shown in the lower-left quadrant of Figure 1. End-to-end fine-tuning of `PHSV-half` features (`ft` variants) also failed to consistently improve results. This suggests that while local confidence signals help, integrating them into a global model remains challenging.

Simple baselines and direct confidence probing are unreliable. Our benchmark confirms that direct confidence estimation methods perform poorly on long-form reasoning tasks. `YVCE`, which relies on the LRM's own self-assessment, is the least calibrated model we tested. Similarly, `PHSV`, which uses a local classifier in an early-exit fashion, performs weakly on both AUROC and ECE. The poor calibration of these methods is evident in Figure 2. The reliability curve for `YVCE` is unstable, showing erratic fluctuations across the confidence spectrum. The curve for `PHSV`, though less volatile, remains misaligned across different LRM architectures. This finding shows that the problem is non-trivial and that more advanced modeling of the full reasoning trace is necessary for reliable confidence estimation.

7 Conclusion

In this work, we address the critical and under-explored challenge of confidence estimation for LRMs by introducing the Reasoning Model Confidence estimation Benchmark (RMCB), a publicly available resource of 347,496 reasoning traces from six popular LRMs across diverse high-stakes domains, each paired with correctness annotations. Using this benchmark, we conducted a comprehensive empirical evaluation of over ten representation-based methods, revealing a persistent trade-off between a method's discriminative power (AUROC) and its calibration (ECE). Our results show that text-based encoders like `ETTIN` achieve the best discrimination, while architectures that explicitly model the reasoning structure, such as `ETTIN-HGA` and simple graph-based models, yield superior calibration. Furthermore, we find that increased architectural complexity does not guarantee improved performance, with sophisticated graph neural networks failing to consistently outperform simpler sequential baselines. Ultimately, our work provides a clear map of the current performance landscape, rigorously quantifies the limitations of relying solely on chunk-level hidden states, and establishes a robust set of baselines that highlight the need for future research to explore alternative signal sources, such as token-level generative signals, to overcome the performance plateau we have identified.

Limitations

While our work provides a comprehensive benchmark for representation-based LRM confidence estimation, its scope is defined by several key methodological choices that present avenues for future research. Primarily, our investigation focuses on methods that operate on chunk-level representations, including both internal hidden states and aggregated token-level generative statistics. This design choice enables a computationally tractable and scalable feature set, avoiding the significant storage and processing overhead required to model full token-level distributions (e.g., complete logit vectors for every token in each of the 347,496 traces). While the inclusion of aggregated logit statistics addresses a simple and computationally efficient form of generative uncertainty, our results indicate that more expressive modeling of reasoning dynamics beyond chunk-level summaries remains an important direction for future work. Furthermore, our methodological scope includes several other boundaries. Our response segmentation into "chunks," while an enhancement of the method proposed by Zhang et al. (2025), remains a heuristic based on linguistic keywords. Our correctness annotations, while scalable, rely on an LLM judge (`GPT-5-nano`) rather than expert human annotation. We also excluded consistency-based methods from our benchmark due to the prohibitive inference cost of generating multiple long-form reasoning traces per query. Finally, the RMCB is currently limited to the English language and a specific suite of LRM architectures. We believe these limitations do not detract from our core findings but instead provide a clear and promising roadmap for future investigations into the reliability of LRMs.

Ethical Considerations

Although the goal of this work is to improve LRM reliability, several ethical considerations are important. First, a key risk is over-relying on any automated confidence score. Our results show that even the best methods have trade-offs and no method is perfect. In high-stakes fields like medicine, finance, or law, using these scores to automatically approve LRM outputs without human oversight could lead to harmful outcomes if a model error goes unnoticed. Second, fairness is a critical issue. The LRMs used in our benchmark may carry biases from their own training data, and the confidence models we trained could inherit or even amplify these issues. As a result, the confidence scores might be less reliable for certain demographic groups or types of questions, which could lead to unfair outcomes. Therefore, any method developed or evaluated on this benchmark should be treated as a tool to assist human experts, not to replace their critical judgment. We strongly recommend that any real-world deployment of these confidence estimators be preceded by thorough fairness testing and ongoing monitoring to ensure they are used responsibly.

Acknowledgments

This work was supported in part by the Henry Ford Health + Michigan State University Health Sciences Cancer Seed Funding Program and by the JPMorgan Chase AI Research Faculty Research Award. The authors are solely responsible for the contents of this paper; the opinions expressed do not necessarily reflect those of the funding organizations. The authors also acknowledge the use of Large Language Models to assist in polishing the language and grammar of this manuscript.

Disclaimer

This paper was prepared for informational purposes by the Artificial Intelligence Research group of JPMorgan Chase & Co and its affiliates ("JP Morgan"), and is not a product of the Research Department of JP Morgan. JP Morgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Mohammad Beigi, Ying Shen, Runing Yang, Zihao Lin, Qifan Wang, Ankith Mohan, Jianfeng He, Ming Jin, Chang-Tien Lu, and Lifu Huang. 2024. InternalInspector i^2: Robust confidence estimation in LLMs through internal states. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12847–12865, Miami, Florida, USA. Association for Computational Linguistics.

Arslan Chaudhry, Sridhar Thiagarajan, and Dilan Gorur. 2024. Finetuning language models to emit linguistic expressions of uncertainty. *Preprint*, arXiv:2409.12180.

Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. 2020. Simple and deep graph convolutional networks. *Preprint*, arXiv:2007.02133.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Jian Du, Shanghang Zhang, Guanhang Wu, Jose M. F. Moura, and Soummya Kar. 2018. Topology adaptive graph convolutional networks. *Preprint*, arXiv:1710.10370.

Yu Fan, Jingwei Ni, Jakob Merane, Etienne Salimbeni, Yang Tian, Yoan Hermstrüwer, Yinya Huang, Mubashara Akhtar, Florian Geering, Oliver Dreyer, Daniel Brunner, Markus Leippold, Mrinmaya Sachan, Alexander Stremitzer, Christoph Engel, Elliott Ash, and Joel Niklaus. 2025. Lexam: Benchmarking legal reasoning on 340 law exams. *Preprint*, arXiv:2505.12864.

Johannes Gasteiger, Aleksandar Bojchevski, and Stephan Günnemann. 2022. Predict then propagate: Graph neural networks meet personalized pagerank. *Preprint*, arXiv:1810.05997.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. 2017. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1263–1272. JMLR.org.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Preprint*, arXiv:2308.11462.

William L. Hamilton, Rex Ying, and Jure Leskovec. 2018. Inductive representation learning on large graphs. *Preprint*, arXiv:1706.02216.

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. *Preprint*, arXiv:2103.03874.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. 2020. Strategies for pre-training graph neural networks. *Preprint*, arXiv:1905.12265.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14).

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Arka Pal, Samuel Dooley, Micah Goldblum, and Andrew Wilson. 2024. Calibration-tuning: Teaching large language models to know what they don't know. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 1–14, St Julians, Malta. Association for Computational Linguistics.

Reza Khanmohammadi, Erfan Miahi, Mehrsa Mardikoraem, Simerjot Kaur, Ivan Brugere, Charese H. Smiley, Kundan Thind, and Mohammad M. Ghassemi. 2025. Calibrating llm confidence by probing perturbed representation stability. *Preprint*, arXiv:2505.21772.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. *Preprint*, arXiv:1609.02907.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Preprint*, arXiv:2205.14334.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *Preprint*, arXiv:2007.08124.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Yunsheng Shi, Zhengjie Huang, Shikun Feng, Hui Zhong, Wenjin Wang, and Yu Sun. 2021. Masked label prediction: Unified message passing model for semi-supervised classification. *Preprint*, arXiv:2009.03509.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *Preprint*, arXiv:2210.09261.

Oyvind Tafjord, Matt Gardner, Kevin Lin, and Peter Clark. 2019. QuaRTz: An open-domain dataset of qualitative relationship questions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5941–5946, Hong Kong, China. Association for Computational Linguistics.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2022. Commonsenseqa 2.0: Exposing the limits of ai through gamification. *Preprint*, arXiv:2201.05320.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. *Preprint*, arXiv:1710.10903.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Preprint*, arXiv:2406.01574.

Dongkeun Yoon, Seungone Kim, Sohee Yang, Sunkyoung Kim, Soyeon Kim, Yongil Kim, Eunbi Choi, Yireun Kim, and Minjoon Seo. 2025. Reasoning models better express their confidence. *Preprint*, arXiv:2505.14489.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. *Preprint*, arXiv:2002.04326.

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. *Preprint*, arXiv:2504.05419.

Xiaoling Zhou, Mingjie Zhang, Zhemg Lee, Wei Ye, and Shikun Zhang. 2025. Hademif: Hallucination detection and mitigation in large language models. In *The Thirteenth International Conference on Learning Representations*.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. *Preprint*, arXiv:2105.07624.

Fengbin Zhu, Ziyang Liu, Fuli Feng, Chao Wang, Moxin Li, and Tat Seng Chua. 2024. Tat-llm: A specialized language model for discrete reasoning over financial tabular and textual data. In *Proceedings of the 5th ACM International Conference on AI in Finance*, ICAIF '24, page 310–318, New York, NY, USA. Association for Computing Machinery.

Appendix

Table of Contents

A Reasoning Language Models **13**

B RMCB Benchmark Construction **13**
B.1 Data Curation and Standard Schema . 14
B.2 Dataset Sources and Versions 14
B.3 Training Datasets 14
B.3.1 GSM8K (Mathematical Reasoning) 15
B.3.2 TAT-QA (Financial Reasoning) . . . 15
B.3.3 MedQA (Medical Reasoning) . . . 17
B.3.4 LEXam (Legal Reasoning) 17
B.3.5 ARC-Challenge (General Science Reasoning) 17
B.3.6 CommonsenseQA 2.0 (General Reasoning) 18
B.3.7 LogiQA (General Reasoning) . . . 19
B.3.8 OpenBookQA (General Science Reasoning) 19
B.3.9 QuaRTz (General Science Reasoning) 20
B.3.10 ReClor (General Reasoning) 21
B.4 Evaluation Datasets 21
B.4.1 MATH (Mathematical Reasoning) . 21
B.4.2 FinQA (Financial Reasoning) . . . 22
B.4.3 MedMCQA (Medical Diagnosis) . . 23
B.4.4 LegalBench (Legal Reasoning) . . . 24
B.4.5 MMLU-PRO (General Understanding) 27
B.4.6 Big-Bench Hard (Complex Reasoning) 27
B.5 Response Generation 27
B.6 Response Segmentation 27
B.7 Response Grading 28
B.8 Example of a Fully Annotated and Segmented Reasoning Trace 34
B.9 Benchmark Availability and Licensing 34

C Confidence Estimation Methods **34**
C.1 Baseline Methods 35
C.1.1 YVCE (Verbalized Confidence Estimation) 35
C.1.2 TLCC (Token-Level Chunk Classification) 35
C.1.3 P(IK) (Probability of "I Know") . . 35
C.1.4 PHSV (Probing Hidden States for Self-Verification) 36
C.2 Novel Benchmarked Methods 36
C.2.1 SFHS (Stacked Final Hidden States) 36
C.2.2 GNNs (Graph Neural Networks) . . 36
C.2.3 CE (Chunk Ensemble) 38
C.2.4 LateFusion 38
C.2.5 ETTIN 38
C.2.6 ETTIN-HGA 38

D Evaluation Metrics **38**
D.1 Expected Calibration Error (ECE) . . 39
D.2 Brier Score 39
D.3 Accuracy (ACC) 39
D.4 F1 Score 39
D.5 Specificity 39
D.6 Area Under the Precision–Recall Curve (AUCPR) 39
D.7 Area Under the ROC Curve (AUROC) 40

E Training and Optimization Details **40**
E.1 Hyperparameter Optimization with Optuna 40
E.2 Hyperparameter Search Spaces 40

F Comprehensive Results **40**

G Supplementary Performance Visualization with Ellipse Plots **74**

H Detailed Calibration Analysis by Dataset **74**

A Reasoning Language Models

We evaluate a set of open-weight reasoning-oriented LLMs from multiple families, including Microsoft's Phi-4 series, Qwen's Qwen3 and QwQ lines, Mistral's Magistral, and LG AI Research's EXAONE. These models range in size from ∼3.85B to ∼32.8B parameters and span both dense and hybrid architectures, with context lengths from 32K up to 131K tokens. Table 3 and Table 4 summarize the key configuration details extracted directly from the model configuration files.

B RMCB Benchmark Construction

This section provides a detailed breakdown of the construction of the RMCB benchmark. We describe the full suite of datasets used to create our training and evaluation splits, the standardized schema for data curation, and the methodology for LRM response generation and annotation. All datasets are in English. For comprehensive information regarding the original construction and domain coverage of the source benchmarks, we refer readers to their respective publications.

Table 3: Architectural parameters of evaluated models.

Hub ID	Params (B)	Hidden Size	Layers	Attention Heads (Q/KV)
microsoft/Phi-4-mini-flash-reasoning	3.85	2560	32	40 / 20
Qwen/Qwen3-8B	8.19	4096	36	32 / 8
Qwen/Qwen3-14B	14.8	5120	40	40 / 8
mistralai/Magistral-Small-2506	23.6	5120	40	32 / 8
LGAI-EXAONE/EXAONE-Deep-32B	32.0	5120	64	40 / 8
Qwen/QwQ-32B	32.8	5120	64	40 / 8

Table 4: Context, feed-forward, and vocabulary parameters of evaluated models.

Hub ID	Feed-Forward Size	Context Length	Vocabulary Size
microsoft/Phi-4-mini-flash-reasoning	10240	64K	200,064
Qwen/Qwen3-8B	12288	32K	151,936
Qwen/Qwen3-14B	17408	32K	151,936
mistralai/Magistral-Small-2506	32768	128K	131,072
LGAI-EXAONE/EXAONE-Deep-32B	27392	32K	102,400
Qwen/QwQ-32B	27648	131K	152,064

B.1 Data Curation and Standard Schema

To create a consistent format for our experiments, all raw datasets were processed into a standardized JSONL format. Each line in the resulting files corresponds to a single reasoning problem and contains the following fields, which are derived from the properties available in each source dataset:

- `prompt`: The input question from the dataset, formatted into a string that is provided to the language model.

- `explanation`: Any relevant supporting information for the ground-truth answer provided by the source dataset. This can range from a detailed, step-by-step reasoning trace to shorter contextual details, or it may be empty if no such information is available.

- `answer`: The ground-truth final answer to the prompt, as specified by the source dataset.

- `category`: A categorization of the sample if it exists in the source dataset (e.g., "Surgery" for a sample in MedMCQA).

- `dataset`: The name of the source dataset.

- `record_id`: A unique and deterministic identifier for each sample, generated using a dataset-specific hashing function applied to the preprocessed input fields used during evaluation. The exact hashing procedures are provided in the dataset reconstruction script to ensure full reproducibility of identifiers.

This standardized schema allows for consistent data handling and feature extraction across all models and datasets in our benchmark.

B.2 Dataset Sources and Versions

Table 6 provides a complete overview of all datasets used to construct the RMCB benchmark, including their domains, splits, access methods, and exact versions or revisions. For datasets hosted on HuggingFace, we record the repository identifier, configuration where applicable, and commit hash to ensure reproducibility. For datasets requiring manual download, we explicitly note the source and provide step-by-step download and preprocessing instructions in the public RMCB code repository associated with this paper. This table is intended to make dataset provenance explicit and verifiable.

B.3 Training Datasets

The trainable confidence models are developed on a balanced training set of 10,000 samples aggregated from ten reasoning datasets, shown in Table 5. From each dataset, we deterministically sampled exactly 1,000 unique examples using `seed=23`, then merged them to form the final training set. These datasets span a variety of high-stakes and general reasoning domains—including mathematical, financial, medical, legal, and general reasoning—ensuring that the training data covers a wide spectrum of reasoning types. Below, we detail the processing for selected training datasets as representative examples of how raw benchmarks were mapped to our standard schema.

Table 5: Overall dataset distribution for training. Each dataset contributes exactly 1,000 unique examples (seed=23) to form a balanced 10,000-sample training set. Domains indicate the primary reasoning type assessed.

Dataset	**Domain**	**Train Samples**	**Train (%)**
GSM8K	Mathematical Reasoning	1,000	10.00
TAT-QA	Financial Reasoning	1,000	10.00
MedQA	Medical Reasoning	1,000	10.00
LEXam	Legal Reasoning	1,000	10.00
ARC	General Reasoning	1,000	10.00
CommonsenseQA2	General Reasoning	1,000	10.00
LogiQA	General Reasoning	1,000	10.00
OpenBookQA	General Reasoning	1,000	10.00
QuaRTz	General Reasoning	1,000	10.00
ReClor	General Reasoning	1,000	10.00

B.3.1 GSM8K (Mathematical Reasoning)

The Grade School Math 8K (GSM8K) dataset (Cobbe et al., 2021), released under the MIT License, is a benchmark designed to test multi-step mathematical reasoning. It consists of high-quality, linguistically diverse word problems whose solutions require a sequence of elementary calculations. We utilize the `socratic` configuration, which provides detailed, chain-of-thought style solutions for each problem. The raw dataset examples are mapped to our standard schema as follows:

- **`prompt`**: The input question is formatted using the following template:

  Prompt Template

  Question: {question}
  Answer:

- **`explanation`**: The full, unaltered reasoning trace from the original `answer` field of the dataset.

- **`answer`**: The final numerical answer is parsed from the text that follows the #### marker at the end of the reasoning trace.

- **`category`**: This field is not applicable to the GSM8K dataset and is set to "N/A".

B.3.2 TAT-QA (Financial Reasoning)

The Tabular and Textual Question Answering (TAT-QA) dataset (Zhu et al., 2021), released for non-commercial use, is a benchmark for financial reasoning over hybrid data. Each sample contains both unstructured text and a structured table extracted from real-world financial reports, requiring a model to synthesize information from both sources. Answering questions correctly often involves complex numerical reasoning, such as addition, subtraction, and comparison. The raw dataset examples are mapped to our standard schema as follows:

- **`prompt`**: A comprehensive prompt is constructed for each question. First, a context is created by verbalizing the structured table into natural language. This process follows the methodology of FinQA (Chen et al., 2021), where each row is converted into a descriptive sentence using a template similar to *'the {column name} of {row name} is {cell value};'*. This verbalized table text is then concatenated with the original paragraphs to form a complete context. This context is then embedded into a template akin to Zhu et al. (2024), which is formatted as follows:

  Prompt Template

  Below is an instruction that describes a question answering task in the finance domain, paired with an input table and its relevant text that provide further context. Generate an appropriate answer to the given question.
  Question: {question}
  Context: {paragraphs and verbalized table}
  Answer:

- **`explanation`**: This field is populated with the `derivation` from the source dataset, which provides a step-by-step reasoning trace for the answer.

- **`answer`**: The ground-truth `answer` from the source dataset is normalized into a consistent string format.

- **`category`**: This field is a composite created by joining the `answer_type` and `answer_from` fields from the source data (e.g., "multi-span[SEP]table-text"). Distributions of `answer_type` and `answer_from` are provided in Tables 7 and 8, respectively.

Table 6: Dataset sources and exact revisions or versions used to construct the RMCB benchmark.

Dataset	**Details**
GSM8K	Source: HuggingFace `openai/gsm8k` (config: `socratic`) Revision: `cc7b047b6e5bb11b4f1af84efc572db110a51b3c`
TAT-QA	Source: GitHub `NExTplusplus/TAT-QA` Revision: not versioned
MedQA	Source: Google Drive release Revision: `ddef95d268cdad413693d634279a9a679d468469`
LEXam	Source: HuggingFace `LEXam-Benchmark/LEXam` Revision: `68f21a324eb0e14837be42f10b644c40847c3ed4`
ARC	Source: HuggingFace `allenai/ai2_arc` (config: `ARC-Challenge`) Revision: `210d026faf9955653af8916fad021475a3f00453`
CommonsenseQA-2	Source: HuggingFace `chiayewken/commonsense-qa-2` Revision: `15e7dc364f7906ad69cbe4a0bed697ba12f07bdf`
LogiQA	Source: HuggingFace `lucasmccabe/logiqa` Revision: `3c19b0488d794d30c36f73d132d8a22e64f42f2e`
OpenBookQA	Source: HuggingFace `allenai/openbookqa` (config: `main`) Revision: `388097ea7776314e93a529163e0fea805b8a6454`
QuaRTz	Source: HuggingFace `allenai/quartz` Revision: `28c1dbb56caf81799296cb17892fa73402e23464`
ReClor	Source: HuggingFace `voidful/ReClor` Revision: `809ebe44b8dde882c4190f4178b27676b941b933`
MATH	Source: Kaggle `awsaf49/math-dataset` Version: 1
FinQA	Source: GitHub `czyssrs/FinQA` Revision: `0f16e2867befa6840783e58be38c9efb9229d742`
MedMCQA	Source: HuggingFace `openlifescienceai/medmcqa` Revision: `91c6572c454088bf71b679ad90aa8dffcd0d5868`
LegalBench	Source: HuggingFace `nguha/legalbench` Revision: `e042ea68c19df12b737fe768572f22ead61e8e37`
MMLU-Pro	Source: HuggingFace `TIGER-Lab/MMLU-Pro` Revision: `dd36ce4b34827164989f100331f82c5a29741747`
BBH	Source: HuggingFace `maveriq/bigbenchhard` Revision: `d53c5b10a77edeb29da195f47e6086b29f2f7f74`

Table 7: TAT-QA answer types distribution (Train).

Answer Type	Train Samples	Train (%)
span	438	43.8
arithmetic	414	41.4
multi-span	121	12.1
count	27	2.7

Table 8: TAT-QA answer sources distribution (Train).

Answer Source	Train Samples	Train (%)
table	448	44.8
table-text	334	33.4
text	218	21.8

B.3.3 MedQA (Medical Reasoning)

The MedQA dataset (Jin et al., 2021), released under a research-use-only license, is a large-scale, multiple-choice question benchmark designed to test professional medical knowledge, with questions sourced from medical board exams in the US, Mainland China, and Taiwan. The questions are varied and often require a deep understanding of clinical scenarios to arrive at the correct diagnosis or treatment. The raw dataset examples are mapped to our standard schema as follows:

- `prompt`: Each example is formatted into a standardized multiple-choice prompt. To ensure consistency, the options are sorted alphabetically by their key (e.g., A, B, C) before being inserted into the following template:

  Prompt Template

  Question: question
  Choices:
  (A) {option_A}
  (B) {option_B}
  ...
  Answer:

- `explanation`: This field is intentionally left empty as the source dataset does not provide reasoning traces for the answers.

- `answer`: The ground-truth `answer` is the single letter corresponding to the correct option (e.g., "B").

- `category`: This field is populated with the `meta_info` from the source dataset, which typically specifies the medical sub-domain (e.g., "Internal Medicine"). In addition, the dataset is categorized by USMLE exam step (Step 1, Step 2 & 3), as shown in Table 9.

Table 9: MedQA categories distribution (Train).

Category	Train Samples	Train (%)
Step 1	552	55.2
Step 2 & 3	448	44.8

B.3.4 LEXam (Legal Reasoning)

The LEXam benchmark (Fan et al., 2025), released under the CC BY 4.0 License, is a collection of legal examination questions designed to test complex legal reasoning across various jurisdictions and legal areas, sourced from 340 real law exams. We process all English-language samples from three distinct variants: `open_question`, `mcq_4_choices`, and `mcq_perturbation`. The raw dataset examples are mapped to our standard schema as follows:

- `prompt`: The prompts are tailored to the question type.

  For the `open_question` variant, a detailed system prompt guides the model to provide a structured, exam-style legal analysis (see Figure 3).

  For the `mcq` variants, a different prompt instructs the model to use a step-by-step, chain-of-thought process to analyze the facts, explain relevant legal rules, and justify its final choice (see Figure 4). Both prompt templates above are reproduced directly from the main manuscript of the LEXam benchmark (Fan et al., 2025).

- `explanation`: This field is intentionally left empty as the source dataset provides reference answers but not step-by-step reasoning traces.

- `answer`: This field contains the full text of the ground-truth answer for open questions and the single correct letter for MCQ variants.

- `category`: This field is a composite created by joining the `area`, `jurisdiction`, and `course` fields from the source data (e.g., "Private[SEP]International[SEP]Comparative Private Law"). Distributions for each of these components are shown in Tables 10, 11, and 12.

B.3.5 ARC-Challenge (General Science Reasoning)

The AI2 Reasoning Challenge (ARC) dataset (Clark et al., 2018), released under the CC BY-

Open Question Prompt Template

You are an expert in {course_name} and address legal issues in a structured, exam-style manner. Assume the applicable jurisdiction unless specifically mentioned; if the course context justifies, address legal issues beyond the stated jurisdiction as well.
Use precise legal language and formal academic tone when answering.
Do NOT state any disclaimer or refer to the need for external legal advice.
Do NOT request the user to consult laws or to research on their own.
Offer focused legal analyses and individualized advice.
Speak directly and authoritatively without mentioning that your response is merely for general information.
Incorporate jurisdiction-specific legal terminology where appropriate.
If you have discovered relevant legal considerations, respond with a concise, clear legal analysis.
Cite the specific legal provision, explicitly indicating sections, subsections, or paragraphs where available (e.g., "Section 74(2)(b) of the Contracts Act").
Avoid vague references without specifying applicable subsections or clauses.
If no relevant considerations are found, explicitly state that no pertinent information is available.
If you do have reliable sources, share practical guidance or insights from them.
Respond in the same language as the question.
If the question specifically requests a short answer, provide a concise response.
If the prompt asks you to analyze a specific case provided in the exam, but the text or details of that case have not been provided in the prompt, explicitly flag that the required case material is missing.
Question:
{question}
Answer:

Figure 3: Open Question Prompt Template for the `open_question` variant.

Table 10: LEXam area distribution (Train).

Area	Train Samples	Train (%)
Interdisciplinary	493	49.3
Private	249	24.9
Public	122	12.2
Unknown	118	11.8
Criminal	18	1.8

Table 11: LEXam jurisdiction distribution (Train).

Jurisdiction	Train Samples	Train (%)
Swiss	471	47.1
International	366	36.6
Unknown	118	11.8
Generic	45	4.5

SA 4.0 License, is a collection of multiple-choice science questions designed to be challenging for AI systems, requiring knowledge and reasoning beyond simple retrieval. We use the more difficult `ARC-Challenge` subset and process examples exclusively from its `train` split. The raw dataset examples are mapped to our standard schema as follows:

- **`prompt`**: Each question is formatted into a standard multiple-choice prompt, with the choices presented in their original order from the dataset. The template is as follows:

Prompt Template

Question: {question}
Choices:
(A) {option_A}
(B) {option_B}
...
Answer:

- **`explanation`**: This field is intentionally left empty as the source dataset does not provide reasoning traces.
- **`answer`**: The ground-truth `answer` is the correct option letter from the original `answerKey` field.
- **`category`**: This field is derived from the prefix of the question's `id` field (e.g., "Mercury", "MCAS"). The category distribution is shown in Table 13.

B.3.6 CommonsenseQA 2.0 (General Reasoning)

The CommonsenseQA 2.0 dataset (Talmor et al., 2022), released under the CC BY 4.0 License, is a challenging benchmark of 14,343 yes/no questions designed to test a model's commonsense reasoning capabilities. The questions were created through a gamified process to be adversarial to language models. The raw dataset examples are mapped to our standard schema as follows:

MCQ Prompt Template

You are an expert in {course_name} and address legal issues in a structured, exam-style manner. You are given a multiple-choice question, where only one choice (e.g., A, B, C, etc.) is correct. Assume the applicable jurisdiction unless specifically stated otherwise. If the context of the course justifies it, consider legal frameworks beyond the stated jurisdiction as well.
Please reason through the question step by step, using a chain-of-thought approach:
- Clarify the facts: Briefly restate or highlight the key facts in the question to anchor your reasoning.
- Issue Identification: What legal issue(s) arise from the facts?
- Rule Explanation: What legal rules or principles are relevant, and what are their sources (e.g., statutes, case law, doctrine)?
- Application and Reasoning: Apply the relevant rules to the facts, carefully weighing any ambiguities, exceptions, or competing interpretations.
- Eliminate Incorrect Answers: Briefly explain why each incorrect answer is wrong or less convincing.
- Conclusion: Clearly state the correct answer choice (e.g., A, B, C, etc.) with a brief justification for why it best fits the legal analysis.
Format your final answer as follows:
Correct Answer: C
Question:
{question_with_choices}
Answer:

Figure 4: Multiple-choice (MCQ) Prompt Template for the `mcq` variants.

- **`prompt`**: The input question is formatted into a standardized yes/no multiple-choice prompt using the following template:

  Prompt Template

  Question: {question}
  Choices:
  (A) yes
  (B) no
  Answer:

- **`explanation`**: This field is populated with the `relational_prompt` from the source dataset (e.g., "is capable of", "causes"), which indicates the type of commonsense reasoning being tested.

- **`answer`**: The ground-truth `answer` is converted to a letter, where "yes" maps to "A" and "no" maps to "B".

- **`category`**: This field is a composite created by joining the `topic_prompt` and boolean flags indicating whether the topic and relational prompts were used in the question's construction (e.g., "world trade center[SEP]True[SEP]True"). The distribution of prompt usage is provided in Table 14.

B.3.7 LogiQA (General Reasoning)

The LogiQA dataset (Liu et al., 2020), released under the CC BY-NC-SA 4.0 License, is a benchmark designed to test a model's capability for logical reasoning within a machine reading comprehension context. It is sourced from expert-written questions for civil servant exams and covers multiple types of deductive reasoning, such as categorical, conditional, and disjunctive reasoning. A key challenge of this dataset is that the correct answer is typically not a direct span in the text but must be inferred through logical steps. The raw dataset examples are mapped to our standard schema as follows:

- **`prompt`**: The input is formatted as a multiple-choice question that includes a context paragraph, the query, and four options. The template is as follows:

  Prompt Template

  Context: {context}
  Question: {query}
  Choices:
  (A) {option_A}
  (B) {option_B}
  (C) {option_C}
  (D) {option_D}
  Answer:

- **`explanation`**: This field is intentionally left empty as the source dataset does not provide explicit reasoning traces.

- **`answer`**: The ground-truth `answer` is the single letter corresponding to the correct option.

- **`category`**: This field is not applicable to the LogiQA dataset and is set to "N/A".

B.3.8 OpenBookQA (General Science Reasoning)

The OpenBookQA dataset (Mihaylov et al., 2018), released under the Apache License 2.0, is a

Table 12: LEXam course distribution (Train).

Course	Train Samples	Train (%)
Swiss Law	582	58.2
US Business Law	96	9.6
International Organisations	34	3.4
Chinese Business Law	29	2.9
International Finance Law	23	2.3
International Commercial Arbitration	23	2.3
Comparative Private Law	20	2.0
European Economic Law	20	2.0
Legal Theory	20	2.0
International Sales Law	18	1.8
Foundations and Trusts	18	1.8
International Criminal Law	18	1.8
History of Business Law	16	1.6
Transnational Public Security Law	15	1.5
Legal Sociology	9	0.9
Principles of Corporate Law	9	0.9
International Financial Law	8	0.8
International Human Rights	8	0.8
Introduction to Sports Law	7	0.7
International Economic Law	7	0.7
Comparative Constitutional Law	7	0.7
Gesellschaftsrecht	7	0.7
Comparative Corporate Law	6	0.6

multiple-choice question answering benchmark designed to test a deeper understanding of elementary-level science. The questions require combining a core science fact from a provided "open book" with broad common knowledge to arrive at the correct answer. We process examples from the `main` configuration's `train` split. The raw dataset examples are mapped to our standard schema as follows:

- **`prompt`**: Each question is formatted into a standard multiple-choice prompt, with the choices presented in their original order from the dataset. The template is as follows:

  Prompt Template

  Question: {question}
  Choices:
  (A) {option_A}
  (B) {option_B}
  ...
  Answer:

- **`explanation`**: This field is intentionally left empty as the source dataset does not provide reasoning traces.

- **`answer`**: The ground-truth `answer` is the correct option letter from the original `answerKey` field.

- **`category`**: This field is not applicable to the OpenBookQA dataset and is set to "N/A".

B.3.9 QuaRTz (General Science Reasoning)

The QuaRTz dataset (Tafjord et al., 2019), released under the CC BY 4.0 License, is a benchmark for reasoning about textual qualitative relationships. Each question is a 2-way multiple-choice problem that is paired with a background sentence expressing a general qualitative relationship (e.g., "More pollutants mean poorer air quality."). Answering correctly requires applying this general knowledge to a specific, novel situation presented in the question. The raw dataset examples are mapped to our standard schema as follows:

- **`prompt`**: The input is formatted as a multiple-choice question that includes a context paragraph, the question itself, and two answer choices. The template is as follows:

  Prompt Template

  Context: {context}
  Question: {question}
  Choices:
  (A) {option_A}
  (B) {option_B}
  ...
  Answer:

- **`explanation`**: This field is intentionally left empty as the source dataset does not provide reasoning traces.

- **`answer`**: The ground-truth `answer` is the correct option letter from the original `answerKey`

Table 13: ARC-Challenge categories distribution (Train).

Category	Train Samples	Train (%)
Mercury	647	64.7
MCAS	89	8.9
NYSEDREGENTS	38	3.8
ACTAAP	36	3.6
MDSA	31	3.1
TIMSS	24	2.4
NCEOGA	23	2.3
AKDE&ED	20	2.0
VASoL	16	1.6
MEA	15	1.5
LEAP	13	1.3
MSA	12	1.2
CSZ	7	0.7
AIMS	5	0.5
TAKS	4	0.4
MEAP	4	0.4
NAEP	4	0.4
OHAT	3	0.3
CSZ30494	1	0.1
CSZ30564	1	0.1
NCEOGA2013	1	0.1
WASL	1	0.1
CSZ10245	1	0.1
CSZ20740	1	0.1
CSZ30771	1	0.1
CSZ20059	1	0.1
FCAT	1	0.1

field.

- `category`: This field is not applicable to the QuaRTz dataset and is set to "N/A".

B.3.10 ReClor (General Reasoning)

The ReClor dataset (Yu et al., 2020), released under the CC BY-NC 4.0 License, is a reading comprehension benchmark that requires complex logical reasoning. The questions are sourced from standardized tests such as the GMAT and LSAT, ensuring a high level of difficulty and quality. The dataset is designed to evaluate a model's ability to understand arguments, identify flaws, and make inferences, rather than simple text matching. The raw dataset examples are mapped to our standard schema as follows:

- `prompt`: The input is formatted as a multiple-choice question that includes a context paragraph, the question itself, and four answer choices. The template is as follows:

Prompt Template

Context: {context}
Question: {question}
Choices:
(A) {option_A}
(B) {option_B}
...
Answer:

- `explanation`: This field is intentionally left empty as the source dataset does not provide explicit reasoning traces.
- `answer`: The ground-truth `answer` is the single letter corresponding to the correct option.
- `category`: This field is not applicable to the ReClor dataset and is set to "N/A".

B.4 Evaluation Datasets

The final evaluation of the confidence estimation models is performed on a challenging suite of 51,951 samples aggregated from six reasoning-intensive datasets. These datasets, shown in Table 15, are entirely disjoint from the training set and are used exclusively for testing. The evaluation suite was selected to probe model generalization in both in-domain and out-of-domain settings. While none of the datasets overlap with those used for training, they span the same high-level reasoning domains—mathematical, financial, medical, legal, and complex reasoning—ensuring that evaluation reflects the model's ability to transfer knowledge. Below, we detail the processing for these evaluation datasets, mapping them to our standard schema.

B.4.1 MATH (Mathematical Reasoning)

The MATH dataset (Hendrycks et al., 2021), released under the MIT License, is a benchmark designed to test mathematical problem-solving ability, consisting of 12,500 problems from high school mathematics competitions. We utilize the full test set of 5,000 records for our evaluation. The problems cover seven subjects, including algebra, geometry, and number theory, and each is accompanied by a full, step-by-step solution in LaTeX. The raw dataset examples are mapped to our standard schema as follows:

- `prompt`: The input question is formatted using the following template:

Prompt Template

Question: {problem}
Answer:

Table 14: CommonsenseQA2 prompt usage distribution (Train).

relational_prompt_used	topic_prompt_used	Train Samples	Train (%)
False	False	21	2.1
False	True	123	12.3
True	False	7	0.7
True	True	849	84.9

Table 15: Overall dataset distribution for testing. Each dataset is entirely disjoint from the training set and used exclusively for evaluation. Domains indicate the primary reasoning type assessed.

Dataset	Domain	Test Samples	Test (%)
MATH	Mathematical Reasoning	5,000	9.62
FinQA	Financial Analysis	1,138	2.19
MedMCQA	Medical Diagnosis	6,150	11.84
LegalBench	Legal Reasoning	21,167	40.74
MMLU-Pro	General Understanding	11,987	23.07
BBH	Complex Reasoning	6,509	12.53

- **`explanation`**: This field is populated with the full, unaltered step-by-step solution from the original `solution` field of the dataset.

- **`answer`**: The ground-truth `answer` is the final result extracted from the `\boxed{. . . }` command within the solution text.

- **`category`**: This field is a composite created by joining the problem `type` and `level` from the source data (e.g., "Prealgebra[SEP]Level 4"). Distributions by type, level, and their combination are reported in Tables 16, 17, and 18, respectively.

Table 16: MATH problem types distribution (Test).

Type	Test Samples	Test (%)
Algebra	1187	23.74
Intermediate Algebra	903	18.06
Prealgebra	871	17.42
Precalculus	546	10.92
Number Theory	540	10.80
Geometry	479	9.58
Counting & Probability	474	9.48

Table 17: MATH problem levels distribution (Test).

Level	Test Samples	Test (%)
Level 5	1324	26.48
Level 4	1214	24.28
Level 3	1131	22.62
Level 2	894	17.88
Level 1	437	8.74

B.4.2 FinQA (Financial Reasoning)

The FinQA dataset (Chen et al., 2021), released under the MIT License, is a large-scale benchmark designed to test numerical reasoning over financial reports. We utilize the full test set of 1,147 records for our evaluation. Each example consists of a question authored by a financial expert, along with both unstructured text and a structured table from S&P 500 company earnings reports. Answering questions correctly requires synthesizing information from both data types and performing complex, multi-step calculations. The raw dataset examples are mapped to our standard schema as follows:

- **`prompt`**: A comprehensive prompt is constructed for each question. First, a context is created by verbalizing the structured table into natural language. This process follows the methodology of the original FinQA paper, where each row is converted into a descriptive sentence using a template similar to *'the {column name} of {row name} is {cell value};'*. This verbalized table text is then concatenated with the original paragraphs to form a complete context, which is then embedded into the same prompt template used for the TAT-QA dataset.

- **`explanation`**: This field is populated with the executable reasoning program (e.g., "divide(100, 100), divide(3.8, #0)") from the source dataset, which provides a fully explainable, step-by-step reasoning trace.

- **`answer`**: The ground-truth `answer` is the final numerical result from the source dataset.

Table 18: MATH combined type–level distribution (Test).

Type	Level	Test Samples	Test (%)
Algebra	Level 5	307	6.14
Algebra	Level 4	283	5.66
Intermediate Algebra	Level 5	280	5.60
Algebra	Level 3	261	5.22
Intermediate Algebra	Level 4	248	4.96
Prealgebra	Level 3	224	4.48
Algebra	Level 2	201	4.02
Intermediate Algebra	Level 3	195	3.90
Prealgebra	Level 5	193	3.86
Prealgebra	Level 4	191	3.82
Prealgebra	Level 2	177	3.54
Number Theory	Level 5	154	3.08
Number Theory	Level 4	142	2.84
Algebra	Level 1	135	2.70
Precalculus	Level 5	135	2.70
Geometry	Level 5	132	2.64
Intermediate Algebra	Level 2	128	2.56
Precalculus	Level 3	127	2.54
Geometry	Level 4	125	2.50
Counting & Probability	Level 5	123	2.46
Number Theory	Level 3	122	2.44
Precalculus	Level 4	114	2.28
Precalculus	Level 2	113	2.26
Counting & Probability	Level 4	111	2.22
Geometry	Level 3	102	2.04
Counting & Probability	Level 2	101	2.02
Counting & Probability	Level 3	100	2.00
Number Theory	Level 2	92	1.84
Prealgebra	Level 1	86	1.72
Geometry	Level 2	82	1.64
Precalculus	Level 1	57	1.14
Intermediate Algebra	Level 1	52	1.04
Counting & Probability	Level 1	39	0.78
Geometry	Level 1	38	0.76
Number Theory	Level 1	30	0.60

- **`category`**: This field indicates whether the necessary information to answer the question was found in the text, the table, or both (e.g., "text_retrieved-table_retrieved"). The category distribution is shown in Table 19.

Table 19: FinQA categories distribution (Test).

Category	Test Samples	Test (%)
table_retrieved	657	57.73
text_retrieved	272	23.90
text_retrieved-table_retrieved	125	10.98
N/A	84	7.38

B.4.3 MedMCQA (Medical Diagnosis)

The MedMCQA dataset (Pal et al., 2022), released under the Apache License 2.0, is a large-scale, multiple-choice question answering benchmark containing over 194,000 questions from Indian medical entrance exams (AIIMS & NEET PG). The dataset is designed to test deep medical knowledge and reasoning across 21 different subjects. We utilize the full test set of 6,150 records for our evaluation. The raw dataset examples are mapped to our standard schema as follows:

- **`prompt`**: Each example is formatted into a standardized multiple-choice prompt. The four options are presented in their original order. The template is as follows:

Prompt Template

Question: {question}
Choices:
(A) {option_A}
(B) {option_B}
(C) {option_C}
(D) {option_D}
Answer:

- **`explanation`**: This field is populated with the expert's explanation for the correct answer, taken directly from the `exp` field of the source dataset.

Table 20: MedMCQA categories distribution (Test).

Category	Test Samples	Test (%)
Dental	1203	19.56
Unknown	682	11.09
Gynaecology & Obstetrics	532	8.65
Surgery	501	8.15
Physiology	388	6.31
Medicine	372	6.05
Biochemistry	352	5.72
Pharmacology	317	5.15
Pathology	305	4.96
Anatomy	259	4.21
Social & Preventive Medicine	243	3.95
Pediatrics	190	3.09
Ophthalmology	177	2.88
Microbiology	167	2.72
Forensic Medicine	132	2.15
Radiology	119	1.93
ENT	86	1.40
Skin	60	0.98
Anaesthesia	59	0.96
Psychiatry	6	0.10

- **`answer`**: The ground-truth `answer` is the single letter corresponding to the correct option, derived from the `cop` (correct option) field.
- **`category`**: This field is populated with the `subject_name` from the source data (e.g., "Pathology", "Anatomy"). The category distribution is shown in Table 20.

B.4.4 LegalBench (Legal Reasoning)

The LegalBench dataset (Guha et al., 2023) is a comprehensive benchmark consisting of 162 tasks designed to evaluate legal reasoning. Given the scale of the benchmark, we curated a representative subset from its test data for a more controlled evaluation. To ensure a balanced and fair selection, we implemented a deterministic, multi-stage sampling algorithm (`seed=23`). First, tasks were stratified into the six core legal reasoning types defined by the original authors (e.g., Issue-spotting, Rule-application, Interpretation). Then, from each reasoning category, we randomly sampled a quarter of the available tasks, up to a maximum of five, to form our evaluation suite. This process resulted in a final test set of 15 distinct tasks, comprising 21,167 test records.

The raw dataset examples are mapped to our standard schema as follows:

- **`prompt`**: Constructed using templates adapted from the official implementation provided for each task in LegalBench. The templates for the selected tasks are summarized in Table 21.
- **`explanation`**: Intentionally left empty, as the source dataset provides final answers but not step-by-step reasoning traces.
- **`answer`**: The ground-truth `answer` is taken directly from the source data for each task.
- **`category`**: Populated with the task identifier from the LegalBench dataset (e.g., "international_citizenship_questions", "opp115_first_party_collection_use"). The category distribution is shown in Table 22.

Table 21: Prompt templates for the selected LegalBench tasks used in our evaluation suite.

Task / Prompt	Records	License
- International citizenship questions Prompt: Answer the following questions considering the state of international law on January 1st, 2020. Question: {{question}} Answer "Yes" or "No". Answer:	9,306	CC BY 4.0
- Learned hands housing Prompt: Does the post discuss issues with paying your rent or mortgage, landlord-tenant issues, housing subsidies and public housing, eviction, and other problems with your apartment, mobile home, or house? Answer "Yes" or "no". Post: {text} Label:	4,494	CC BY-NC-SA 4.0
- Opp115 first party collection use Prompt: Does the clause describe how and why a service provider collects user information? Answer "Yes" or "no". Clause: {text} Label:	2,086	CC BY-NC
- Cuad license grant Prompt: Does the clause contain a license granted by one party to its counterparty? Answer "Yes" or "no". Clause: {text} Label:	1,396	CC BY 4.0
- Definition classification Prompt: Identify if the sentence defines a term. Answer "Yes" or "no". Sentence: {text} Label:	1,337	CC BY-SA 4.0
- Opp115 international and specific audiences Prompt: Does the clause describe practices that pertain only to a specific group of users (e.g., children, Europeans, or California residents)? Answer "Yes" or "no". Clause: {text} Label:	980	CC BY-NC
- Learned hands torts Prompt: Does the post discuss problems that one person has with another person (or animal), like when there is a car accident, a dog bite, bullying or possible harassment, or neighbors treating each other badly? Answer "Yes" or "no". Post: {text} Label:	432	CC BY-NC-SA 4.0
- Diversity 5 Prompt: Diversity jurisdiction exists when there is (1) complete diversity between plaintiffs and defendants, and (2) the amount-in-controversy (AiC) is greater than $75k. Q: {text} Is there diversity jurisdiction? A:	300	CC BY 4.0
- Diversity 6 Prompt: Diversity jurisdiction exists when there is (1) complete diversity between plaintiffs and defendants, and (2) the amount-in-controversy (AiC) is greater than $75k. Q: {text} Is there diversity jurisdiction? A:	300	CC BY 4.0
- Learned hands domestic violence Prompt: Does the post discuss dealing with domestic violence and abuse, including getting protective orders, enforcing them, understanding abuse, reporting abuse, and getting resources and status if there is abuse? Answer "Yes" or "no". Post: {text} Label:	174	CC BY-NC-SA 4.0

(Continued on next page)

(Continued from previous page)

Task / Prompt	Records	License
- UCC v common law Prompt: The UCC (through Article 2) governs the sale of goods, which are defined as moveable tangible things (cars, apples, books, etc.), whereas the common law governs contracts for real estate and services. For the following contracts, determine if they are governed by the UCC or by common law. Contract: {contract} Is this contract governed by the UCC or the common law? Governed by:	94	CC BY 4.0
- Maud cor standard (intervening event) Prompt: Instruction: Read the segment of a merger agreement and answer the multiple-choice question by choosing the option that best characterizes the agreement. Question: What standard should the board follow when determining whether to change its recommendation in response to an intervening event? Option A: “Breach” of fiduciary duties ... Option I: Other specified standard Merger Agreement: {text} Answer:	84	CC BY 4.0
- Cuad third party beneficiary Prompt: Does the clause specify that there is a non-contracting party who is a beneficiary to some or all of the clauses in the contract and therefore can enforce its rights against a contracting party? Answer “Yes” or “no”. Clause: {text} Label:	68	CC BY 4.0
- Learned hands benefits Prompt: Does the post discuss public benefits and social services that people can get from the government, like for food, disability, old age, housing, medical help, unemployment, child care, or other social needs? Answer “Yes” or “no”. Post: {text} Label:	66	CC BY-NC-SA 4.0
- Personal jurisdiction Prompt: There is personal jurisdiction over a defendant in the state where the defendant is domiciled, or when (1) the defendant has sufficient contacts with the state, such that they have availed itself of the privileges of the state and (2) the claim arises out of the nexus of the defendant’s contacts with the state. Q: {text} Is there personal jurisdiction? A:	50	CC BY 4.0

B.4.5 MMLU-PRO (General Understanding)

The MMLU-Pro dataset (Wang et al., 2024), released under the Apache-2.0 License, is a more challenging version of the MMLU benchmark, designed to elevate the assessment of multi-task language understanding by incorporating more complex, reasoning-intensive questions across 14 challenging tasks. We utilize the full test set of 12,032 records for our evaluation. The raw dataset examples are mapped to our standard schema as follows:

- **`prompt`**: Each question is formatted into a standard multiple-choice prompt, with the choices presented in their original order. The template is as follows:

Prompt Template

Question: {question}
Choices:
(A) {option_A}
(B) {option_B}
...
Answer:

- **`explanation`**: This field is populated with the `cot_content` from the source dataset, which provides a ground-truth chain-of-thought reasoning trace.
- **`answer`**: The ground-truth `answer` is the single letter corresponding to the correct option.
- **`category`**: This field is populated with the `category` field from the source data (e.g., “math”, “physics”). The distribution of categories in the test set is shown in Table 23.

B.4.6 Big-Bench Hard (Complex Reasoning)

The Big-Bench Hard (BBH) dataset (Suzgun et al., 2022), released under the MIT License, is a suite of 27 challenging tasks designed to be beyond the capabilities of contemporary language models. The tasks are diverse, covering areas such as logical deduction, causal judgment, and tracking shuffled objects. We utilize the full test set, combining all 27 sub-tasks for a total of 6,511 records in our evaluation. The raw dataset examples are mapped to our standard schema as follows:

- **`prompt`**: The input question is formatted using the following simple template:

Prompt Template

Question: {input}
Answer:

- **`explanation`**: This field is intentionally left empty as the source dataset does not provide reasoning traces.
- **`answer`**: The ground-truth `answer` is the value from the original `target` field.
- **`category`**: This field contains the name of the specific BBH sub-task (e.g., “boolean_expressions”, “causal_judgement”). The category distribution is shown in Table 24.

B.5 Response Generation

All model outputs were generated using vLLM (Kwon et al., 2023) servers hosted locally and accessed through the OpenAI-compatible API interface. Each reasoning prompt was submitted as a user message, and the model’s completion was recorded as the response. To ensure deterministic and reproducible outputs, inference was performed with a temperature of 0.0 and a maximum generation length of 4096 tokens. To mitigate repetitive or looping text generation, which some reasoning models can exhibit even under deterministic decoding, we applied model-specific penalties following recommendations from their original releases. Specifically, we set the frequency penalty to 1.5 for `Magistral-Small-2506` and the frequency penalty to 0.8 with a presence penalty of 1.5 for `Phi-4-mini-flash-reasoning`. All responses were produced in batch mode via parallel threaded API calls to the vLLM endpoint, using a compute setup with 8 NVIDIA H200 GPUs, enabling efficient large-scale inference.

B.6 Response Segmentation

To analyze the step-by-step reasoning, each generated `model_response` was segmented into coherent units of thought, or “chunks.” This process follows the methodology of Zhang et al. (2025), which first splits the reasoning trace into paragraphs and then groups them based on a set of keywords that signal a new reasoning path. Based on our development experiments, we found that an expanded set of keywords provided more robust segmentation across the diverse model families in our study. Each chunk is designed to represent a single, coherent line of reasoning that often ends with an intermediate conclusion (see Table 25).

Table 22: LegalBench categories distribution (Test).

Category	**Test Samples**	**Test (%)**
international_citizenship_questions	9306	43.96
learned_hands_housing	4494	21.23
opp115_first_party_collection_use	2086	9.85
cuad_license_grant	1396	6.60
definition_classification	1337	6.32
opp115_international_and_specific_audiences	980	4.63
learned_hands_torts	432	2.04
diversity_5	300	1.42
diversity_6	300	1.42
learned_hands_domestic_violence	174	0.82
ucc_v_common_law	94	0.44
maud_cor_standard_(intervening_event)	84	0.40
cuad_third_party_beneficiary	68	0.32
learned_hands_benefits	66	0.31
personal_jurisdiction	50	0.24

Table 23: MMLU-Pro categories distribution (Test).

Category	**Test Samples**	**Test (%)**
math	1351	11.27
physics	1299	10.84
chemistry	1132	9.44
law	1080	9.01
engineering	969	8.08
other	924	7.71
economics	843	7.03
health	807	6.73
psychology	798	6.66
business	788	6.57
biology	707	5.90
philosophy	499	4.16
computer science	409	3.41
history	381	3.18

B.7 Response Grading

Each generated response was subsequently labeled as correct or incorrect. For multiple-choice questions, correctness was determined by a straightforward string match. For open-ended responses, we employed a state-of-the-art language model as a judge to assess semantic equivalence, a practice that has been validated and widely adopted in recent literature (Kapoor et al., 2024; Zhang et al., 2025; Chaudhry et al., 2024; Khanmohammadi et al., 2025). The reliability of using powerful LLMs for the constrained task of comparing a generated answer to a ground-truth answer has been demonstrated by Kapoor et al. (2024), who found that GPT-4's judgments closely align with human assessments (4.5% average difference). Building upon their findings, we utilized the more recent and capable `GPT-5-nano` model to ensure the highest quality labels. This approach does not use the grader as a knowledge oracle; rather, it provides the grader with the question, the ground-truth answer, and the model's generated answer, asking only if the two are semantically equivalent.

To provide a more granular analysis, we applied this grading on a per-chunk basis, similar to the approach in Zhang et al. (2025). This allows us to obtain a correctness label for each intermediate step in the reasoning process. The prompt used for this per-chunk grading task, which is an adjusted version of the one used in prior work, is detailed in Table 26. The total cost for this automated labeling process—calculated across all training and testing datasets and all LRMs—was approximately $746. The final sample- and chunk-level statistics for each LRM after grading are provided in Table 27 for the train split and Table 28 for the test split.

Table 24: Big-Bench Hard (BBH) categories distribution (Test).

Category	Test Samples	Test (%)
boolean_expressions	250	3.84
multistep_arithmetic_two	250	3.84
web_of_lies	250	3.84
tracking_shuffled_objects_three_objects	250	3.84
tracking_shuffled_objects_seven_objects	250	3.84
tracking_shuffled_objects_five_objects	250	3.84
temporal_sequences	250	3.84
salient_translation_error_detection	250	3.84
ruin_names	250	3.84
reasoning_about_colored_objects	250	3.84
object_counting	250	3.84
navigate	250	3.84
movie_recommendation	250	3.84
logical_deduction_three_objects	250	3.84
logical_deduction_seven_objects	250	3.84
logical_deduction_five_objects	250	3.84
hyperbaton	250	3.84
geometric_shapes	250	3.84
formal_fallacies	250	3.84
dyck_languages	250	3.84
disambiguation_qa	250	3.84
date_understanding	250	3.84
word_sorting	250	3.84
sports_understanding	248	3.81
causal_judgement	187	2.87
snarks	178	2.73
penguins_in_a_table	146	2.24

Table 25: Reasoning-path keywords used to detect reconsideration, verification, or alternate exploration behaviors in generated responses.

Category	Keywords
Verification	wait, double-check, make sure, verify, to confirm, let me verify, let me double-check, let me confirm
Alternative Approach	alternatively, another way, another approach, different approach
Reconsideration	but let me, let me try, on second thought, let me reconsider, let me check, hold on, wait a minute, let me think again, but what if

Table 26: System and user prompts employed for GPT-5-nano to assign correctness labels to individual model responses.

Prompt Type	**Content**
System Prompt	You are a meticulous grading assistant. A teacher has asked a student a question, and the student provided a step-by-step answer as a series of 'chunks'. Your task is to assist the teacher by evaluating each chunk of the student's reasoning and provide an overall assessment. You must follow the instructions precisely and provide your output only in the specified XML format.
User Prompt	**### Instruction** For each reasoning chunk from the student, evaluate whether its intermediate result exactly matches the Final Ground-Truth Answer. Mark each chunk with: - 1 if the chunk's intermediate result matches the ground-truth answer. - 0 if the chunk's intermediate result does not match the ground-truth answer. - null if the chunk does not contain any intermediate result (e.g., pure reflection/setup). After grading each chunk, provide a final grade that evaluates whether the model's final answer/conclusion matches the ground truth: - 1 if the final answer/conclusion matches the ground truth. - 0 if the final answer/conclusion does not match the ground truth. Your output must be a series of chunk evaluations in XML format, followed by a final grade: <chunk id="1">0/1/null</chunk> <chunk id="2">0/1/null</chunk> ... <final_grade>0/1</final_grade> — **### Context** * Question: "{prompt}" * Final Ground-Truth Answer: "{answer}" — **### Task: Grade Each Chunk** {reasoning_chunks}

Table 27: A breakdown of sample- and chunk-level train statistics across various datasets, aggregated by LRMs. **Notation:** S = number of samples (problems); C = chunks per sample; p = proportion; labels $L \in \{✓, ✗, \varnothing\}$ denote correct, incorrect, and no-result. **Columns:** S; ✓ = correct samples (#S, p); ✗ = incorrect samples (#S, p); $C(\mu \pm \sigma)$ = mean±sd chunks per sample; $\sum C$ = total chunks; ✓ = correct chunks (#C, p); ✗ = incorrect chunks (#C, p); ∅ = no-result chunks (#C, p). Proportions p are typeset tiny as $(.xx)$.

Dataset	S	✓ (#S, p)	✗ (#S, p)	C (μ±σ)	$\sum C$	✓ (#C, p)	✗ (#C, p)	∅ (#C, p)
`microsoft-Phi-4-mini-flash-reasoning`								
GSM8K	849	824 (.97)	25 (.03)	6.40±6.67	5,436	2,668 (.49)	1,185 (.22)	1,583 (.29)
TAT-QA	747	371 (.50)	376 (.50)	9.55±11.34	7,137	1,300 (.18)	3,476 (.49)	2,361 (.33)
MedQA	791	252 (.32)	539 (.68)	3.73±9.31	2,948	415 (.14)	1,671 (.57)	862 (.29)
LEXam	562	144 (.26)	418 (.74)	10.25±10.53	5,761	435 (.08)	3,902 (.68)	1,424 (.25)
ARC	811	661 (.81)	150 (.18)	4.95±11.16	4,016	1,296 (.32)	1,629 (.41)	1,091 (.27)
CommonsenseQA2	814	569 (.70)	245 (.30)	8.58±11.68	6,985	1,511 (.22)	2,311 (.33)	3,163 (.45)
LogiQA	760	338 (.44)	422 (.56)	3.23±7.29	2,453	484 (.20)	1,330 (.54)	639 (.26)
OpenBookQA	790	565 (.72)	225 (.28)	13.17±21.77	10,401	2,072 (.20)	4,109 (.40)	4,220 (.41)
QuaRTz	815	644 (.79)	171 (.21)	3.07±6.74	2,500	1,128 (.45)	747 (.30)	625 (.25)
ReClor	800	452 (.56)	348 (.43)	2.36±5.87	1,885	584 (.31)	937 (.50)	364 (.19)
TOTAL	**7,739**	**4,820** (.62)	**2,919** (.38)	**6.40±11.64**	**49,522**	**11,893** (.24)	**21,297** (.43)	**16,332** (.33)
`Qwen-Qwen3-8B`								
GSM8K	956	932 (.97)	24 (.03)	12.49±17.23	11,939	5,461 (.46)	2,322 (.19)	4,156 (.35)
TAT-QA	936	619 (.66)	317 (.34)	6.48±6.80	6,067	1,935 (.32)	2,454 (.40)	1,678 (.28)
MedQA	927	720 (.78)	207 (.22)	20.42±20.04	18,928	4,409 (.23)	7,427 (.39)	7,092 (.37)
LEXam	639	295 (.46)	344 (.54)	23.53±19.38	15,033	1,830 (.12)	9,429 (.63)	3,774 (.25)
ARC	984	945 (.96)	39 (.04)	8.03±10.75	7,904	3,406 (.43)	2,373 (.30)	2,125 (.27)
CommonsenseQA2	988	802 (.81)	186 (.19)	6.77±6.04	6,692	2,018 (.30)	1,726 (.26)	2,948 (.44)
LogiQA	906	683 (.75)	223 (.25)	13.73±13.87	12,441	2,610 (.21)	5,514 (.44)	4,317 (.35)
OpenBookQA	983	886 (.90)	97 (.10)	10.18±10.99	10,007	3,106 (.31)	3,716 (.37)	3,185 (.32)
QuaRTz	990	922 (.93)	68 (.07)	5.30±6.12	5,249	2,609 (.50)	1,140 (.22)	1,500 (.29)
ReClor	918	858 (.93)	60 (.07)	13.24±13.78	12,156	3,821 (.31)	4,690 (.39)	3,645 (.30)
TOTAL	**9,227**	**7,662** (.83)	**1,565** (.17)	**11.53±14.21**	**106,416**	**31,205** (.29)	**40,791** (.38)	**34,420** (.32)
`Qwen-Qwen3-14B`								
GSM8K	978	948 (.97)	30 (.03)	8.80±7.38	8,605	4,219 (.49)	1,668 (.19)	2,718 (.32)
TAT-QA	911	599 (.66)	312 (.34)	4.15±3.87	3,783	1,326 (.35)	1,528 (.40)	929 (.25)
MedQA	955	790 (.83)	165 (.17)	10.05±10.12	9,594	2,786 (.29)	3,235 (.34)	3,573 (.37)
LEXam	687	313 (.46)	374 (.54)	13.94±14.89	9,580	1,117 (.12)	6,592 (.69)	1,871 (.20)
ARC	976	932 (.95)	44 (.05)	4.33±5.76	4,225	1,958 (.46)	1,339 (.32)	928 (.22)
CommonsenseQA2	986	781 (.79)	205 (.21)	4.42±4.41	4,362	1,426 (.33)	1,295 (.30)	1,641 (.38)
LogiQA	895	664 (.74)	231 (.26)	7.59±8.49	6,794	1,584 (.23)	3,088 (.45)	2,122 (.31)
OpenBookQA	975	893 (.92)	82 (.08)	5.22±5.86	5,093	1,870 (.37)	1,748 (.34)	1,475 (.29)
QuaRTz	991	922 (.93)	69 (.07)	3.43±3.07	3,401	1,964 (.58)	706 (.21)	731 (.21)
ReClor	964	904 (.94)	60 (.06)	6.16±6.77	5,938	2,157 (.36)	2,243 (.38)	1,538 (.26)
TOTAL	**9,318**	**7,746** (.83)	**1,572** (.17)	**6.59±8.03**	**61,375**	**20,407** (.33)	**23,442** (.38)	**17,526** (.29)
`mistralai-Magistral-Small-2506`								
GSM8K	909	126 (.14)	783 (.86)	7.84±2.40	7,129	266 (.04)	2,370 (.33)	4,493 (.63)
TAT-QA	921	204 (.22)	717 (.78)	6.50±2.35	5,983	409 (.07)	1,336 (.22)	4,238 (.71)
MedQA	970	694 (.72)	276 (.28)	4.73±1.85	4,590	1,788 (.39)	1,404 (.31)	1,398 (.30)
LEXam	928	212 (.23)	716 (.77)	5.69±2.65	5,282	450 (.09)	3,374 (.64)	1,458 (.28)
ARC	986	855 (.87)	131 (.13)	3.89±1.95	3,839	1,986 (.52)	1,019 (.27)	834 (.22)
CommonsenseQA2	985	677 (.69)	308 (.31)	3.95±2.04	3,887	1,305 (.34)	1,244 (.32)	1,338 (.34)
LogiQA	965	510 (.53)	455 (.47)	5.23±2.24	5,051	1,223 (.24)	2,363 (.47)	1,465 (.29)
OpenBookQA	985	835 (.85)	150 (.15)	3.81±1.82	3,750	1,873 (.50)	900 (.24)	977 (.26)
QuaRTz	992	877 (.88)	115 (.12)	3.69±1.94	3,663	2,096 (.57)	824 (.23)	743 (.20)
ReClor	975	749 (.77)	226 (.23)	4.88±1.97	4,755	1,859 (.39)	1,695 (.36)	1,201 (.25)
TOTAL	**9,616**	**5,739** (.60)	**3,877** (.40)	**4.98±2.48**	**47,929**	**13,255** (.28)	**16,529** (.34)	**18,145** (.38)
`LGAI-EXAONE-EXAONE-Deep-32B`								
GSM8K	977	921 (.94)	56 (.06)	18.82±15.58	18,391	6,750 (.37)	4,540 (.25)	7,101 (.39)
TAT-QA	973	558 (.57)	415 (.43)	15.35±13.68	14,931	2,559 (.17)	6,182 (.41)	6,190 (.41)
MedQA	972	652 (.67)	320 (.33)	26.09±20.96	25,362	3,639 (.14)	10,765 (.42)	10,958 (.43)
LEXam	946	246 (.26)	700 (.74)	28.41±19.03	26,874	1,559 (.06)	17,653 (.66)	7,662 (.29)
ARC	992	933 (.94)	59 (.06)	9.22±11.80	9,142	3,233 (.35)	2,811 (.31)	3,098 (.34)
CommonsenseQA2	984	768 (.78)	216 (.22)	13.04±15.57	12,832	2,715 (.21)	3,755 (.29)	6,362 (.50)
LogiQA	970	513 (.53)	457 (.47)	21.45±17.01	20,811	2,246 (.11)	10,050 (.48)	8,515 (.41)
OpenBookQA	979	886 (.91)	93 (.10)	14.74±18.58	14,427	3,572 (.25)	4,589 (.32)	6,266 (.43)
QuaRTz	990	933 (.94)	57 (.06)	6.93±7.70	6,861	3,170 (.46)	1,179 (.17)	2,512 (.37)
ReClor	975	823 (.84)	152 (.16)	13.56±12.15	13,217	3,113 (.24)	5,583 (.42)	4,521 (.34)
TOTAL	**9,758**	**7,233** (.74)	**2,525** (.26)	**16.69±16.95**	**162,848**	**32,556** (.20)	**67,107** (.41)	**63,185** (.39)
`Qwen-QwQ-32B`								
GSM8K	984	946 (.96)	38 (.04)	10.52±12.47	10,348	4,458 (.43)	2,492 (.24)	3,398 (.33)
TAT-QA	983	627 (.64)	356 (.36)	7.65±9.09	7,518	1,974 (.26)	3,098 (.41)	2,446 (.33)
MedQA	975	790 (.81)	185 (.19)	15.27±19.66	14,886	3,292 (.22)	6,350 (.43)	5,244 (.35)
LEXam	947	304 (.32)	643 (.68)	28.89±27.30	27,359	2,239 (.08)	19,250 (.70)	5,870 (.21)

(Continued on next page)

(Continued from previous page)

Dataset	S	✓ (#S, p)	✗ (#S, p)	C ($\mu\pm\sigma$)	$\sum C$	✓ (#C, p)	✗ (#C, p)	∅ (#C, p)
ARC	987	945 (.96)	42 (.04)	6.11±12.05	6,027	2,449 (.41)	2,016 (.33)	1,562 (.26)
CommonsenseQA2	990	809 (.82)	181 (.18)	6.42±7.58	6,351	1,906 (.30)	1,762 (.28)	2,683 (.42)
LogiQA	988	687 (.70)	301 (.30)	15.95±18.96	15,758	2,484 (.16)	7,959 (.51)	5,315 (.34)
OpenBookQA	985	907 (.92)	78 (.08)	8.22±15.27	8,095	2,543 (.31)	2,899 (.36)	2,653 (.33)
QuaRTz	996	936 (.94)	60 (.06)	4.79±6.38	4,766	2,367 (.50)	1,058 (.22)	1,341 (.28)
ReClor	979	914 (.93)	65 (.07)	8.10±11.77	7,934	2,492 (.31)	3,145 (.40)	2,297 (.29)
TOTAL	**9,814**	**7,865** (.80)	**1,949** (.20)	**11.11±16.68**	**109,042**	**26,204** (.24)	**50,029** (.46)	**32,809** (.30)

Table 28: A breakdown of sample- and chunk-level test statistics across various datasets, aggregated by RLMs. **Notation:** S = number of samples (problems); C = chunks per sample; p = proportion; labels $L \in \{\checkmark, ✗, \varnothing\}$ denote correct, incorrect, and no-result. **Columns:** S; ✓ = correct samples (#S, p); ✗ = incorrect samples (#S, p); $C(\mu \pm \sigma)$ = mean±sd chunks per sample; $\sum C$ = total chunks; ✓ = correct chunks (#C, p); ✗ = incorrect chunks (#C, p); ∅ = no-result chunks (#C, p). Proportions p are typeset tiny as $(.xx)$.

Dataset	S	✓ (#S, p)	✗ (#S, p)	C ($\mu\pm\sigma$)	$\sum C$	✓ (#C, p)	✗ (#C, p)	∅ (#C, p)
`microsoft-Phi-4-mini-flash-reasoning`								
MATH	4,560	4,179 (.92)	381 (.08)	14.95±15.16	68,152	22,273 (.33)	21,197 (.31)	24,682 (.36)
FinQA	979	256 (.26)	723 (.74)	10.31±13.41	10,097	719 (.07)	5,834 (.58)	3,544 (.35)
MedMCQA	5,505	1,285 (.23)	4,220 (.77)	12.37±18.04	68,076	5,200 (.08)	37,607 (.55)	25,269 (.37)
LegalBench	18,321	8,649 (.47)	9,672 (.53)	5.35±7.38	97,965	19,841 (.20)	47,817 (.49)	30,307 (.31)
MMLU-Pro	9,404	5,063 (.54)	4,341 (.46)	15.64±19.81	147,104	22,850 (.16)	73,686 (.50)	50,568 (.34)
BBH	5,640	4,161 (.74)	1,479 (.26)	14.93±17.22	84,177	21,582 (.26)	31,061 (.37)	31,534 (.37)
TOTAL	**44,409**	**23,593** (.53)	**20,816** (.47)	**10.71±15.25**	**475,571**	**92,465** (.19)	**217,202** (.46)	**165,904** (.35)
`Qwen-Qwen3-8B`								
MATH	4,384	4,122 (.94)	262 (.06)	27.35±16.27	119,893	33,878 (.28)	36,069 (.30)	49,946 (.42)
FinQA	1,038	382 (.37)	656 (.63)	10.46±8.83	10,862	1,339 (.12)	5,927 (.55)	3,596 (.33)
MedMCQA	5,893	1,634 (.28)	4,259 (.72)	18.72±16.48	110,315	10,511 (.10)	60,755 (.55)	39,049 (.35)
LegalBench	20,720	13,997 (.68)	6,723 (.32)	5.39±5.97	111,704	26,217 (.23)	43,053 (.39)	42,434 (.38)
MMLU-Pro	9,962	7,624 (.77)	2,338 (.23)	24.49±18.81	243,933	48,678 (.20)	111,128 (.46)	84,127 (.34)
BBH	6,068	4,967 (.82)	1,101 (.18)	16.71±15.87	101,377	31,156 (.31)	31,110 (.31)	39,111 (.39)
TOTAL	**48,065**	**32,726** (.68)	**15,339** (.32)	**14.52±15.95**	**698,084**	**151,779** (.22)	**288,042** (.41)	**258,263** (.37)
`Qwen-Qwen3-14B`								
MATH	4,489	4,241 (.94)	248 (.06)	22.40±14.04	100,551	30,626 (.30)	29,685 (.30)	40,240 (.40)
FinQA	1,075	382 (.36)	693 (.64)	5.91±5.28	6,349	847 (.13)	3,450 (.54)	2,052 (.32)
MedMCQA	5,895	1,726 (.29)	4,169 (.71)	9.78±9.45	57,654	6,120 (.11)	31,640 (.55)	19,894 (.35)
LegalBench	20,463	14,433 (.71)	6,030 (.29)	3.77±3.90	77,123	20,958 (.27)	28,372 (.37)	27,793 (.36)
MMLU-Pro	10,312	8,061 (.78)	2,251 (.22)	15.46±13.78	159,452	35,322 (.22)	69,048 (.43)	55,082 (.35)
BBH	6,281	5,182 (.82)	1,099 (.17)	9.31±9.71	58,501	20,293 (.35)	16,519 (.28)	21,689 (.37)
TOTAL	**48,515**	**34,025** (.70)	**14,490** (.30)	**9.47±11.22**	**459,630**	**114,166** (.25)	**178,714** (.39)	**166,750** (.36)
`mistralai-Magistral-Small-2506`								
MATH	4,589	193 (.04)	4,396 (.96)	7.54±2.33	34,622	436 (.01)	8,874 (.26)	25,312 (.73)
FinQA	1,022	32 (.03)	990 (.97)	7.46±1.84	7,627	43 (.01)	1,896 (.25)	5,688 (.75)
MedMCQA	5,968	1,386 (.23)	4,582 (.77)	4.58±1.94	27,344	3,481 (.13)	16,642 (.61)	7,221 (.26)
LegalBench	20,810	10,845 (.52)	9,965 (.48)	3.56±1.54	74,066	19,700 (.27)	33,515 (.45)	20,851 (.28)
MMLU-Pro	11,347	3,783 (.33)	7,564 (.67)	6.39±2.82	72,552	9,290 (.13)	26,815 (.37)	36,447 (.50)
BBH	6,057	2,064 (.34)	3,993 (.66)	6.93±2.58	41,966	5,089 (.12)	15,461 (.37)	21,416 (.51)
TOTAL	**49,793**	**18,303** (.37)	**31,490** (.63)	**5.18±2.66**	**258,177**	**38,039** (.15)	**103,203** (.40)	**116,935** (.45)
`LGAI-EXAONE-EXAONE-Deep-32B`								
MATH	4,784	4,040 (.84)	744 (.16)	27.90±14.58	133,466	30,697 (.23)	48,691 (.36)	54,078 (.41)
FinQA	1,107	380 (.34)	727 (.66)	24.62±15.38	27,249	2,027 (.07)	13,416 (.49)	11,806 (.43)
MedMCQA	5,999	1,581 (.26)	4,418 (.74)	32.53±27.43	195,162	13,983 (.07)	96,255 (.49)	84,924 (.44)
LegalBench	20,879	14,648 (.70)	6,231 (.30)	12.56±15.41	262,286	42,549 (.16)	92,368 (.35)	127,369 (.49)
MMLU-Pro	11,473	7,119 (.62)	4,354 (.38)	32.72±20.99	375,342	47,973 (.13)	185,339 (.49)	142,030 (.38)
BBH	6,208	4,902 (.79)	1,306 (.21)	24.95±17.54	154,890	35,943 (.23)	52,970 (.34)	65,977 (.43)
TOTAL	**50,450**	**32,670** (.65)	**17,780** (.35)	**22.76±20.77**	**1,148,395**	**173,172** (.15)	**489,039** (.43)	**486,184** (.42)
`Qwen-QwQ-32B`								
MATH	4,813	4,177 (.87)	636 (.13)	24.56±16.65	118,207	30,404 (.26)	41,809 (.35)	45,994 (.39)
FinQA	1,112	406 (.37)	706 (.63)	13.16±15.00	14,630	1,411 (.10)	7,701 (.53)	5,518 (.38)
MedMCQA	6,040	1,695 (.28)	4,345 (.72)	21.06±27.68	127,195	10,465 (.08)	69,914 (.55)	46,816 (.37)
LegalBench	20,908	13,353 (.64)	7,555 (.36)	7.63±12.19	159,603	26,789 (.17)	60,468 (.38)	72,346 (.45)
MMLU-Pro	11,597	8,165 (.70)	3,432 (.30)	26.58±22.85	308,197	47,887 (.16)	153,966 (.50)	106,344 (.35)
BBH	6,322	5,266 (.83)	1,056 (.17)	14.24±13.61	90,051	26,997 (.30)	29,290 (.33)	33,764 (.37)
TOTAL	**50,792**	**33,062** (.65)	**17,730** (.35)	**16.10±19.76**	**817,883**	**143,953** (.18)	**363,148** (.44)	**310,782** (.38)

B.8 Example of a Fully Annotated and Segmented Reasoning Trace

To make the structure of the RMCB dataset concrete, we provide a fully worked example of a single reasoning trace. This example illustrates the raw input, the model-generated reasoning segmented into coherent regions, and the final answer produced after the internal reasoning phase. The example is shown exactly as stored in the dataset, with visual grouping used to indicate segmentation.

Table 29: Illustrative example of a single RMCB reasoning trace. Gray denotes the input prompt. Colored regions indicate distinct segments of the model's reasoning.

Input

> The pH scale measures acidity. A higher concentration of hydrogen ions (H^+) corresponds to a lower pH value. What would be the pH value of a solution with fewer hydrogen ions?

Model Reasoning

> Okay, so the question is asking about the pH value of something with fewer hydrogen ions. Let me think. The context says that a higher concentration of hydrogen ions makes a solution more acidic and has a lower pH. So if there are more H^+ ions, pH is lower. Then, the opposite should be true, right? If there are fewer hydrogen ions, the pH should be higher. Because pH is a scale where lower numbers are more acidic and higher numbers are more basic or alkaline.

> Wait, let me make sure I'm not mixing this up. The pH scale goes from 0 to 14. 7 is neutral. Below 7 is acidic, above is basic. Each number represents a tenfold difference in hydrogen ion concentration. So, for example, pH 3 has ten times more H^+ than pH 4. So if something has less H^+ ions, moving up the pH scale, like from 3 to 4, the pH increases. Therefore, less H^+ means higher pH. So the answer should be (A) high.

> But let me check again. The question says "something with less hydrogen ions would have what pH value?" Since more H^+ equals lower pH, then less H^+ would mean the pH is higher. So yes, the answer is A. The options are high (A) or low (B). The context directly states that higher H^+ leads to lower pH, so inversely, lower H^+ leads to higher pH. Therefore, the correct answer is A.

B.9 Benchmark Availability and Licensing

The **RMCB** benchmark is a composite resource built upon 16 distinct public datasets, each governed by its own license. These licenses range from highly permissive (e.g., MIT, Apache 2.0) to more restrictive terms that include NonCommercial (NC), ShareAlike (SA), or research-use-only clauses. To ensure full compliance with the terms of every source dataset, we cannot distribute the complete, aggregated benchmark as a single downloadable package. Instead, we adopt a standard, script-based reconstruction approach that separates our original contributions from the source data. We make the following components publicly available:

- **The Derived Data File:** A comprehensive JSONL file containing our novel contributions is released under the **CC BY 4.0 license**. Each line includes: (1) a unique, deterministic `record_id` created by hashing the source content; (2) the full, model-generated reasoning trace; and (3) the associated correctness annotations.

- **The Reconstruction Script:** A Python script that automates the process of building the full RMCB benchmark on a user's local machine. All of our code, including this script, is released under the **MIT license**.

To construct the full benchmark, users must first download the original source datasets, thereby agreeing to their respective licenses. Our provided script then processes these local files, generates the corresponding `record_id` hashes, and merges the source data with our derived data file to create the complete benchmark. This method ensures that the original data is never redistributed by us.

Important Disclaimer

Users are solely responsible for acquiring the source datasets from their official distributors and for adhering to their original license terms. The final, reconstructed RMCB dataset is a derivative work. As such, it is governed by the most restrictive terms of its constituent components. This means the complete benchmark is intended for **non-commercial, research-use only** and is subject to all applicable *ShareAlike (SA)* provisions inherited from its sources. The information provided here does not constitute legal advice.

C Confidence Estimation Methods

This section details the confidence estimation methods evaluated in our benchmark. We categorize them into two groups: established baseline methods adapted from the literature (C.1), and novel methods we developed to specifically address the challenges of multi-step reasoning (C.2). For all trainable models, hyperparameters were systematically optimized using Optuna with a consistent trial budget to ensure a fair and rigorous comparison.

C.1 Baseline Methods

We selected a representative set of baseline methods from recent literature, covering both simple probing techniques and more complex, reasoning-focused approaches.

C.1.1 YVCE (Verbalized Confidence Estimation)

Yoon et al. (2025) introduced a verbalized confidence estimation approach, where a reasoning model is prompted to assess its own solution after completing the reasoning process. For clarity, we refer to this method as YVCE. The procedure operates in two stages: first, the model generates its full reasoning trace and final answer; second, this output is re-used to construct a new conversational turn. A detailed system prompt, defining a 10-point confidence scale, is combined with a "nudge" phrase that encourages the model to select one of these labels as a continuation of its thought process. The final confidence score is then parsed from this verbalized self-assessment.

YVCE System Prompt

First, reason through the question step by step to arrive at an answer.
Then, thoroughly assess your confidence in that answer by evaluating your thinking process so far.
Finally, classify your confidence into one of the following classes based on how likely your answer is to be correct:
- "Almost no chance" (0.0–0.1)
- "Highly unlikely" (0.1–0.2)
- "Chances are slight" (0.2–0.3)
- "Unlikely" (0.3–0.4)
- "Less than even" (0.4–0.5)
- "Better than even" (0.5–0.6)
- "Likely" (0.6–0.7)
- "Very good chance" (0.7–0.8)
- "Highly likely" (0.8–0.9)
- "Almost certain" (0.9–1.0)

Each category reflects the probability that your answer is correct.
At the very end of your output, format your answer and confidence as
Answer: $ANSWER
Confidence: $CLASS
where CLASS is one of the names (only the names without the probability ranges) of the classes above.

YVCE Nudge Prompt

Now, finally, if I were to briefly mention my confidence among the given classes in the system prompt, I would choose

C.1.2 TLCC (Token-Level Chunk Classification)

TLCC investigates whether low-dimensional statistics derived purely from output logits can serve as a computationally efficient proxy for deep hidden-state representations. While methods like SFHS require storing and processing high-dimensional embedding vectors (often $d = 4096+$), TLCC operates on a compact set of uncertainty metrics calculated during the generation pass.

Token-Level Feature Extraction For every token x_t generated in the reasoning trace, we extract a vector of ten statistical features designed to capture various aspects of the model's local predictive uncertainty—ranging from confidence in the top choice to the dispersion of the probability mass. These features are detailed in Table 30.

Chunk Aggregation and Classification To map these variable-length sequences of token features to fixed-size chunk representations, we apply four aggregation statistics—mean, standard deviation, minimum, and maximum—across the tokens within each chunk. Additionally, we include the normalized token count as a feature to capture the length of the reasoning step. This process yields a fixed-dimensional vector $v_c \in \mathbb{R}^{41}$ for each chunk (10 features × 4 statistics + 1 length feature).

Similar to the SFHS framework, the sequence of chunk vectors $[v_1, v_2, \ldots, v_n]$ is treated as a time-series input. We evaluate three architectural variants to model the dependencies between these statistical summaries:

- **TLCC-MLP**: A simple baseline that mean-pools the chunk vectors and classifies them using a multi-layer perceptron.
- **TLCC-CONV**: Applies 1D convolutions over the chunk sequence to capture local temporal patterns in uncertainty dynamics.
- **TLCC-LSTM**: Uses a bidirectional LSTM to model the global evolution of confidence statistics across the reasoning trace.

All variants are trained via binary cross-entropy to predict the final correctness of the answer.

C.1.3 P(IK) (Probability of "I Know")

This method tests the hypothesis that an LRM's initial comprehension of a question contains a signal about its likelihood of answering correctly. It

Table 30: Token-level confidence features extracted for the TLCC method. For a vocabulary size V, let $\mathbf{z} \in \mathbb{R}^V$ be the logit vector and $\mathbf{p} = \mathrm{softmax}(\mathbf{z})$ be the probability distribution. Indices $(1), (2), \ldots$ denote rank-ordered elements such that $p_{(1)} \geq p_{(2)} \geq \cdots \geq p_{(V)}$.

Feature	**Definition / Formulation**	**Intuition**
Top-1 Probability	$p_{(1)}$	The model's raw confidence in its selected token.
Log Top-1 Prob	$\log(p_{(1)})$	Equivalent to the negative log-likelihood (NLL); sensitive to extremely low probabilities.
Logit Margin	$z_{(1)} - z_{(2)}$	The distance between the best and runner-up logits (unnormalized confidence).
Probability Gap	$p_{(1)} - p_{(2)}$	The gap between the best and runner-up probabilities; indicates competition between top candidates.
Entropy	$H(\mathbf{p}) = -\sum p_i \log p_i$	Measures total uncertainty; high entropy implies a "flat" distribution.
Normalized Entropy	$H(\mathbf{p}) / \log(V)$	Entropy scaled by vocabulary size, making it comparable across different models.
Top-k Mass	$\sum_{i=1}^{k=5} p_{(i)}$	Cumulative probability of the top 5 tokens; measures concentration in the head of the distribution.
Tail Mass	$1 - \sum_{i=1}^{k=5} p_{(i)}$	Probability mass assigned to unlikely tokens; captures the "long tail" risk.
L2 Concentration	$\sum p_i^2$	The Herfindahl index; approaches 1.0 for certainty and $1/V$ for uniform uncertainty.
Logit Std Dev	$\sigma(\mathbf{z})$	Standard deviation of raw logits; indicates the sharpness of the pre-softmax distribution.

trains a lightweight classifier, an MLP, on the hidden state corresponding to the final token of the *input prompt*. The goal is to predict the correctness of the eventual answer based solely on the model's state *before* it begins to generate a solution.

C.1.4 PHSV (Probing Hidden States for Self-Verification)

Zhang et al. (2025) introduced Probing Hidden States for Self-Verification (PHSV), which monitors a reasoning model's internal states throughout the reasoning process by training a lightweight classifier on the hidden representation of each intermediate reasoning chunk. The classifier is implemented as an MLP whose architecture (depth, width, and dropout) is tuned to predict the correctness of each step. To address class imbalance, the model is trained with a weighted binary cross-entropy loss. While originally proposed for early-exit classification, we adapt PHSV to our benchmark by using the confidence prediction from the **final chunk** as the overall score. In addition, we define `PHSV-half`, a variant trained on only 50% of the data, which we employ as a crucial feature extractor for our two-stage models.

C.2 Novel Benchmarked Methods

Building on these baselines, we developed and benchmarked a suite of novel architectures designed to better capture the sequential, relational, and dynamic properties of a reasoning trace.

C.2.1 SFHS (Stacked Final Hidden States)

A natural extension of PHSV, this family of models treats the entire reasoning trace as a single sequence. Instead of classifying each chunk independently, it feeds the full stack of chunk hidden states $[h_1, h_2, \ldots, h_n]$ into a more powerful sequential model to make a single, globally informed prediction. We evaluate three architectural variants: `SFHS-MLP`, `SFHS-CONV`, and `SFHS-LSTM`.

C.2.2 GNNs (Graph Neural Networks)

We reframe the reasoning trace as a graph where each chunk is a node, allowing us to test whether the structural properties of an argument can predict its correctness. We introduce three variants, each designed to isolate a different potential signal: the chronological flow, the logical coherence between all steps, and the evolution of the model's confidence.

GNN-SB (Sequential Binary) This method establishes a simple baseline by representing the reasoning trace as a sequential chain and testing whether the chronological flow of reasoning alone encodes enough signal to distinguish correct from incorrect answers. Each reasoning chunk is modeled as a node connected to its immediate successor, capturing local, step-to-step dependencies without incorporating richer relational structure. Three architectural variants are implemented, corresponding to different backbone operators from the PyTorch Geometric library: `GNN-SB-GAT`, `GNN-SB-GCN`, and `GNN-SB-GraphSAGE`. These variants differ only in how they aggregate neighborhood information.

- **Graph Construction** : Temporal chain with directed edges $i \to i + 1$.
- **Nodes** : Raw chunk hidden state h_t from the reasoning model.
- **Edges** : Chronological and unweighted (no additional attributes).
- **Backbones** : `GAT` (Veličković et al., 2018) (attention-based), `GCN` (Kipf and Welling, 2017) (spectral convolution), and `GraphSAG`

`E` (Hamilton et al., 2018) (sample-based message passing).

Together, these variants probe whether local sequential connectivity alone can account for reasoning correctness.

GNN-SR (Relational Graph) This model extends the sequential baseline by explicitly encoding semantic and logical relationships between all forward pairs of reasoning chunks. The hypothesis is that correct reasoning forms a globally coherent structure where distant steps reinforce one another, whereas incorrect reasoning exhibits contradictions or semantic drift. The reasoning trace is thus modeled as a fully connected, directed acyclic graph (DAG) in which every earlier chunk can influence all later ones. We evaluate three backbone operators—`GNN-SR-GINE`, `GNN-SR-NNConv`, and `GNN-SR-Transformer`—each differing in how it leverages edge features to modulate message passing.

- **Graph Construction**: Directed, fully connected DAG with edges $i \to j$ for all $i < j$.
- **Nodes**: Hidden state h_t of each reasoning chunk.
- **Edge Features**: Five-dimensional vector per edge $\langle i, j \rangle$ comprising: (1) NLI-style entailment, contradiction, and neutral probabilities; (2) *proximity* $= 1 -$ normalized distance; and (3) cosine similarity between h_i and h_j.
- **Backbones**: `GINEConv` (Hu et al., 2020) (MLP-injected messages), `NNConv` (Gilmer et al., 2017) (edge-conditioned kernels), and `TransformerConv` (Shi et al., 2021) (edge-aware attention).

Collectively, these three variants—`GNN-SR-GINE`, `GNN-SR-NNConv`, and `GNN-SR-Transformer`—test whether relational and semantic coherence across all reasoning steps enhances confidence prediction.

GNN-CD (Confidence Dynamics Graph) This variant shifts the focus from the semantic content of the reasoning to the meta-level *dynamics of confidence*. It tests the intuition that correct reasoning is characterized by a stable or increasing sense of certainty, whereas flawed reasoning involves sharp, erratic shifts in confidence. By using features from a probe trained on separate data, this model isolates the predictive power of the confidence trajectory itself.

- **Graph Construction**: Directed, fully connected forward DAG with edges $i \to j$ for all $i < j$ (identical reachability pattern to GNN-SR).
- **Nodes**: Concatenation of per-chunk `PHSV-half` confidence c_i and the probe's penultimate layer representation.
- **Edge Weights**: Scalar values for *every* forward pair $\langle i, j \rangle$ computed as the distributional distance (e.g., Wasserstein or KL) between the chunk-level token log-probability distributions of chunks i and j; this quantifies how much the model's belief state shifts between any two steps, not only adjacent ones.
- **Backbones**: `GCN2Conv` (Chen et al., 2020), `APPNP` (Gasteiger et al., 2022), `TAGConv` (Du et al., 2018).

The choice of GNN backbone differs across these variants due to the nature of their edge information. Architectures that natively support scalar *edge weights* (such as `GCN2Conv` and `APPNP`) do not simultaneously support rich, multi-dimensional *edge features*. Conversely, operators designed for edge attributes (like `GINEConv` and `TransformerConv`) do not accommodate scalar edge weights. This distinction explains why GNN-CD and GNN-SR employ different families of models. For a concise overview of which operators support edge weights versus edge attributes, we refer the reader to the PyTorch Geometric (PyG) documentation[1], which guided our model selection.

We instantiate GNN-CD with two feature strategies—`noft` (frozen `PHSV-half` features) and `ft` (end-to-end fine-tuning of `PHSV-half`)—and three backbones, with `GCN2Conv` further split into `same` vs. `dual` initial-residual schemes. The exact names appearing in tables and plots follow the pattern `GNN-CD-{noft,ft}-{GCN2Conv,APPNP,TAGConv}[-{same,dual}]`, yielding:

- `GNN-CD-noft-GCN2Conv-same`
- `GNN-CD-ft-GCN2Conv-same`
- `GNN-CD-noft-GCN2Conv-dual`
- `GNN-CD-ft-GCN2Conv-dual`
- `GNN-CD-noft-APPNP`

---

[1] `https://pytorch-geometric.readthedocs.io/en/latest/cheatsheet/gnn_cheatsheet.html`

- `GNN-CD-ft-APPNP`
- `GNN-CD-noft-TAGConv`
- `GNN-CD-ft-TAGConv`

Here, `same` uses each layer's input as the initial residual ($x_0{=}x$), whereas `dual` supplements message passing with a skip from the LRM's original hidden states, providing a second information stream.

C.2.3 CE (Chunk Ensemble)

This method approaches confidence estimation as a classical machine learning problem, training simple classifiers on high-level features derived from the reasoning trace. It follows a strict two-stage training protocol to prevent data leakage. A `PHSV-half` model is first trained on the initial 50% of the data and then frozen as a feature extractor. Using its outputs, we derive per-chunk confidence scores and form a fixed-length trajectory vector $[c_1, c_2, \ldots, c_L]$. Standard classifiers—`LogisticRegression`, `RandomForest`, `DecisionTree`, `KNN`, and `XGBoost`—are trained on this representation, and the best-performing model is reported for each LRM.

C.2.4 LateFusion

This model implements a hybrid, dual-stream architecture designed to integrate semantic and confidence-based signals before final prediction. It adheres to a two-stage protocol: a pre-trained `PHSV-half` model serves as a feature extractor on the first half of the data, and the `LateFusion` model is trained on the held-out half. Each reasoning trace is processed by two parallel streams. The *semantic stream* models the logical flow and content of the argument using the raw hidden states $[h_1, h_2, \ldots, h_L]$, while the *dynamics stream* models the evolution of certainty by processing the concatenation of each chunk's confidence score c_i and its penultimate-layer representation. The fused representations are concatenated and passed through a shared classifier. Variants differ along two axes: (i) whether the `PHSV-half` features are frozen (`noft`) or fine-tuned jointly (`ft`), and (ii) the network architecture used in both streams—`MLP`, `1D Conv`, or bidirectional `LSTM`. The resulting six variants (`LateFusion-noft-MLP`, `LateFusion-noft-Conv`, `LateFusion-noft-LSTM`, `LateFusion-ft-MLP`, `LateFusion-ft-Conv`, and `LateFusion-ft-LSTM`) test the contribution of fine-tuning and temporal modeling.

C.2.5 ETTIN

This model adapts the `jhu-clsp/ettin-encoder-17m`, originally designed for hallucination detection, to the task of reasoning confidence estimation. It treats the concatenated prompt and reasoning trace as a single text sequence. The text is fed into the ETTIN encoder, whose token embeddings are mean-pooled to form a single representation summarizing the reasoning process. This vector is passed to an MLP head trained with a binary cross-entropy loss on the overall correctness label. The model thus performs holistic, text-level confidence prediction without explicit step structure, providing a strong baseline for text-encoder-based calibration.

C.2.6 ETTIN-HGA

This variant extends ETTIN with a hierarchical architecture that explicitly models the structure and quality of intermediate reasoning steps. The reasoning trace is segmented into chunks and concatenated with `[SEP]` delimiters (`[CLS]` prompt `[SEP]` chunk_1 `[SEP]` ... `[SEP]` chunk_n). The ETTIN encoder processes the structured input, and hidden states at each `[SEP]` token serve as chunk-level embeddings. A hierarchical gated attention (HGA) module combines two components: an attention head that models inter-chunk dependencies and a quality head that predicts per-chunk correctness scores, which act as gating weights to emphasize coherent, reliable chunks. The gated, context-aware representations are then mean-pooled and passed through an MLP classifier. Training employs a composite loss combining the final correctness loss with an auxiliary per-chunk loss supervising the quality head. This design allows ETTIN-HGA to explicitly reason about the consistency and reliability of intermediate steps, yielding stronger, structure-aware confidence estimates.

D Evaluation Metrics

To comprehensively assess the quality of our confidence estimation framework, we employ a diverse collection of metrics that capture both calibration and discriminative performance. This combination enables a balanced interpretation of model behavior beyond simple accuracy, especially under class imbalance conditions that commonly arise in correctness prediction.

D.1 Expected Calibration Error (ECE)

Calibration reflects how well a model's predicted confidences correspond to actual empirical frequencies. In a well-calibrated system, predictions with confidence p should be correct approximately p fraction of the time. We compute the Expected Calibration Error (ECE) using a standard binning approach. Specifically, the confidence scores of all n samples are divided into b uniform bins $\{B_j\}_{j=1}^{b}$, and the deviation between mean confidence and empirical accuracy within each bin is aggregated as:

$$\text{ECE} = \sum_{j=1}^{b} \frac{|B_j|}{n} \left|\text{conf}(B_j) - \text{acc}(B_j)\right|$$

where $\text{conf}(B_j)$ denotes the average predicted confidence in bin B_j and $\text{acc}(B_j)$ the observed accuracy therein. Unless stated otherwise, we use $b = 10$. Lower values indicate better alignment between predicted and observed probabilities.

D.2 Brier Score

The Brier Score quantifies the mean squared distance between each predicted probability p_k and its ground-truth correctness label $o_k \in \{0, 1\}$:

$$\text{Brier Score} = \frac{1}{N} \sum_{k=1}^{N} (p_k - o_k)^2$$

This measure simultaneously captures aspects of calibration and sharpness, penalizing both over- and under-confident predictions. A smaller value reflects superior overall reliability.

D.3 Accuracy (ACC)

Accuracy represents the percentage of cases in which the model's predicted answer is correct:

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

While confidence estimation methods typically do not alter the base model's predictions, reporting ACC provides useful reference for the inherent difficulty of the task and contextualizes other confidence-related metrics.

D.4 F1 Score

To comprehensively evaluate discriminative performance, we report the F1 score, which captures the balance between precision and recall. Precision quantifies the proportion of predicted correct cases that are actually correct:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

while recall (or sensitivity) measures the proportion of truly correct cases that are successfully identified by the model:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

The F1 score is then defined as the harmonic mean of precision and recall:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

A higher F1 score indicates that the confidence estimator achieves both strong precision and high recall, effectively distinguishing correct predictions while minimizing false alarms and missed detections.

D.5 Specificity

Specificity (Spec), also known as the True Negative Rate (TNR), measures the ability of the confidence estimation model to correctly identify instances that are truly incorrect. In other words, it quantifies how effectively the model avoids assigning high confidence to wrong predictions. It is defined as:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

where TN denotes the number of correctly identified incorrect cases and FP represents the number of incorrect cases mistakenly classified as correct. High specificity indicates that the confidence estimator is conservative—rarely overconfident in wrong predictions—thereby complementing sensitivity (or recall) to provide a more complete view of discriminative reliability.

D.6 Area Under the Precision–Recall Curve (AUCPR)

AUCPR condenses the trade-off between precision $(\text{TP}/(\text{TP} + \text{FP}))$ and recall $(\text{TP}/(\text{TP} + \text{FN}))$ into a single score by integrating over varying discrimination thresholds. It is especially sensitive to performance on the positive (correct) class and therefore more informative than AUROC when class distributions are skewed.

D.7 Area Under the ROC Curve (AUROC)

AUROC evaluates the ability of the confidence estimator to discriminate between correct and incorrect answers. It plots the true positive rate ($\mathrm{TPR} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$) against the false positive rate ($\mathrm{FPR} = \mathrm{FP}/(\mathrm{FP} + \mathrm{TN})$) as the confidence threshold varies. An AUROC of 1.0 denotes perfect separability, whereas 0.5 corresponds to chance-level discrimination.

E Training and Optimization Details

All trainable confidence estimation models in our benchmark were developed using a consistent and rigorous training protocol to ensure a fair comparison. Training was conducted with a fixed batch size of 32. Hyperparameter tuning was performed using the Optuna framework (Akiba et al., 2019), with each model variant undergoing up to 100 trials to find an optimal configuration.

E.1 Hyperparameter Optimization with Optuna

We utilized `Optuna` to systematically explore architectural, training, and regularization parameters. A Tree-structured Parzen Estimator (TPE) sampler was employed to intelligently suggest new configurations based on past results. The objective optimized in each study was a composite score designed to address the central trade-off between discrimination and calibration:

$$\mathrm{CompositeScore} = \alpha \cdot \mathrm{AUROC} + (1 - \alpha) \cdot (1 - \mathrm{ECE})$$

where we set $\alpha = 0.6$ to place a slight emphasis on AUROC, prioritizing the discovery of models with strong discriminative power while still imposing a significant penalty for poor calibration.

For final model selection, we imposed an additional practical constraint. A trial was only considered "feasible" if its best-performing epoch also achieved a minimum sensitivity and specificity of 0.50 at its Youden's J optimal threshold. This ensures our final reported models are not only well-balanced in terms of AUROC and ECE, but also demonstrate a tangible predictive ability better than random chance. Among all feasible trials, the one with the highest composite score was selected.

Each trial was trained for a maximum of 200 epochs with an early-stopping patience of 20 epochs based on the composite validation score. To accelerate the search, Optuna's Median Pruner was applied with a "moderate" schedule to terminate unpromising trials early.

E.2 Hyperparameter Search Spaces

The specific hyperparameter search spaces for each model family are detailed in Table 31. All methods shared a common search space for the final classifier's architecture (e.g., `classifier_layers`, `classifier_dropout`), learning rate, and weight decay. To ensure a fair comparison of architectural efficiency and prevent model complexity from being conflated with raw parameter count, all configurations were constrained to a maximum of 3.2 million trainable parameters. Any trial suggesting a model outside this budget was immediately discarded. Upon completion of each 100-trial study, the best feasible configuration was selected as the final model for evaluation.

F Comprehensive Results

This appendix presents complementary views of performance so readers can inspect both granular and aggregated behavior across models, datasets, and methods. We report *ECE*, *Brier*, *Acc*, *F1*, *Spec*, *AUCPR*, and *AUROC*. Lower is better for ECE and Brier; higher is better for the others. Within any comparison group (e.g., a given dataset inside a per-LLM table), the best value per metric is shown in **bold** using the appropriate direction. Tables 32, 33, 34, 35, 36, and 37 enumerate, for each LLM, every method's performance on each test dataset without aggregation, supporting fine-grained, within-model comparisons. Tables 38, 39, 40, 41, 42, and 43 summarize, for each LLM, the *unweighted* mean ± standard deviation of each method across all datasets, providing a concise per-model overview that averages out dataset-level variability. Table 44 aggregates each *dataset–method* pair across LLMs, reporting the *unweighted* mean ± standard deviation to indicate which methods generalize well on a given dataset independent of the underlying model. Table 45 groups each *LLM–method* pair across datasets as *unweighted* mean ± standard deviation, emphasizing which methods work best for a particular LLM after averaging out dataset effects. Finally, Table 46 provides a method-only view using a two-stage *unweighted* average: first average a method across datasets within each LLM, then average those LLM-level means across LLMs; we report the corresponding standard deviations.

Table 31: Hyperparameter search space. All methods inherit the shared space; rows list only method-specific additions.

Method	**Variant**	**Search space**
Shared (all methods)		learning_rate $\in$ {1e-4, 1e-3}; weight_decay $\in$ {1e-5, 1e-4}; classifier_layers $\in$ {128,64; 128,32; 64,32; 32,16; 128; 64; 32; 0; 256,128; 512,256; 256; 512}; classifier_dropout $\in \{0.1,\ 0.25,\ 0.4\}$; budget $[1,\ 3.2$M params]
PIK	–	– (no additional hyperparameters)
PHSV	–	– (no additional hyperparameters)
SFHS	MLP	– (uses only shared)
	Conv	conv_layers $\in$ {32,64; 64,128}; kernel_sizes $\in$ {3,3; 5,3}; dropout $\in \{0.1,\ 0.25,\ 0.4\}$
	LSTM	hidden_dim $\in \{16,\ 32,\ 64\}$; num_layers $\in \{1, 2\}$; bidirectional $\in$ {`True`, `False`}; dropout $\in \{0.1,\ 0.25,\ 0.4\}$
TLCC	MLP	– (uses only shared)
	Conv	conv_layers $\in$ {32,64; 64,128}; kernel_sizes $\in$ {3,3; 5,3}; dropout $\in \{0.1,\ 0.25,\ 0.4\}$
	LSTM	hidden_dim $\in \{16,\ 32,\ 64\}$; num_layers $\in \{1, 2\}$; bidirectional $\in$ {`True`, `False`}; dropout $\in \{0.1,\ 0.25,\ 0.4\}$
LateFusion	MLP	semantic_hidden $\in$ {128,64; 64; 0}; dynamics_hidden $\in$ {64,32; 32; 0}
	Conv	semantic_conv $\in$ {32,64}; semantic_kernels $\in$ {3,3; 5,3}; dynamics_conv $\in$ {16,32; 32,32}; dynamics_kernels $\in$ {3,3}; dropout $\in \{0.1,\ 0.25,\ 0.4\}$
	LSTM	semantic_hidden_dim $\in \{32,\ 64\}$; semantic_num_layers $= 1$; semantic_bidirectional $=$ `True`; dynamics_hidden_dim $\in \{16,\ 32\}$; dynamics_num_layers $= 1$; dynamics_bidirectional $=$ `True`; dropout $\in \{0.1,\ 0.25,\ 0.4\}$
GNN_EdgeAttr	GINE	hidden_dim $\in \{64,\ 128,\ 256\}$; num_layers $\in \{1, 2\}$; pooling $\in$ {`mean`, `max`, `sum`, `attention`, `last_node`}; edge_nn $\in$ {`linear`, `mlp_small`, `mlp_medium`}
	Transformer	hidden_dim $\in \{64,\ 128,\ 256\}$; num_layers $\in \{1, 2\}$; pooling as above; heads $\in \{2, 4, 8\}$; concat $\in$ {`True`, `False`}
	NNConv	hidden_dim $\in \{32,\ 64,\ 128,\ 256\}$; num_layers $\in \{1, 2\}$; pooling as above; edge_nn=`linear`
GNN_Conf	GCN2_Same	hidden_dim $\in \{128,\ 256,\ 512\}$; num_layers $\in \{1, 2\}$; pooling as above; $\alpha \in \{0.1, 0.3, 0.5\}$; $\theta \in \{1.0, 1.5, 2.0\}$; shared_weights=True
	GCN2_Dual	hidden_dim $\in \{128,\ 256\}$; num_layers $\in \{1, 2\}$; pooling as above; α, θ as above; shared_weights=True
	TAGConv	hidden_dim $\in \{128,\ 256,\ 512\}$; num_layers $\in \{1, 2\}$; $K \in \{2, 3\}$; pooling as above
	APPNP	hidden_dim $\in \{128,\ 256,\ 512\}$; num_layers $\in \{1, 2\}$; $K \in \{2, 3\}$; appnp_alpha $\in \{0.1, 0.5, 0.9\}$; pooling as above
GNN_SB	GCN	gnn_type=gcn; hidden_dim $\in \{64, 128, 256\}$; num_layers $\in \{1, 2, 3, 4\}$; pooling $\in$ {`mean`, `max`, `sum`}
	GAT	gnn_type=gat; hidden_dim as above; num_layers as above; pooling as above; heads $\in \{1, 2, 4\}$; concat $\in$ {`True`, `False`}
	GraphSAGE	gnn_type=graphsage; hidden_dim as above; num_layers as above; pooling as above; aggr $\in$ {`mean`, `max`, `add`}
ETTIN	–	– (no additional hyperparameters)
ETTIN-HGA	–	attention_dropout $\in \{0.1,\ 0.25,\ 0.4\}$; quality_layers same as classifier_layers above

Table 32: Performance metrics for `Phi-4-mini-flash-reasoning` showing results per method within each test dataset. Each metric value represents the performance of the specified method on the specified dataset for this LLM. **Bold** entries mark the best-performing method for each metric within each dataset.

Dataset	Method	ECE↓	Brier↓	Acc↑	F1↑	Spec↑	AUCPR↑	AUROC↑
BBH	YVCE	0.171	0.245	0.648	0.767	0.295	0.845	0.625
	TLCC-MLP	0.110	0.171	0.723	0.809	0.577	0.906	0.755
	TLCC-CONV	0.099	0.169	0.719	0.807	0.557	0.907	0.756
	TLCC-LSTM	0.083	0.164	0.741	0.827	0.530	0.905	0.757
	P(IK)	0.033	0.171	**0.758**	**0.862**	0.012	0.868	0.682
	PHSV-half	0.078	0.197	0.710	0.817	0.212	0.830	0.683
	PHSV	0.117	0.204	0.691	0.801	0.231	0.837	0.693
	SFHS-MLP	0.076	0.165	0.719	0.813	0.449	0.889	0.744
	SFHS-Conv	0.072	0.173	0.708	0.804	0.454	0.890	0.729
	SFHS-LSTM	0.082	0.168	0.726	0.817	0.477	0.901	0.753
	GNN-SB-GAT	0.072	0.166	0.717	0.811	0.454	0.896	0.743
	GNN-SB-GCN	0.058	0.171	0.714	0.812	0.408	0.887	0.720
	GNN-SB-GraphSAGE	0.061	0.165	0.719	0.814	0.431	0.902	0.748
	CE-DT	0.098	0.217	0.690	0.797	0.294	0.850	0.664
	CE-KNN	0.086	0.198	0.701	0.800	0.361	0.841	0.681
	CE-LogReg	0.102	0.200	0.721	0.838	0.001	0.848	0.700
	CE-RF	**0.031**	0.182	0.721	0.821	0.299	0.856	0.709
	CE-XGB	0.055	0.184	0.718	0.817	0.310	0.857	0.710
	LateFusion-noft-MLP	0.115	0.181	0.709	0.802	0.490	0.885	0.720
	LateFusion-noft-Conv	0.176	0.189	0.688	0.770	0.690	0.899	0.755
	LateFusion-noft-LSTM	0.096	0.172	0.732	0.822	0.460	0.896	0.750
	LateFusion-ft-MLP	0.116	0.176	0.724	0.817	0.436	0.886	0.727
	LateFusion-ft-Conv	0.342	0.285	0.544	0.590	0.893	0.887	0.739
	LateFusion-ft-LSTM	0.085	0.176	0.727	0.820	0.431	0.895	0.740
	GNN-SR-GINE	0.111	0.173	0.717	0.808	0.507	0.886	0.738
	GNN-SR-NNConv	0.311	0.265	0.468	0.481	**0.917**	0.886	0.726
	GNN-SR-Transformer	0.127	0.171	0.730	0.819	0.498	**0.909**	**0.764**
	GNN-CD-noft-GCN2Conv-same	0.188	0.215	0.743	0.844	0.202	0.873	0.680
	GNN-CD-noft-GCN2Conv-dual	0.068	0.166	0.725	0.814	0.520	0.897	0.752
	GNN-CD-noft-APPNP	0.258	0.260	0.733	0.831	0.308	0.870	0.678
	GNN-CD-noft-TAGConv	0.201	0.217	0.721	0.822	0.315	0.881	0.660

(Continued on next page)

(Continued from previous page)

Dataset	Method	ECE↓	Brier↓	Acc↑	F1↑	Spec↑	AUCPR↑	AUROC↑
	GNN-CD-ft-GCN2Conv-same	0.099	0.172	0.730	0.819	0.490	0.882	0.738
	GNN-CD-ft-GCN2Conv-dual	0.242	0.244	0.736	0.841	0.165	0.882	0.601
	GNN-CD-ft-APPNP	0.088	0.185	0.716	0.818	0.330	0.849	0.670
	GNN-CD-ft-TAGConv	0.149	0.224	0.720	0.817	0.389	0.878	0.679
	ETTIN	0.150	0.187	0.707	0.798	0.534	0.892	0.722
	ETTIN-HGA	0.053	**0.160**	0.751	0.844	0.317	0.905	0.753
FinQA	YVCE	0.339	0.354	0.496	0.358	0.489	0.339	0.521
	TLCC-MLP	0.359	0.341	0.495	0.387	0.451	0.311	0.567
	TLCC-CONV	0.346	0.342	0.489	0.357	0.468	0.312	0.556
	TLCC-LSTM	0.371	0.355	0.468	0.385	0.406	0.313	0.566
	P(IK)	0.114	0.214	0.688	0.185	0.897	0.299	0.539
	PHSV-half	0.308	0.293	0.461	0.378	0.394	0.288	0.549
	PHSV	0.310	0.298	0.454	0.370	0.390	0.292	0.550
	SFHS-MLP	0.189	0.238	0.628	0.336	0.734	0.344	0.574
	SFHS-Conv	0.158	0.227	0.633	0.333	0.745	0.342	0.574
	SFHS-LSTM	0.192	0.244	0.622	0.378	0.697	0.335	0.596
	GNN-SB-GAT	0.174	0.232	0.631	0.245	0.785	0.305	0.532
	GNN-SB-GCN	0.181	0.237	0.649	0.262	0.807	0.310	0.527
	GNN-SB-GraphSAGE	0.174	0.233	0.611	0.274	0.739	0.330	0.547
	CE-DT	0.251	0.273	0.493	0.387	0.442	0.318	0.575
	CE-KNN	0.229	0.263	0.583	0.367	0.614	0.282	0.571
	CE-LogReg	0.354	0.310	0.249	0.397	0.003	0.313	0.559
	CE-RF	0.245	0.250	0.532	0.375	0.519	0.299	0.572
	CE-XGB	0.236	0.245	0.551	0.379	0.550	0.305	0.578
	LateFusion-noft-MLP	0.106	0.208	0.713	0.224	0.923	0.349	0.559
	LateFusion-noft-Conv	0.132	0.213	0.686	0.332	0.835	0.349	0.581
	LateFusion-noft-LSTM	0.133	0.224	0.672	0.410	0.768	0.354	**0.619**
	LateFusion-ft-MLP	0.201	0.235	0.683	0.283	0.854	0.353	0.588
	LateFusion-ft-Conv	0.102	0.209	0.727	0.009	0.998	0.314	0.548
	LateFusion-ft-LSTM	0.090	0.215	0.654	0.298	0.798	0.330	0.590
	GNN-SR-GINE	0.179	0.228	0.705	0.199	0.919	0.324	0.570
	GNN-SR-NNConv	**0.085**	**0.203**	**0.729**	0.009	**1.000**	0.331	0.562
	GNN-SR-Transformer	0.179	0.225	0.681	0.296	0.844	0.362	0.602
	GNN-CD-noft-GCN2Conv-same	0.459	0.451	0.443	**0.441**	0.306	0.369	0.591
	GNN-CD-noft-GCN2Conv-dual	0.121	0.217	0.677	0.237	0.862	0.316	0.545
	GNN-CD-noft-APPNP	0.403	0.406	0.573	0.391	0.600	0.372	0.577
	GNN-CD-noft-TAGConv	0.427	0.445	0.521	0.415	0.482	0.377	0.575
	GNN-CD-ft-GCN2Conv-same	0.159	0.230	0.666	0.305	0.815	0.334	0.568
	GNN-CD-ft-GCN2Conv-dual	0.522	0.548	0.429	0.399	0.329	**0.429**	0.516
	GNN-CD-ft-APPNP	0.230	0.260	0.570	0.385	0.598	0.349	0.570
	GNN-CD-ft-TAGConv	0.244	0.306	0.614	0.378	0.682	0.362	0.553
	ETTIN	0.194	0.229	0.593	0.363	0.655	0.366	0.610
	ETTIN-HGA	0.130	0.210	0.694	0.321	0.855	0.363	0.603
LegalBench	YVCE	0.273	0.339	0.491	0.476	0.505	0.497	0.497
	TLCC-MLP	0.074	0.257	0.550	0.457	0.696	0.547	0.554
	TLCC-CONV	0.097	0.260	**0.553**	0.460	0.698	0.545	0.555
	TLCC-LSTM	0.079	0.256	0.548	0.476	0.662	0.547	0.557
	P(IK)	0.104	0.267	0.496	0.584	0.276	0.463	0.492
	PHSV-half	0.117	0.269	0.521	0.530	0.466	0.500	0.543
	PHSV	0.111	0.265	0.541	0.518	0.543	0.498	0.551
	SFHS-MLP	0.110	0.266	0.531	0.349	0.786	0.519	0.542
	SFHS-Conv	0.108	0.268	0.540	0.427	0.716	0.516	0.546
	SFHS-LSTM	0.122	0.271	0.536	0.361	0.785	0.518	0.538
	GNN-SB-GAT	0.058	0.255	0.539	0.429	0.711	0.530	0.549
	GNN-SB-GCN	0.070	0.256	0.537	0.416	0.723	0.524	0.545
	GNN-SB-GraphSAGE	0.085	0.258	0.543	0.432	0.716	0.529	0.553
	CE-DT	0.102	0.282	0.524	0.503	0.523	0.486	0.531
	CE-KNN	0.166	0.288	0.529	0.438	0.639	0.483	0.527
	CE-LogReg	0.138	0.266	0.461	**0.630**	0.007	0.498	0.544
	CE-RF	0.076	0.257	0.535	0.455	0.631	0.498	0.540
	CE-XGB	0.088	0.259	0.534	0.455	0.629	0.496	0.540
	LateFusion-noft-MLP	0.121	0.264	0.539	0.300	0.853	0.536	0.559
	LateFusion-noft-Conv	0.083	0.260	0.526	0.237	0.878	0.518	0.540
	LateFusion-noft-LSTM	0.152	0.283	0.541	0.405	0.748	0.525	0.545
	LateFusion-ft-MLP	**0.020**	0.249	0.545	0.370	0.798	0.534	0.555
	LateFusion-ft-Conv	0.290	0.335	0.515	0.005	0.997	0.509	0.538
	LateFusion-ft-LSTM	0.173	0.285	0.529	0.397	0.726	0.519	0.536
	GNN-SR-GINE	0.021	**0.248**	0.537	0.332	0.819	0.534	0.555
	GNN-SR-NNConv	0.121	0.262	0.515	0.003	**0.998**	0.534	**0.561**
	GNN-SR-Transformer	0.035	0.249	0.539	0.363	0.792	0.530	0.551
	GNN-CD-noft-GCN2Conv-same	0.291	0.358	0.532	0.567	0.438	0.539	0.547
	GNN-CD-noft-GCN2Conv-dual	0.096	0.265	0.542	0.381	0.778	0.536	0.556
	GNN-CD-noft-APPNP	0.431	0.439	0.545	0.505	0.608	0.545	0.557
	GNN-CD-noft-TAGConv	0.226	0.333	0.537	0.484	0.622	0.546	0.547
	GNN-CD-ft-GCN2Conv-same	0.126	0.277	0.536	0.313	0.836	0.525	0.544
	GNN-CD-ft-GCN2Conv-dual	0.428	0.443	0.517	0.604	0.290	**0.622**	0.540
	GNN-CD-ft-APPNP	0.098	0.260	0.538	0.424	0.713	0.523	0.551
	GNN-CD-ft-TAGConv	0.206	0.325	0.527	0.462	0.629	0.526	0.539
	ETTIN	0.090	0.261	0.531	0.453	0.655	0.505	0.538
	ETTIN-HGA	0.071	0.257	0.517	0.484	0.565	0.504	0.525
MATH	YVCE	0.225	0.185	0.761	0.860	0.395	0.960	0.619
	TLCC-MLP	0.096	0.084	0.877	0.934	0.213	0.960	0.691
	TLCC-CONV	0.109	0.090	0.845	0.913	0.420	0.962	0.701

(Continued on next page)

(Continued from previous page)

Dataset	Method	ECE↓	Brier↓	Acc↑	F1↑	Spec↑	AUCPR↑	AUROC↑
	TLCC-LSTM	0.072	0.077	0.884	0.937	0.253	0.958	0.689
	P(IK)	0.074	0.065	**0.937**	**0.967**	0.004	0.960	0.632
	PHSV-half	0.067	0.137	0.831	0.905	0.152	0.889	0.623
	PHSV	0.098	0.153	0.803	0.887	0.146	0.877	0.629
	SFHS-MLP	0.054	0.075	0.879	0.935	0.197	0.963	0.669
	SFHS-Conv	0.062	0.072	0.900	0.947	0.176	0.969	0.708
	SFHS-LSTM	0.050	0.073	0.894	0.943	0.151	0.965	0.685
	GNN-SB-GAT	0.087	0.078	0.875	0.932	0.210	0.968	0.700
	GNN-SB-GCN	0.091	0.081	0.875	0.932	0.206	0.965	0.691
	GNN-SB-GraphSAGE	0.092	0.077	0.876	0.933	0.202	0.970	0.722
	CE-DT	0.134	0.129	0.866	0.926	0.275	0.951	0.659
	CE-KNN	0.091	0.098	0.880	0.934	0.328	0.957	0.735
	CE-LogReg	0.253	0.141	0.900	0.948	0.000	0.957	**0.763**
	CE-RF	0.088	0.092	0.895	0.943	0.214	0.955	0.747
	CE-XGB	0.069	0.091	0.893	0.942	0.241	0.953	0.734
	LateFusion-noft-MLP	0.183	0.101	0.880	0.935	0.319	0.969	0.719
	LateFusion-noft-Conv	0.197	0.107	0.872	0.930	0.374	0.967	0.711
	LateFusion-noft-LSTM	0.066	0.077	0.888	0.940	0.172	0.964	0.684
	LateFusion-ft-MLP	0.231	0.114	0.878	0.934	0.202	0.968	0.715
	LateFusion-ft-Conv	0.380	0.236	0.630	0.764	0.538	0.956	0.624
	LateFusion-ft-LSTM	0.152	0.094	0.885	0.938	0.218	0.968	0.701
	GNN-SR-GINE	0.121	0.087	0.875	0.933	0.189	0.953	0.621
	GNN-SR-NNConv	0.428	0.241	0.682	0.802	**0.601**	0.966	0.680
	GNN-SR-Transformer	0.188	0.097	0.887	0.939	0.176	0.970	0.711
	GNN-CD-noft-GCN2Conv-same	0.069	0.067	0.923	0.959	0.092	0.975	0.707
	GNN-CD-noft-GCN2Conv-dual	0.066	0.075	0.883	0.937	0.248	0.965	0.693
	GNN-CD-noft-APPNP	0.085	0.085	0.912	0.953	0.261	0.974	0.700
	GNN-CD-noft-TAGConv	0.067	0.065	0.917	0.956	0.181	**0.975**	0.658
	GNN-CD-ft-GCN2Conv-same	0.070	0.077	0.896	0.944	0.172	0.950	0.609
	GNN-CD-ft-GCN2Conv-dual	0.095	0.091	0.896	0.945	0.101	0.972	0.591
	GNN-CD-ft-APPNP	0.206	0.102	0.896	0.944	0.231	0.964	0.681
	GNN-CD-ft-TAGConv	0.095	0.094	0.897	0.945	0.122	0.960	0.529
	ETTIN	0.104	0.075	0.921	0.958	0.113	0.966	0.686
	ETTIN-HGA	**0.038**	**0.063**	0.936	0.967	0.008	0.967	0.694
MedMCQA	YVCE	0.405	0.378	0.374	0.359	0.260	0.269	0.507
	TLCC-MLP	0.299	0.296	0.505	0.321	0.507	0.234	0.502
	TLCC-CONV	0.297	0.294	0.493	0.330	0.481	0.238	0.510
	TLCC-LSTM	0.323	0.313	0.460	0.335	0.423	0.237	0.502
	P(IK)	0.376	0.327	0.274	**0.374**	0.074	0.232	0.491
	PHSV-half	0.458	0.414	0.319	0.349	0.177	0.238	0.515
	PHSV	0.418	0.382	0.361	0.341	0.252	0.234	0.511
	SFHS-MLP	0.306	0.307	0.518	0.326	0.525	0.245	0.515
	SFHS-Conv	0.324	0.324	0.479	0.330	0.458	0.244	0.508
	SFHS-LSTM	0.326	0.324	0.485	0.344	0.457	0.248	**0.523**
	GNN-SB-GAT	0.298	0.285	0.508	0.328	0.506	0.238	0.508
	GNN-SB-GCN	0.328	0.312	0.481	0.346	0.448	0.244	0.522
	GNN-SB-GraphSAGE	0.322	0.314	0.482	0.341	0.453	0.243	0.515
	CE-DT	0.384	0.364	0.357	0.351	0.235	0.271	0.513
	CE-KNN	0.358	0.345	0.441	0.333	0.388	0.238	0.517
	CE-LogReg	0.393	0.328	0.223	0.364	0.000	0.237	0.507
	CE-RF	0.371	0.332	0.377	0.348	0.272	0.239	0.514
	CE-XGB	0.373	0.337	0.383	0.343	0.285	0.240	0.513
	LateFusion-noft-MLP	0.261	0.274	0.544	0.311	0.576	0.248	0.513
	LateFusion-noft-Conv	0.201	0.233	0.626	0.260	0.730	0.242	0.505
	LateFusion-noft-LSTM	0.298	0.309	0.501	0.332	0.492	0.241	0.513
	LateFusion-ft-MLP	0.302	0.278	0.525	0.331	0.532	0.246	0.514
	LateFusion-ft-Conv	**0.086**	**0.194**	0.740	0.058	0.956	0.235	0.504
	LateFusion-ft-LSTM	0.258	0.288	0.524	0.299	0.552	0.231	0.496
	GNN-SR-GINE	0.290	0.269	0.516	0.335	0.514	0.245	0.518
	GNN-SR-NNConv	0.156	0.205	**0.761**	0.048	**0.985**	0.248	0.519
	GNN-SR-Transformer	0.278	0.264	0.498	0.332	0.487	0.243	0.516
	GNN-CD-noft-GCN2Conv-same	0.580	0.571	0.335	0.366	0.187	0.301	0.521
	GNN-CD-noft-GCN2Conv-dual	0.289	0.293	0.520	0.320	0.532	0.238	0.507
	GNN-CD-noft-APPNP	0.581	0.582	0.403	0.352	0.315	0.290	0.511
	GNN-CD-noft-TAGConv	0.585	0.598	0.380	0.354	0.274	0.344	0.501
	GNN-CD-ft-GCN2Conv-same	0.287	0.294	0.560	0.299	0.608	0.241	0.506
	GNN-CD-ft-GCN2Conv-dual	0.607	0.620	0.365	0.368	0.236	**0.471**	0.516
	GNN-CD-ft-APPNP	0.331	0.314	0.466	0.339	0.429	0.244	0.512
	GNN-CD-ft-TAGConv	0.324	0.352	0.502	0.310	0.508	0.280	0.499
	ETTIN	0.297	0.290	0.442	0.321	0.405	0.226	0.491
	ETTIN-HGA	0.285	0.280	0.516	0.316	0.527	0.229	0.497
MMLU-Pro	YVCE	0.199	0.274	0.560	0.662	0.288	0.713	0.646
	TLCC-MLP	**0.027**	0.203	0.663	0.674	0.697	0.792	0.743
	TLCC-CONV	0.036	0.203	0.661	0.670	0.699	0.794	0.741
	TLCC-LSTM	0.055	0.206	0.660	0.684	0.647	0.793	0.741
	P(IK)	0.099	0.224	0.612	0.726	0.214	0.760	0.719
	PHSV-half	0.168	0.254	0.576	0.675	0.278	0.701	0.689
	PHSV	0.185	0.262	0.553	0.657	0.249	0.696	0.683
	SFHS-MLP	0.079	0.212	0.655	0.686	0.619	0.796	0.741
	SFHS-Conv	0.049	0.204	0.683	0.711	0.648	0.799	0.750
	SFHS-LSTM	0.080	0.213	0.661	0.700	0.587	0.799	0.740
	GNN-SB-GAT	0.057	0.206	0.642	0.685	0.560	0.796	0.727
	GNN-SB-GCN	0.076	0.215	0.644	0.683	0.578	0.776	0.711

(Continued on next page)

(Continued from previous page)

Dataset	Method	ECE↓	Brier↓	Acc↑	F1↑	Spec↑	AUCPR↑	AUROC↑
	GNN-SB-GraphSAGE	0.053	0.208	0.645	0.685	0.575	0.794	0.732
	CE-DT	0.117	0.249	0.597	0.678	0.360	0.722	0.673
	CE-KNN	0.125	0.243	0.621	0.673	0.486	0.715	0.680
	CE-LogReg	0.105	0.248	0.523	0.686	0.002	0.751	0.710
	CE-RF	0.093	0.220	0.620	0.695	0.392	0.756	0.722
	CE-XGB	0.098	0.221	0.620	0.691	0.407	0.755	0.718
	LateFusion-noft-MLP	0.029	0.205	0.673	0.682	0.716	0.788	0.736
	LateFusion-noft-Conv	0.069	0.207	0.677	0.645	0.853	0.802	0.750
	LateFusion-noft-LSTM	0.097	0.214	0.657	0.697	0.586	0.798	0.735
	LateFusion-ft-MLP	0.077	0.212	0.669	0.691	0.660	0.785	0.740
	LateFusion-ft-Conv	0.232	0.268	0.585	0.428	0.952	0.762	0.707
	LateFusion-ft-LSTM	0.088	0.218	0.650	0.678	0.625	0.780	0.723
	GNN-SR-GINE	0.059	0.205	0.668	0.688	0.669	0.797	0.745
	GNN-SR-NNConv	0.130	0.242	0.579	0.400	**0.975**	0.792	0.730
	GNN-SR-Transformer	0.078	0.203	0.675	0.701	0.652	**0.815**	0.758
	GNN-CD-noft-GCN2Conv-same	0.282	0.309	0.625	0.722	0.306	0.760	0.708
	GNN-CD-noft-GCN2Conv-dual	0.058	0.205	0.661	0.692	0.620	0.804	0.739
	GNN-CD-noft-APPNP	0.330	0.337	0.646	0.712	0.463	0.759	0.707
	GNN-CD-noft-TAGConv	0.266	0.323	0.628	0.707	0.397	0.777	0.700
	GNN-CD-ft-GCN2Conv-same	0.105	0.214	0.681	0.700	0.682	0.796	0.740
	GNN-CD-ft-GCN2Conv-dual	0.438	0.441	0.543	0.685	0.105	0.745	0.560
	GNN-CD-ft-APPNP	0.033	0.218	0.663	0.701	0.592	0.737	0.712
	GNN-CD-ft-TAGConv	0.173	0.268	0.640	0.695	0.512	0.739	0.675
	ETTIN	0.062	0.201	**0.693**	0.744	0.549	0.797	0.759
	ETTIN-HGA	0.058	**0.196**	0.693	**0.745**	0.541	0.814	**0.771**

Table 33: Performance metrics for `Qwen3-8B` showing results per method within each test dataset. Each metric value represents the performance of the specified method on the specified dataset for this LLM. **Bold** entries mark the best-performing method for each metric within each dataset.

Dataset	Method	ECE↓	Brier↓	Acc↑	F1↑	Spec↑	AUCPR↑	AUROC↑
BBH	YVCE	**0.018**	0.140	0.824	0.903	0.032	0.897	0.637
	TLCC-MLP	0.173	0.181	0.724	0.824	0.424	0.883	0.654
	TLCC-CONV	0.139	0.179	0.732	0.831	0.391	0.888	0.650
	TLCC-LSTM	0.027	0.137	0.814	0.894	0.171	0.892	0.697
	P(IK)	0.034	0.142	0.821	0.901	0.012	0.881	0.634
	PHSV-half	0.095	0.187	0.741	0.846	0.127	0.833	0.672
	PHSV	0.060	0.173	0.754	0.854	0.140	0.849	0.685
	SFHS-MLP	0.072	0.135	0.815	0.891	0.359	0.915	0.742
	SFHS-Conv	0.068	0.137	0.808	0.885	0.364	0.910	0.727
	SFHS-LSTM	0.078	0.147	0.799	0.882	0.276	0.901	0.701
	GNN-SB-GAT	0.083	0.151	0.801	0.884	0.248	0.882	0.660
	GNN-SB-GCN	0.032	0.131	0.823	0.900	0.173	0.897	0.719
	GNN-SB-GraphSAGE	0.036	**0.125**	**0.830**	0.903	0.225	0.918	**0.762**
	CE-DT	0.093	0.174	0.772	0.867	0.118	0.882	0.681
	CE-KNN	0.081	0.166	0.777	0.873	0.039	0.875	0.692
	CE-LogReg	0.076	0.170	0.774	0.873	0.000	0.873	0.710
	CE-RF	0.067	0.160	0.776	0.873	0.015	0.879	0.722
	CE-XGB	0.072	0.163	0.781	0.874	0.089	0.874	0.710
	LateFusion-noft-MLP	0.083	0.137	0.807	0.883	0.450	0.910	0.751
	LateFusion-noft-Conv	0.125	0.162	0.786	0.870	0.387	0.897	0.693
	LateFusion-noft-LSTM	0.149	0.181	0.772	0.864	0.263	0.884	0.644
	LateFusion-ft-MLP	0.161	0.161	0.773	0.856	0.568	0.911	0.747
	LateFusion-ft-Conv	0.074	0.148	0.805	0.887	0.225	0.895	0.667
	LateFusion-ft-LSTM	0.151	0.186	0.757	0.853	0.306	0.893	0.655
	GNN-SR-GINE	0.426	0.320	0.303	0.275	**0.959**	0.920	0.742
	GNN-SR-NNConv	0.311	0.231	0.774	0.858	0.495	0.926	0.760
	GNN-SR-Transformer	0.080	0.140	0.814	0.891	0.300	0.897	0.709
	GNN-CD-noft-GCN2Conv-same	0.372	0.350	0.568	0.670	0.725	0.904	0.676
	GNN-CD-noft-GCN2Conv-dual	0.039	0.139	0.803	0.886	0.233	0.914	0.727
	GNN-CD-noft-APPNP	0.248	0.248	0.749	0.842	0.446	0.920	0.652
	GNN-CD-noft-TAGConv	0.157	0.382	0.606	0.727	0.461	0.879	0.560
	GNN-CD-ft-GCN2Conv-same	0.137	0.162	0.742	0.832	0.590	0.907	0.733
	GNN-CD-ft-GCN2Conv-dual	0.146	0.167	0.764	0.852	0.483	0.897	0.703
	GNN-CD-ft-APPNP	0.376	0.372	0.614	0.717	0.699	0.910	0.654
	GNN-CD-ft-TAGConv	0.212	0.270	0.728	0.821	0.589	**0.926**	0.687
	ETTIN	0.033	0.131	0.829	**0.903**	0.158	0.910	0.721
	ETTIN-HGA	0.034	0.134	0.822	0.902	0.040	0.894	0.690
FinQA	YVCE	0.403	0.398	0.382	0.545	0.019	0.372	0.500
	TLCC-MLP	0.238	0.314	0.476	0.522	0.300	0.366	0.516
	TLCC-CONV	0.252	0.312	0.500	0.528	0.350	0.389	0.544
	TLCC-LSTM	0.368	0.378	0.404	0.542	0.084	0.391	0.535
	P(IK)	0.277	0.309	0.367	0.525	0.029	0.423	0.523
	PHSV-half	0.241	0.297	0.497	0.494	0.381	0.399	0.577
	PHSV	0.232	0.292	0.483	0.497	0.346	0.365	0.545
	SFHS-MLP	0.185	0.292	0.556	0.436	0.610	0.395	0.554
	SFHS-Conv	0.159	0.267	0.589	0.493	0.618	0.421	0.593
	SFHS-LSTM	0.164	0.268	0.582	0.423	0.681	0.414	0.560
	GNN-SB-GAT	0.162	0.276	0.539	0.453	0.553	0.385	0.540
	GNN-SB-GCN	0.068	0.240	0.603	0.361	0.780	0.425	0.586
	GNN-SB-GraphSAGE	0.099	0.240	0.606	0.485	0.669	0.462	0.617
	CE-DT	0.341	0.347	0.477	0.535	0.269	0.390	0.583
	CE-KNN	0.365	0.357	0.386	0.521	0.079	0.438	0.596
	CE-LogReg	0.456	0.427	0.347	0.515	0.000	0.457	0.633
	CE-RF	0.373	0.357	0.357	0.516	0.021	0.455	0.626
	CE-XGB	0.341	0.348	0.419	0.521	0.158	0.399	0.569
	LateFusion-noft-MLP	0.160	0.265	0.572	0.206	0.820	0.371	0.528
	LateFusion-noft-Conv	0.273	0.332	0.558	0.407	0.644	0.396	0.549
	LateFusion-noft-LSTM	0.447	0.465	0.425	0.496	0.226	0.385	0.516
	LateFusion-ft-MLP	0.145	0.262	0.541	0.368	0.647	0.373	0.530
	LateFusion-ft-Conv	0.177	0.268	0.549	0.501	0.512	0.438	0.588
	LateFusion-ft-LSTM	0.309	0.365	0.491	0.495	0.384	0.385	0.530
	GNN-SR-GINE	0.101	0.240	**0.630**	0.000	**1.000**	0.395	0.561
	GNN-SR-NNConv	**0.067**	**0.235**	0.613	0.295	0.845	0.408	0.559
	GNN-SR-Transformer	0.103	0.254	0.573	0.455	0.627	0.404	0.577
	GNN-CD-noft-GCN2Conv-same	0.352	0.360	0.630	0.132	0.955	0.430	0.581
	GNN-CD-noft-GCN2Conv-dual	0.101	0.245	0.593	0.345	0.771	0.416	0.590
	GNN-CD-noft-APPNP	0.378	0.377	0.622	0.310	0.853	0.464	0.543
	GNN-CD-noft-TAGConv	0.325	0.433	0.535	0.436	0.565	0.452	0.520
	GNN-CD-ft-GCN2Conv-same	0.109	0.247	0.614	0.075	0.950	0.389	0.550
	GNN-CD-ft-GCN2Conv-dual	0.184	0.291	0.539	0.402	0.610	0.392	0.532
	GNN-CD-ft-APPNP	0.418	0.420	0.575	0.323	0.752	0.463	0.544
	GNN-CD-ft-TAGConv	0.247	0.392	0.607	0.110	0.926	0.413	0.539
	ETTIN	0.236	0.275	0.535	**0.559**	0.382	0.498	**0.650**
	ETTIN-HGA	0.230	0.274	0.445	0.551	0.167	**0.529**	0.645
LegalBench	YVCE	0.168	0.260	0.642	0.780	0.023	0.746	0.506
	TLCC-MLP	0.098	0.230	0.621	0.723	0.387	0.734	0.588
	TLCC-CONV	0.134	0.238	0.633	0.739	0.347	0.740	0.595
	TLCC-LSTM	0.116	0.229	0.658	0.776	0.200	0.741	0.598
	P(IK)	0.151	0.243	0.545	0.643	0.415	0.737	0.535

(Continued on next page)

(Continued from previous page)

Dataset	Method	ECE↓	Brier↓	Acc↑	F1↑	Spec↑	AUCPR↑	AUROC↑
	PHSV-half	0.157	0.273	0.541	0.645	0.352	0.663	0.508
	PHSV	0.158	0.272	0.546	0.655	0.336	0.680	0.528
	SFHS-MLP	0.280	0.317	0.529	0.573	0.658	0.734	0.598
	SFHS-Conv	0.196	0.274	0.588	0.664	0.558	0.740	0.603
	SFHS-LSTM	0.215	0.283	0.567	0.647	0.525	0.727	0.578
	GNN-SB-GAT	0.184	0.262	0.611	0.710	0.414	0.728	0.581
	GNN-SB-GCN	0.193	0.266	0.581	0.661	0.530	0.725	0.590
	GNN-SB-GraphSAGE	0.196	0.265	0.592	0.667	0.566	0.750	0.608
	CE-DT	0.092	0.237	0.652	0.768	0.219	0.689	0.557
	CE-KNN	0.088	0.236	0.655	0.786	0.057	0.693	0.560
	CE-LogReg	0.148	0.245	0.655	0.792	0.000	0.697	0.562
	CE-RF	0.078	0.226	0.656	0.791	0.012	0.706	0.587
	CE-XGB	**0.074**	0.229	0.652	0.773	0.165	0.700	0.576
	LateFusion-noft-MLP	0.318	0.340	0.486	0.487	0.747	0.732	0.593
	LateFusion-noft-Conv	0.221	0.281	0.623	0.728	0.370	0.722	0.580
	LateFusion-noft-LSTM	0.265	0.312	0.591	0.697	0.373	0.705	0.547
	LateFusion-ft-MLP	0.323	0.338	0.486	0.484	0.755	0.736	0.603
	LateFusion-ft-Conv	0.139	0.242	0.646	0.754	0.326	0.725	0.581
	LateFusion-ft-LSTM	0.270	0.321	0.566	0.647	0.517	0.718	0.568
	GNN-SR-GINE	0.358	0.342	0.325	0.004	**0.998**	0.744	0.603
	GNN-SR-NNConv	0.277	0.292	0.488	0.497	0.725	0.737	0.592
	GNN-SR-Transformer	0.153	0.247	0.597	0.676	0.544	0.742	0.604
	GNN-CD-noft-GCN2Conv-same	0.577	0.569	0.390	0.247	0.894	0.739	0.582
	GNN-CD-noft-GCN2Conv-dual	0.173	0.264	0.567	0.656	0.479	0.726	0.571
	GNN-CD-noft-APPNP	0.513	0.513	0.485	0.497	0.710	0.758	0.555
	GNN-CD-noft-TAGConv	0.328	0.427	0.524	0.582	0.593	0.714	0.558
	GNN-CD-ft-GCN2Conv-same	0.366	0.363	0.433	0.368	0.826	0.728	0.591
	GNN-CD-ft-GCN2Conv-dual	0.291	0.331	0.544	0.598	0.627	0.721	0.583
	GNN-CD-ft-APPNP	0.550	0.550	0.445	0.400	0.801	0.746	0.548
	GNN-CD-ft-TAGConv	0.404	0.528	0.468	0.452	0.765	**0.764**	0.590
	ETTIN	0.133	0.232	0.664	0.784	0.163	0.756	0.607
	ETTIN-HGA	0.101	**0.224**	**0.674**	**0.803**	0.035	0.762	**0.615**
MATH	YVCE	0.111	0.068	0.945	0.972	0.005	0.966	0.596
	TLCC-MLP	0.066	0.068	0.921	0.958	0.117	0.954	0.578
	TLCC-CONV	0.054	0.058	0.936	0.967	0.066	0.950	0.556
	TLCC-LSTM	0.079	0.065	0.927	0.962	0.103	0.954	0.599
	P(IK)	0.046	0.055	**0.945**	**0.972**	0.000	0.962	0.632
	PHSV-half	0.180	0.226	0.728	0.835	0.132	0.796	0.639
	PHSV	0.120	0.183	0.770	0.861	0.232	0.851	0.682
	SFHS-MLP	0.051	0.057	0.933	0.965	0.131	0.970	0.689
	SFHS-Conv	0.040	0.061	0.925	0.961	0.042	0.967	0.679
	SFHS-LSTM	0.043	0.059	0.933	0.965	0.070	0.966	0.651
	GNN-SB-GAT	0.026	0.051	0.943	0.971	0.023	0.967	0.697
	GNN-SB-GCN	**0.016**	0.052	0.944	0.971	0.000	0.965	0.672
	GNN-SB-GraphSAGE	0.027	**0.051**	0.945	0.972	0.009	0.970	0.683
	CE-DT	0.106	0.140	0.826	0.899	0.271	0.933	0.744
	CE-KNN	0.091	0.126	0.819	0.898	0.114	0.940	0.815
	CE-LogReg	0.103	0.138	0.804	0.891	0.000	0.946	0.849
	CE-RF	0.079	0.122	0.805	0.892	0.014	0.947	**0.852**
	CE-XGB	0.107	0.136	0.811	0.894	0.084	0.940	0.832
	LateFusion-noft-MLP	0.023	0.051	0.943	0.971	0.009	0.969	0.688
	LateFusion-noft-Conv	0.141	0.133	0.827	0.904	0.305	0.967	0.665
	LateFusion-noft-LSTM	0.094	0.090	0.890	0.941	0.169	0.956	0.597
	LateFusion-ft-MLP	0.058	0.060	0.941	0.970	0.089	0.967	0.680
	LateFusion-ft-Conv	0.105	0.074	0.915	0.955	0.113	0.966	0.661
	LateFusion-ft-LSTM	0.050	0.056	0.939	0.969	0.042	0.966	0.650
	GNN-SR-GINE	0.308	0.173	0.716	0.830	0.404	0.954	0.581
	GNN-SR-NNConv	0.396	0.208	0.937	0.967	0.085	0.966	0.672
	GNN-SR-Transformer	0.113	0.068	0.932	0.964	0.099	0.972	0.697
	GNN-CD-noft-GCN2Conv-same	0.126	0.116	0.860	0.923	0.343	**0.976**	0.634
	GNN-CD-noft-GCN2Conv-dual	0.109	0.090	0.878	0.935	0.174	0.967	0.648
	GNN-CD-noft-APPNP	0.074	0.073	0.925	0.961	0.174	0.975	0.575
	GNN-CD-noft-TAGConv	0.093	0.441	0.558	0.705	**0.502**	0.962	0.528
	GNN-CD-ft-GCN2Conv-same	0.044	0.059	0.925	0.961	0.188	0.966	0.677
	GNN-CD-ft-GCN2Conv-dual	0.053	0.058	0.939	0.969	0.042	0.967	0.668
	GNN-CD-ft-APPNP	0.196	0.193	0.797	0.885	0.268	0.971	0.565
	GNN-CD-ft-TAGConv	0.067	0.070	0.929	0.963	0.117	0.975	0.560
	ETTIN	0.019	0.055	0.937	0.967	0.023	0.966	0.639
	ETTIN-HGA	0.113	0.070	0.945	0.972	0.000	0.969	0.670
MedMCQA	YVCE	0.536	0.493	0.283	**0.440**	0.002	0.307	0.514
	TLCC-MLP	0.276	0.313	0.469	0.384	0.422	0.280	0.504
	TLCC-CONV	0.308	0.345	0.466	0.389	0.413	0.286	0.514
	TLCC-LSTM	0.406	0.401	0.382	0.424	0.216	0.279	0.506
	P(IK)	0.580	0.542	0.282	0.440	0.000	0.272	0.482
	PHSV-half	0.541	0.512	0.298	0.401	0.085	0.244	0.482
	PHSV	0.465	0.441	0.313	0.418	0.091	0.258	0.485
	SFHS-MLP	0.375	0.403	0.460	0.406	0.384	0.291	0.526
	SFHS-Conv	0.405	0.425	0.432	0.411	0.326	0.290	0.520
	SFHS-LSTM	0.415	0.423	0.406	0.415	0.272	0.284	0.510
	GNN-SB-GAT	0.423	0.428	0.402	0.423	0.254	0.289	0.523
	GNN-SB-GCN	0.450	0.439	0.341	0.423	0.138	0.284	0.504
	GNN-SB-GraphSAGE	0.501	0.478	0.336	0.438	0.107	0.297	0.529
	CE-DT	0.528	0.510	0.321	0.422	0.101	0.342	0.511

(Continued on next page)

(Continued from previous page)

Dataset	**Method**	**ECE↓**	**Brier↓**	**Acc↑**	**F1↑**	**Spec↑**	**AUCPR↑**	**AUROC↑**
	CE-KNN	0.516	0.485	0.291	0.432	0.030	0.313	0.518
	CE-LogReg	0.550	0.502	0.276	0.433	0.000	0.311	**0.547**
	CE-RF	0.525	0.486	0.280	0.433	0.006	0.290	0.522
	CE-XGB	0.535	0.506	0.298	0.431	0.045	0.279	0.509
	LateFusion-noft-MLP	0.384	0.399	0.424	0.412	0.308	0.291	0.516
	LateFusion-noft-Conv	0.457	0.468	0.420	0.409	0.305	0.290	0.511
	LateFusion-noft-LSTM	0.551	0.555	0.356	0.422	0.170	0.278	0.500
	LateFusion-ft-MLP	0.298	0.338	0.494	0.395	0.457	0.290	0.521
	LateFusion-ft-Conv	0.384	0.381	0.382	0.419	0.221	0.283	0.514
	LateFusion-ft-LSTM	0.512	0.515	0.383	0.422	0.219	0.290	0.519
	GNN-SR-GINE	**0.066**	**0.208**	**0.716**	0.008	**0.996**	0.287	0.509
	GNN-SR-NNConv	0.208	0.248	0.455	0.392	0.389	0.285	0.507
	GNN-SR-Transformer	0.383	0.376	0.400	0.431	0.240	0.294	0.531
	GNN-CD-noft-GCN2Conv-same	0.439	0.458	0.515	0.351	0.535	0.389	0.502
	GNN-CD-noft-GCN2Conv-dual	0.479	0.469	0.361	0.433	0.163	0.283	0.515
	GNN-CD-noft-APPNP	0.571	0.572	0.425	0.416	0.306	**0.520**	0.514
	GNN-CD-noft-TAGConv	0.448	0.525	0.461	0.378	0.414	0.392	0.498
	GNN-CD-ft-GCN2Conv-same	0.299	0.364	0.518	0.378	0.517	0.289	0.521
	GNN-CD-ft-GCN2Conv-dual	0.335	0.368	0.466	0.396	0.405	0.287	0.511
	GNN-CD-ft-APPNP	0.397	0.401	0.589	0.289	0.705	0.373	0.505
	GNN-CD-ft-TAGConv	0.417	0.460	0.537	0.363	0.565	0.444	0.520
	ETTIN	0.526	0.502	0.336	0.436	0.111	0.284	0.519
	ETTIN-HGA	0.465	0.443	0.330	0.432	0.104	0.292	0.518
MMLU-Pro	YVCE	0.041	0.160	0.792	0.883	0.011	0.875	0.616
	TLCC-MLP	0.175	0.178	0.712	0.797	0.687	0.920	0.769
	TLCC-CONV	0.114	0.163	0.744	0.827	0.623	0.924	0.775
	TLCC-LSTM	0.049	0.142	0.784	0.868	0.356	0.920	0.779
	P(IK)	0.108	0.169	0.790	0.883	0.001	0.897	0.691
	PHSV-half	0.136	0.227	0.685	0.790	0.267	0.761	0.671
	PHSV	0.072	0.203	0.707	0.805	0.313	0.798	0.695
	SFHS-MLP	0.097	0.152	0.778	0.854	0.593	0.933	0.802
	SFHS-Conv	0.075	0.148	0.770	0.851	0.551	0.923	0.785
	SFHS-LSTM	0.075	0.155	0.776	0.859	0.449	0.925	0.774
	GNN-SB-GAT	0.048	0.146	0.785	0.868	0.363	0.919	0.767
	GNN-SB-GCN	**0.022**	0.138	0.795	0.877	0.321	0.927	0.788
	GNN-SB-GraphSAGE	0.041	**0.133**	**0.803**	0.883	0.275	0.938	0.808
	CE-DT	0.097	0.181	0.755	0.844	0.328	0.877	0.741
	CE-KNN	0.079	0.169	0.735	0.841	0.115	0.900	0.783
	CE-LogReg	0.117	0.187	0.719	0.837	0.000	0.907	0.802
	CE-RF	0.081	0.166	0.726	0.839	0.043	0.914	0.811
	CE-XGB	0.092	0.174	0.735	0.839	0.154	0.904	0.783
	LateFusion-noft-MLP	0.077	0.141	0.785	0.860	0.590	**0.941**	**0.816**
	LateFusion-noft-Conv	0.146	0.191	0.748	0.840	0.415	0.887	0.698
	LateFusion-noft-LSTM	0.163	0.195	0.751	0.846	0.324	0.892	0.698
	LateFusion-ft-MLP	0.180	0.174	0.725	0.803	0.771	0.939	0.812
	LateFusion-ft-Conv	0.085	0.163	0.760	0.853	0.303	0.914	0.737
	LateFusion-ft-LSTM	0.133	0.177	0.764	0.853	0.387	0.905	0.730
	GNN-SR-GINE	0.408	0.319	0.349	0.310	**0.964**	0.917	0.775
	GNN-SR-NNConv	0.286	0.231	0.761	0.843	0.579	0.930	0.789
	GNN-SR-Transformer	0.080	0.147	0.780	0.864	0.383	0.925	0.777
	GNN-CD-noft-GCN2Conv-same	0.346	0.332	0.608	0.689	0.828	0.920	0.750
	GNN-CD-noft-GCN2Conv-dual	0.087	0.164	0.746	0.839	0.393	0.914	0.740
	GNN-CD-noft-APPNP	0.269	0.270	0.727	0.815	0.594	0.915	0.703
	GNN-CD-noft-TAGConv	0.162	0.404	0.587	0.690	0.599	0.865	0.593
	GNN-CD-ft-GCN2Conv-same	0.164	0.181	0.697	0.777	0.802	0.933	0.799
	GNN-CD-ft-GCN2Conv-dual	0.119	0.162	0.753	0.835	0.609	0.928	0.783
	GNN-CD-ft-APPNP	0.370	0.366	0.617	0.697	0.847	0.918	0.728
	GNN-CD-ft-TAGConv	0.243	0.300	0.697	0.782	0.744	0.922	0.731
	ETTIN	0.039	0.146	0.775	0.863	0.309	0.921	0.760
	ETTIN-HGA	0.044	0.141	0.801	**0.885**	0.165	0.926	0.778

Table 34: Performance metrics for `Qwen3-14B` showing results per method within each test dataset. Each metric value represents the performance of the specified method on the specified dataset for this LLM. **Bold** entries mark the best-performing method for each metric within each dataset.

Dataset	Method	ECE↓	Brier↓	Acc↑	F1↑	Spec↑	AUCPR↑	AUROC↑
BBH	YVCE	0.078	0.133	**0.833**	**0.908**	0.027	**0.933**	0.686
	TLCC-MLP	0.134	0.154	0.788	0.874	0.308	0.911	0.689
	TLCC-CONV	0.082	0.151	0.791	0.878	0.234	0.908	0.676
	TLCC-LSTM	0.078	0.143	0.796	0.879	0.308	0.923	0.720
	P(IK)	0.114	0.149	0.831	0.908	0.000	0.889	0.643
	PHSV-half	0.057	0.171	0.745	0.847	0.170	0.862	0.690
	PHSV	0.075	0.180	0.737	0.842	0.152	0.848	0.678
	SFHS-MLP	0.078	0.141	0.802	0.885	0.244	0.921	0.724
	SFHS-Conv	0.071	0.137	0.809	0.888	0.281	0.924	0.734
	SFHS-LSTM	0.081	0.150	0.791	0.878	0.232	0.908	0.687
	GNN-SB-GAT	0.059	0.132	0.807	0.888	0.259	0.928	0.745
	GNN-SB-GCN	0.058	0.134	0.808	0.887	0.307	0.924	0.745
	GNN-SB-GraphSAGE	0.057	0.134	0.809	0.888	0.307	0.926	0.748
	CE-DT	0.097	0.175	0.769	0.866	0.089	0.857	0.662
	CE-KNN	0.079	0.165	0.774	0.872	0.026	0.881	0.701
	CE-LogReg	0.082	0.163	0.775	0.873	0.029	0.884	0.720
	CE-RF	0.067	0.163	0.775	0.872	0.029	0.876	0.698
	CE-XGB	0.076	0.166	0.773	0.870	0.050	0.875	0.696
	LateFusion-noft-MLP	0.080	0.139	0.802	0.881	0.401	0.912	0.725
	LateFusion-noft-Conv	0.226	0.191	0.706	0.807	0.538	0.907	0.697
	LateFusion-noft-LSTM	0.101	0.160	0.776	0.866	0.312	0.907	0.684
	LateFusion-ft-MLP	0.110	0.147	0.794	0.876	0.387	0.901	0.702
	LateFusion-ft-Conv	0.139	0.145	0.813	0.891	0.282	0.924	0.747
	LateFusion-ft-LSTM	0.113	0.162	0.766	0.853	0.497	0.921	0.736
	GNN-SR-GINE	0.058	0.134	0.811	0.891	0.237	0.924	0.736
	GNN-SR-NNConv	0.319	0.230	0.653	0.751	**0.766**	0.926	**0.749**
	GNN-SR-Transformer	0.072	0.129	0.824	0.901	0.175	0.923	0.749
	GNN-CD-noft-GCN2Conv-same	0.132	0.159	0.756	0.846	0.504	0.915	0.721
	GNN-CD-noft-GCN2Conv-dual	0.048	0.131	0.824	0.899	0.230	0.902	0.707
	GNN-CD-noft-APPNP	0.084	0.136	0.811	0.890	0.240	0.915	0.722
	GNN-CD-noft-TAGConv	0.192	0.208	0.724	0.823	0.501	0.905	0.674
	GNN-CD-ft-GCN2Conv-same	0.187	0.193	0.726	0.821	0.577	0.912	0.720
	GNN-CD-ft-GCN2Conv-dual	**0.037**	**0.129**	0.823	0.900	0.166	0.909	0.719
	GNN-CD-ft-APPNP	0.130	0.156	0.770	0.858	0.442	0.899	0.689
	GNN-CD-ft-TAGConv	0.189	0.219	0.737	0.832	0.509	0.916	0.694
	ETTIN	0.069	0.133	0.832	0.908	0.028	0.919	0.727
	ETTIN-HGA	0.062	0.134	0.824	0.902	0.060	0.920	0.715
FinQA	YVCE	0.533	0.509	0.384	0.538	0.041	**0.635**	0.583
	TLCC-MLP	0.267	0.320	0.401	0.502	0.156	0.344	0.492
	TLCC-CONV	0.238	0.306	0.480	0.496	0.349	0.377	0.529
	TLCC-LSTM	0.335	0.368	0.448	0.512	0.246	0.375	0.542
	P(IK)	0.278	0.310	0.411	0.524	0.134	0.409	0.559
	PHSV-half	0.162	0.254	0.531	0.450	0.502	0.368	0.563
	PHSV	0.154	0.247	0.544	0.432	0.548	0.378	0.569
	SFHS-MLP	0.148	0.257	0.637	0.302	0.872	0.421	0.579
	SFHS-Conv	0.203	0.284	0.604	0.352	0.775	0.391	0.553
	SFHS-LSTM	0.196	0.284	0.563	0.380	0.669	0.383	0.539
	GNN-SB-GAT	0.059	0.231	0.644	0.271	0.901	0.434	0.558
	GNN-SB-GCN	0.077	0.238	0.622	0.313	0.837	0.427	0.560
	GNN-SB-GraphSAGE	0.082	0.241	0.602	0.311	0.800	0.414	0.567
	CE-DT	0.356	0.351	0.425	0.495	0.212	0.373	0.559
	CE-KNN	0.366	0.352	0.340	0.487	0.039	0.381	0.582
	CE-LogReg	0.354	0.347	0.411	0.500	0.172	0.378	0.578
	CE-RF	0.365	0.347	0.356	0.495	0.060	0.381	0.590
	CE-XGB	0.364	0.352	0.382	0.501	0.108	0.378	0.570
	LateFusion-noft-MLP	0.110	0.243	0.637	0.289	0.879	0.434	0.569
	LateFusion-noft-Conv	**0.054**	**0.229**	0.628	0.239	0.888	0.422	0.598
	LateFusion-noft-LSTM	0.225	0.297	0.568	0.357	0.700	0.386	0.530
	LateFusion-ft-MLP	0.093	0.240	0.641	0.301	0.879	0.426	0.562
	LateFusion-ft-Conv	0.115	0.245	0.595	0.391	0.725	0.427	0.556
	LateFusion-ft-LSTM	0.159	0.255	0.620	0.236	0.876	0.414	0.566
	GNN-SR-GINE	0.157	0.257	0.572	0.463	0.605	0.436	0.585
	GNN-SR-NNConv	0.063	0.230	0.642	0.021	**0.996**	0.451	0.578
	GNN-SR-Transformer	0.117	0.244	0.581	0.416	0.673	0.441	0.568
	GNN-CD-noft-GCN2Conv-same	0.090	0.241	0.642	0.270	0.898	0.436	0.548
	GNN-CD-noft-GCN2Conv-dual	0.182	0.268	0.608	0.413	0.733	0.426	0.547
	GNN-CD-noft-APPNP	0.140	0.247	0.592	0.365	0.741	0.428	0.529
	GNN-CD-noft-TAGConv	0.155	0.278	0.608	0.300	0.817	0.420	0.552
	GNN-CD-ft-GCN2Conv-same	0.213	0.276	**0.648**	0.122	0.972	0.411	0.535
	GNN-CD-ft-GCN2Conv-dual	0.183	0.264	0.574	0.433	0.642	0.440	0.542
	GNN-CD-ft-APPNP	0.060	0.235	0.641	0.258	0.903	0.417	0.550
	GNN-CD-ft-TAGConv	0.235	0.295	0.627	0.224	0.894	0.426	0.559
	ETTIN	0.291	0.312	0.369	0.530	0.022	0.407	0.555
	ETTIN-HGA	0.201	0.256	0.527	**0.542**	0.384	0.518	**0.656**
LegalBench	YVCE	0.157	0.218	**0.706**	**0.827**	0.004	**0.856**	**0.664**
	TLCC-MLP	0.153	0.235	0.620	0.707	0.549	0.781	0.617
	TLCC-CONV	0.143	0.238	0.634	0.725	0.522	0.778	0.620
	TLCC-LSTM	0.192	0.265	0.608	0.687	0.603	0.799	0.632
	P(IK)	0.172	0.238	0.658	0.791	0.037	0.639	0.382

(Continued on next page)

(Continued from previous page)

Dataset	Method	ECE↓	Brier↓	Acc↑	F1↑	Spec↑	AUCPR↑	AUROC↑
	PHSV-half	0.207	0.263	0.556	0.621	0.616	0.749	0.575
	PHSV	0.219	0.274	0.558	0.626	0.597	0.744	0.567
	SFHS-MLP	0.364	0.355	0.426	0.390	0.824	0.777	0.619
	SFHS-Conv	0.316	0.317	0.458	0.452	0.796	0.778	0.618
	SFHS-LSTM	0.259	0.290	0.583	0.656	0.626	0.771	0.614
	GNN-SB-GAT	0.304	0.297	0.441	0.394	0.881	0.786	0.599
	GNN-SB-GCN	0.309	0.314	0.545	0.590	0.743	0.787	0.620
	GNN-SB-GraphSAGE	0.321	0.310	0.514	0.536	0.795	0.799	0.633
	CE-DT	0.053	0.217	0.662	0.770	0.304	0.718	0.561
	CE-KNN	0.031	0.212	0.678	0.803	0.075	0.749	0.603
	CE-LogReg	0.028	0.211	0.674	0.795	0.130	0.762	0.610
	CE-RF	**0.016**	0.210	0.682	0.807	0.054	0.751	0.607
	CE-XGB	0.033	0.211	0.678	0.798	0.133	0.751	0.604
	LateFusion-noft-MLP	0.304	0.308	0.487	0.498	0.790	0.782	0.629
	LateFusion-noft-Conv	0.309	0.304	0.505	0.536	0.744	0.760	0.600
	LateFusion-noft-LSTM	0.297	0.314	0.475	0.515	0.669	0.750	0.581
	LateFusion-ft-MLP	0.291	0.299	0.546	0.594	0.729	0.778	0.616
	LateFusion-ft-Conv	0.209	0.248	0.601	0.683	0.581	0.787	0.622
	LateFusion-ft-LSTM	0.385	0.378	0.403	0.337	0.854	0.772	0.612
	GNN-SR-GINE	0.182	0.243	0.567	0.629	0.685	0.792	0.630
	GNN-SR-NNConv	0.287	0.285	0.298	0.011	**0.999**	0.806	0.630
	GNN-SR-Transformer	0.265	0.271	0.498	0.515	0.789	0.791	0.633
	GNN-CD-noft-GCN2Conv-same	0.396	0.368	0.350	0.192	0.928	0.770	0.601
	GNN-CD-noft-GCN2Conv-dual	0.223	0.268	0.592	0.662	0.653	0.793	0.635
	GNN-CD-noft-APPNP	0.233	0.262	0.403	0.341	0.844	0.744	0.544
	GNN-CD-noft-TAGConv	0.389	0.360	0.351	0.184	0.943	0.775	0.612
	GNN-CD-ft-GCN2Conv-same	0.433	0.411	0.401	0.324	0.876	0.769	0.610
	GNN-CD-ft-GCN2Conv-dual	0.245	0.271	0.572	0.637	0.669	0.803	0.634
	GNN-CD-ft-APPNP	0.280	0.297	0.478	0.491	0.770	0.765	0.605
	GNN-CD-ft-TAGConv	0.424	0.411	0.393	0.318	0.855	0.766	0.604
	ETTIN	0.063	0.201	0.697	0.812	0.148	0.800	0.643
	ETTIN-HGA	0.049	**0.197**	0.696	0.810	0.151	0.818	0.656
MATH	YVCE	0.020	**0.040**	**0.955**	0.977	0.007	**0.982**	0.614
	TLCC-MLP	0.042	0.051	0.947	0.973	0.049	0.968	0.602
	TLCC-CONV	0.067	0.062	0.924	0.960	0.105	0.967	0.606
	TLCC-LSTM	0.030	0.048	0.945	0.972	0.049	0.959	0.599
	P(IK)	**0.002**	0.043	0.955	**0.977**	0.000	0.964	0.599
	PHSV-half	0.060	0.139	0.826	0.899	0.253	0.892	0.688
	PHSV	0.081	0.149	0.819	0.895	0.230	0.889	0.687
	SFHS-MLP	0.047	0.052	0.939	0.968	0.119	0.973	0.667
	SFHS-Conv	0.036	0.043	0.952	0.975	0.077	0.973	0.694
	SFHS-LSTM	0.038	0.050	0.944	0.971	0.112	0.971	0.657
	GNN-SB-GAT	0.035	0.046	0.953	0.976	0.014	0.974	0.696
	GNN-SB-GCN	0.025	0.043	0.955	0.977	0.042	0.971	0.681
	GNN-SB-GraphSAGE	0.021	0.042	0.954	0.977	0.042	0.969	0.668
	CE-DT	0.066	0.101	0.871	0.929	0.271	0.949	0.716
	CE-KNN	0.026	0.077	0.886	0.939	0.057	0.958	0.812
	CE-LogReg	0.021	0.075	0.897	0.945	0.182	0.965	**0.832**
	CE-RF	0.044	0.081	0.888	0.940	0.073	0.958	0.806
	CE-XGB	0.024	0.083	0.887	0.940	0.084	0.955	0.790
	LateFusion-noft-MLP	0.069	0.051	0.941	0.969	0.126	0.975	0.696
	LateFusion-noft-Conv	0.164	0.087	0.900	0.946	0.336	0.972	0.668
	LateFusion-noft-LSTM	0.045	0.056	0.930	0.963	0.189	0.973	0.665
	LateFusion-ft-MLP	0.146	0.069	0.927	0.962	0.238	0.973	0.688
	LateFusion-ft-Conv	0.147	0.075	0.915	0.955	0.203	0.975	0.696
	LateFusion-ft-LSTM	0.071	0.051	0.940	0.969	0.175	0.972	0.697
	GNN-SR-GINE	0.081	0.053	0.941	0.970	0.049	0.969	0.646
	GNN-SR-NNConv	0.389	0.194	0.836	0.909	**0.420**	0.974	0.687
	GNN-SR-Transformer	0.140	0.063	0.948	0.973	0.077	0.975	0.690
	GNN-CD-noft-GCN2Conv-same	0.100	0.059	0.931	0.964	0.322	0.971	0.681
	GNN-CD-noft-GCN2Conv-dual	0.009	0.042	0.954	0.976	0.007	0.975	0.706
	GNN-CD-noft-APPNP	0.092	0.051	0.954	0.976	0.056	0.971	0.678
	GNN-CD-noft-TAGConv	0.124	0.101	0.872	0.931	0.294	0.970	0.624
	GNN-CD-ft-GCN2Conv-same	0.124	0.073	0.902	0.947	0.350	0.974	0.687
	GNN-CD-ft-GCN2Conv-dual	0.101	0.052	0.955	0.977	0.000	0.972	0.669
	GNN-CD-ft-APPNP	0.160	0.068	0.937	0.967	0.210	0.971	0.666
	GNN-CD-ft-TAGConv	0.082	0.086	0.907	0.950	0.259	0.977	0.624
	ETTIN	0.035	0.043	0.955	0.977	0.000	0.972	0.664
	ETTIN-HGA	0.060	0.051	0.953	0.976	0.014	0.968	0.632
MedMCQA	YVCE	0.556	0.523	0.300	0.458	0.006	**0.463**	**0.533**
	TLCC-MLP	0.288	0.323	0.445	0.403	0.366	0.293	0.502
	TLCC-CONV	0.368	0.382	0.407	0.426	0.265	0.295	0.505
	TLCC-LSTM	0.304	0.342	0.460	0.399	0.398	0.292	0.500
	P(IK)	0.536	0.499	0.297	**0.458**	0.000	0.285	0.491
	PHSV-half	0.463	0.444	0.325	0.438	0.088	0.292	0.503
	PHSV	0.447	0.425	0.320	0.434	0.084	0.286	0.500
	SFHS-MLP	0.504	0.500	0.356	0.449	0.132	0.304	0.510
	SFHS-Conv	0.497	0.499	0.363	0.443	0.156	0.301	0.507
	SFHS-LSTM	0.429	0.439	0.383	0.434	0.208	0.296	0.496
	GNN-SB-GAT	0.451	0.435	0.322	0.447	0.067	0.293	0.501
	GNN-SB-GCN	0.453	0.451	0.363	0.439	0.161	0.300	0.510
	GNN-SB-GraphSAGE	0.475	0.467	0.346	0.449	0.115	0.298	0.504
	CE-DT	0.535	0.518	0.317	0.446	0.059	0.315	0.510

(Continued on next page)

(Continued from previous page)

Dataset	Method	ECE↓	Brier↓	Acc↑	F1↑	Spec↑	AUCPR↑	AUROC↑
	CE-KNN	0.524	0.498	0.297	0.448	0.016	0.318	0.514
	CE-LogReg	0.541	0.509	0.296	0.448	0.015	0.294	0.511
	CE-RF	0.526	0.496	0.300	0.449	0.022	0.299	0.514
	CE-XGB	0.531	0.504	0.305	0.448	0.031	0.293	0.508
	LateFusion-noft-MLP	0.399	0.414	0.414	0.434	0.269	0.314	0.519
	LateFusion-noft-Conv	0.194	0.267	0.513	0.366	0.530	0.297	0.510
	LateFusion-noft-LSTM	0.439	0.446	0.384	0.430	0.216	0.299	0.499
	LateFusion-ft-MLP	0.363	0.380	0.423	0.427	0.296	0.306	0.513
	LateFusion-ft-Conv	0.347	0.339	0.340	0.451	0.099	0.299	0.511
	LateFusion-ft-LSTM	0.359	0.390	0.454	0.409	0.376	0.296	0.507
	GNN-SR-GINE	0.456	0.439	0.346	0.445	0.118	0.296	0.509
	GNN-SR-NNConv	**0.189**	**0.249**	**0.535**	0.341	**0.590**	0.288	0.494
	GNN-SR-Transformer	0.431	0.405	0.319	0.453	0.053	0.309	0.515
	GNN-CD-noft-GCN2Conv-same	0.316	0.348	0.470	0.415	0.402	0.312	0.520
	GNN-CD-noft-GCN2Conv-dual	0.429	0.429	0.370	0.443	0.171	0.308	0.517
	GNN-CD-noft-APPNP	0.392	0.381	0.343	0.447	0.111	0.310	0.518
	GNN-CD-noft-TAGConv	0.353	0.378	0.476	0.409	0.419	0.336	0.520
	GNN-CD-ft-GCN2Conv-same	0.288	0.338	0.510	0.388	0.506	0.303	0.514
	GNN-CD-ft-GCN2Conv-dual	0.451	0.433	0.335	0.450	0.089	0.305	0.511
	GNN-CD-ft-APPNP	0.314	0.354	0.470	0.413	0.405	0.303	0.517
	GNN-CD-ft-TAGConv	0.410	0.419	0.493	0.406	0.455	0.348	0.515
	ETTIN	0.573	0.553	0.304	0.458	0.015	0.294	0.505
	ETTIN-HGA	0.538	0.520	0.308	0.455	0.027	0.300	0.510
MMLU-Pro	YVCE	0.085	0.145	0.797	0.887	0.016	**0.941**	0.792
	TLCC-MLP	0.181	0.173	0.730	0.813	0.688	0.927	0.782
	TLCC-CONV	0.167	0.179	0.733	0.818	0.644	0.919	0.761
	TLCC-LSTM	0.151	0.173	0.736	0.819	0.682	0.928	0.780
	P(IK)	0.071	0.154	0.794	0.885	0.007	0.909	0.717
	PHSV-half	0.051	0.184	0.718	0.820	0.264	0.842	0.708
	PHSV	0.075	0.195	0.707	0.813	0.236	0.815	0.683
	SFHS-MLP	0.075	0.152	0.784	0.866	0.406	0.920	0.763
	SFHS-Conv	0.058	0.141	0.795	0.873	0.435	0.929	0.786
	SFHS-LSTM	0.099	0.168	0.755	0.843	0.461	0.914	0.741
	GNN-SB-GAT	0.067	0.141	0.781	0.866	0.360	0.936	0.790
	GNN-SB-GCN	0.054	0.142	0.787	0.868	0.432	0.929	0.785
	GNN-SB-GraphSAGE	0.052	0.140	0.788	0.868	0.442	0.935	0.793
	CE-DT	0.083	0.181	0.755	0.851	0.213	0.840	0.686
	CE-KNN	0.056	0.157	0.753	0.856	0.070	0.903	0.778
	CE-LogReg	0.061	0.158	0.764	0.860	0.171	0.914	0.788
	CE-RF	0.049	0.160	0.756	0.858	0.091	0.898	0.767
	CE-XGB	0.061	0.164	0.755	0.856	0.112	0.890	0.755
	LateFusion-noft-MLP	0.092	0.148	0.770	0.850	0.589	0.933	0.793
	LateFusion-noft-Conv	0.245	0.207	0.672	0.759	0.764	0.925	0.767
	LateFusion-noft-LSTM	0.102	0.171	0.754	0.841	0.487	0.915	0.748
	LateFusion-ft-MLP	0.130	0.157	0.758	0.840	0.608	0.923	0.777
	LateFusion-ft-Conv	0.152	0.167	0.765	0.854	0.383	0.928	0.766
	LateFusion-ft-LSTM	0.170	0.185	0.727	0.810	0.708	0.934	0.793
	GNN-SR-GINE	0.055	0.146	0.786	0.872	0.295	0.922	0.763
	GNN-SR-NNConv	0.298	0.236	0.606	0.684	**0.872**	0.927	0.764
	GNN-SR-Transformer	0.070	0.148	0.792	0.879	0.196	0.918	0.754
	GNN-CD-noft-GCN2Conv-same	0.182	0.180	0.709	0.793	0.736	0.927	0.774
	GNN-CD-noft-GCN2Conv-dual	0.045	0.139	0.795	0.873	0.433	0.931	0.793
	GNN-CD-noft-APPNP	0.101	0.148	0.782	0.863	0.449	0.928	0.782
	GNN-CD-noft-TAGConv	0.239	0.246	0.679	0.770	0.691	0.901	0.716
	GNN-CD-ft-GCN2Conv-same	0.265	0.233	0.672	0.756	0.796	0.930	0.784
	GNN-CD-ft-GCN2Conv-dual	0.050	0.137	0.801	0.881	0.331	0.929	0.788
	GNN-CD-ft-APPNP	0.173	0.175	0.719	0.804	0.701	0.913	0.764
	GNN-CD-ft-TAGConv	0.256	0.269	0.686	0.772	0.760	0.916	0.742
	ETTIN	0.036	**0.134**	**0.805**	**0.888**	0.167	0.934	**0.796**
	ETTIN-HGA	**0.034**	0.138	0.805	0.885	0.261	0.927	0.781

Table 35: Performance metrics for `Magistral-Small-2506` showing results per method within each test dataset. Each metric value represents the performance of the specified method on the specified dataset for this LLM. **Bold** entries mark the best-performing method for each metric within each dataset.

Dataset	Method	ECE↓	Brier↓	Acc↑	F1↑	Spec↑	AUCPR↑	AUROC↑
BBH	YVCE	0.507	0.517	0.398	0.546	0.059	0.598	0.466
	TLCC-MLP	0.041	0.173	0.764	0.669	0.839	0.676	0.800
	TLCC-CONV	0.055	0.173	0.765	0.670	0.840	0.686	0.805
	TLCC-LSTM	0.045	0.170	0.766	0.677	0.833	0.697	0.807
	P(IK)	0.075	0.217	0.649	0.608	0.608	0.511	0.693
	PHSV-half	0.087	0.214	0.696	0.517	0.817	0.537	0.675
	PHSV	0.068	0.209	0.685	0.510	0.798	0.537	0.681
	SFHS-MLP	0.036	0.183	0.739	0.631	0.829	0.681	0.774
	SFHS-Conv	0.044	0.185	0.734	0.643	0.795	0.680	0.778
	SFHS-LSTM	0.053	0.194	0.718	0.601	0.815	0.628	0.755
	GNN-SB-GAT	0.038	0.186	0.724	0.634	0.780	0.651	0.776
	GNN-SB-GCN	0.036	0.181	0.735	0.647	0.792	0.669	0.787
	GNN-SB-GraphSAGE	**0.012**	0.176	0.744	0.643	0.829	0.691	0.794
	CE-DT	0.127	0.240	0.636	0.535	0.653	0.489	0.652
	CE-KNN	0.103	0.224	0.664	0.532	0.725	0.517	0.680
	CE-LogReg	0.058	0.211	0.687	0.476	0.835	0.522	0.665
	CE-RF	0.037	0.205	0.700	0.514	0.828	0.528	0.689
	CE-XGB	0.095	0.219	0.664	0.532	0.725	0.530	0.691
	LateFusion-noft-MLP	0.034	0.183	0.731	0.599	0.854	0.660	0.783
	LateFusion-noft-Conv	0.108	0.227	0.640	0.462	0.782	0.559	0.655
	LateFusion-noft-LSTM	0.080	0.204	0.709	0.583	0.814	0.624	0.739
	LateFusion-ft-MLP	0.054	0.185	0.730	0.580	0.874	0.655	0.778
	LateFusion-ft-Conv	0.110	0.217	0.698	0.570	0.802	0.597	0.691
	LateFusion-ft-LSTM	0.050	0.194	0.720	0.601	0.818	0.628	0.754
	GNN-SR-GINE	0.126	0.208	0.734	0.631	0.816	0.657	0.773
	GNN-SR-NNConv	0.127	0.202	0.717	0.659	0.715	0.670	0.780
	GNN-SR-Transformer	0.119	0.192	0.736	0.583	**0.888**	0.685	0.796
	GNN-CD-noft-GCN2Conv-same	0.026	0.194	0.707	0.593	0.796	0.612	0.750
	GNN-CD-noft-GCN2Conv-dual	0.045	0.188	0.728	0.621	0.812	0.643	0.770
	GNN-CD-noft-APPNP	0.036	0.200	0.691	0.566	0.788	0.590	0.737
	GNN-CD-noft-TAGConv	0.200	0.306	0.643	0.523	0.720	0.515	0.591
	GNN-CD-ft-GCN2Conv-same	0.108	0.205	0.671	0.649	0.591	0.636	0.765
	GNN-CD-ft-GCN2Conv-dual	0.086	0.193	0.723	0.553	0.888	0.660	0.782
	GNN-CD-ft-APPNP	0.086	0.214	0.666	0.553	0.739	0.570	0.716
	GNN-CD-ft-TAGConv	0.309	0.328	0.669	0.600	0.678	0.601	0.683
	ETTIN	0.046	**0.163**	**0.769**	**0.690**	0.823	**0.739**	**0.834**
	ETTIN-HGA	0.049	0.177	0.747	0.642	0.838	0.685	0.803
FinQA	YVCE	0.860	0.809	0.080	0.065	0.050	**0.410**	0.449
	TLCC-MLP	0.129	0.055	0.952	0.258	0.974	0.124	0.626
	TLCC-CONV	0.125	0.058	0.950	0.200	0.974	0.097	0.608
	TLCC-LSTM	0.101	0.050	0.946	0.188	0.970	0.094	0.651
	P(IK)	**0.028**	**0.032**	0.965	0.000	**0.998**	0.155	**0.718**
	PHSV-half	0.274	0.120	0.933	0.115	0.962	0.058	0.501
	PHSV	0.218	0.096	0.933	0.115	0.962	0.057	0.576
	SFHS-MLP	0.122	0.054	0.964	0.242	0.991	0.186	0.562
	SFHS-Conv	0.116	0.051	0.959	**0.263**	0.985	0.176	0.640
	SFHS-LSTM	0.106	0.050	0.958	0.121	0.988	0.124	0.625
	GNN-SB-GAT	0.141	0.061	0.958	0.121	0.988	0.122	0.609
	GNN-SB-GCN	0.130	0.058	0.956	0.167	0.985	0.100	0.608
	GNN-SB-GraphSAGE	0.104	0.045	0.966	0.207	0.995	0.198	0.709
	CE-DT	0.219	0.098	0.872	0.043	0.900	0.044	0.544
	CE-KNN	0.222	0.100	0.905	0.085	0.933	0.051	0.507
	CE-LogReg	0.273	0.118	0.934	0.082	0.965	0.042	0.562
	CE-RF	0.223	0.087	0.955	0.162	0.985	0.061	0.567
	CE-XGB	0.230	0.100	0.901	0.081	0.929	0.045	0.519
	LateFusion-noft-MLP	0.105	0.047	0.966	0.148	0.997	0.128	0.617
	LateFusion-noft-Conv	0.260	0.137	0.780	0.062	0.799	0.075	0.502
	LateFusion-noft-LSTM	0.108	0.055	0.945	0.136	0.973	0.111	0.658
	LateFusion-ft-MLP	0.079	0.042	0.965	0.143	0.995	0.132	0.546
	LateFusion-ft-Conv	0.220	0.105	0.895	0.122	0.918	0.099	0.562
	LateFusion-ft-LSTM	0.175	0.067	0.946	0.140	0.974	0.071	0.655
	GNN-SR-GINE	0.368	0.166	0.942	0.200	0.967	0.181	0.685
	GNN-SR-NNConv	0.309	0.130	0.928	0.109	0.956	0.144	0.657
	GNN-SR-Transformer	0.264	0.100	**0.968**	0.154	0.998	0.208	0.677
	GNN-CD-noft-GCN2Conv-same	0.180	0.078	0.921	0.069	0.950	0.067	0.540
	GNN-CD-noft-GCN2Conv-dual	0.149	0.063	0.956	0.118	0.986	0.122	0.568
	GNN-CD-noft-APPNP	0.198	0.081	0.949	0.103	0.979	0.057	0.569
	GNN-CD-noft-TAGConv	0.216	0.133	0.873	0.155	0.891	0.073	0.531
	GNN-CD-ft-GCN2Conv-same	0.125	0.064	0.924	0.103	0.952	0.093	0.626
	GNN-CD-ft-GCN2Conv-dual	0.200	0.080	0.966	0.148	0.997	0.156	0.612
	GNN-CD-ft-APPNP	0.194	0.086	0.895	0.077	0.921	0.045	0.541
	GNN-CD-ft-TAGConv	0.067	0.067	0.931	0.113	0.959	0.054	0.572
	ETTIN	0.081	0.038	0.962	0.235	0.989	0.190	0.697
	ETTIN-HGA	0.131	0.050	0.964	0.194	0.992	0.165	0.709
LegalBench	YVCE	0.190	0.266	**0.595**	0.694	0.301	**0.753**	**0.661**
	TLCC-MLP	0.204	0.288	0.561	**0.700**	0.109	0.604	0.571
	TLCC-CONV	0.224	0.300	0.563	0.699	0.124	0.605	0.570
	TLCC-LSTM	0.204	0.292	0.563	0.699	0.123	0.599	0.567
	P(IK)	0.052	0.252	0.557	0.686	0.161	0.561	0.522

(Continued on next page)

(Continued from previous page)

Dataset	**Method**	**ECE↓**	**Brier↓**	**Acc↑**	**F1↑**	**Spec↑**	**AUCPR↑**	**AUROC↑**
	PHSV-half	0.128	0.268	0.554	0.482	0.649	0.528	0.566
	PHSV	0.085	0.259	0.565	0.486	0.673	0.533	0.575
	SFHS-MLP	0.085	0.256	0.565	0.597	0.540	0.640	0.589
	SFHS-Conv	0.102	0.254	0.582	0.625	0.520	0.653	0.608
	SFHS-LSTM	0.146	0.277	0.543	0.587	0.485	0.607	0.556
	GNN-SB-GAT	0.081	0.253	0.564	0.600	0.530	0.635	0.589
	GNN-SB-GCN	0.083	0.252	0.568	0.603	0.531	0.642	0.595
	GNN-SB-GraphSAGE	0.100	0.256	0.570	0.585	0.595	0.637	0.597
	CE-DT	0.199	0.301	0.547	0.530	0.546	0.526	0.557
	CE-KNN	0.183	0.289	0.554	0.506	0.609	0.517	0.562
	CE-LogReg	0.110	0.263	0.559	0.424	**0.744**	0.526	0.565
	CE-RF	0.139	0.268	0.557	0.466	0.682	0.520	0.564
	CE-XGB	0.172	0.282	0.552	0.508	0.602	0.521	0.565
	LateFusion-noft-MLP	0.143	0.272	0.550	0.511	0.699	0.630	0.590
	LateFusion-noft-Conv	0.148	0.275	0.523	0.448	0.734	0.612	0.567
	LateFusion-noft-LSTM	0.171	0.283	0.554	0.568	0.579	0.619	0.578
	LateFusion-ft-MLP	0.183	0.286	0.534	0.472	0.724	0.620	0.583
	LateFusion-ft-Conv	0.154	0.283	0.540	0.511	0.665	0.619	0.569
	LateFusion-ft-LSTM	0.149	0.268	0.563	0.560	0.634	0.636	0.595
	GNN-SR-GINE	0.055	0.248	0.557	0.654	0.309	0.615	0.567
	GNN-SR-NNConv	**0.028**	0.242	0.576	0.634	0.462	0.637	0.597
	GNN-SR-Transformer	0.084	0.247	0.551	0.503	0.721	0.641	0.599
	GNN-CD-noft-GCN2Conv-same	0.119	0.266	0.550	0.558	0.594	0.610	0.568
	GNN-CD-noft-GCN2Conv-dual	0.142	0.269	0.544	0.513	0.673	0.617	0.581
	GNN-CD-noft-APPNP	0.137	0.271	0.528	0.496	0.657	0.595	0.560
	GNN-CD-noft-TAGConv	0.235	0.361	0.533	0.531	0.598	0.606	0.528
	GNN-CD-ft-GCN2Conv-same	0.153	0.275	0.561	0.606	0.497	0.619	0.582
	GNN-CD-ft-GCN2Conv-dual	0.073	0.249	0.542	0.492	0.711	0.622	0.581
	GNN-CD-ft-APPNP	0.204	0.296	0.526	0.469	0.703	0.602	0.562
	GNN-CD-ft-TAGConv	0.443	0.459	0.535	0.652	0.222	0.626	0.534
	ETTIN	0.048	**0.240**	0.590	0.674	0.369	0.664	0.625
	ETTIN-HGA	0.076	0.253	0.574	0.640	0.433	0.621	0.588
MATH	YVCE	0.874	0.831	0.078	0.113	0.020	**0.530**	0.512
	TLCC-MLP	0.108	0.057	0.949	0.457	0.985	0.393	0.806
	TLCC-CONV	0.124	0.066	0.948	0.477	0.981	0.412	0.848
	TLCC-LSTM	0.095	0.058	0.938	0.438	0.969	0.410	0.796
	P(IK)	**0.015**	0.049	0.941	0.039	**0.999**	0.262	0.807
	PHSV-half	0.257	0.128	0.908	0.207	0.955	0.193	0.606
	PHSV	0.198	0.107	0.909	0.244	0.953	0.195	0.631
	SFHS-MLP	0.141	0.070	0.939	0.382	0.978	0.369	0.777
	SFHS-Conv	0.122	0.068	0.929	0.344	0.967	0.299	0.777
	SFHS-LSTM	0.128	0.073	0.916	0.326	0.951	0.254	0.759
	GNN-SB-GAT	0.140	0.074	0.928	0.312	0.968	0.279	0.788
	GNN-SB-GCN	0.126	0.067	0.931	0.328	0.972	0.338	0.837
	GNN-SB-GraphSAGE	0.111	0.061	0.941	0.368	0.981	0.363	0.823
	CE-DT	0.214	0.114	0.831	0.194	0.864	0.153	0.608
	CE-KNN	0.212	0.112	0.874	0.194	0.915	0.168	0.675
	CE-LogReg	0.256	0.133	0.909	0.161	0.959	0.140	0.512
	CE-RF	0.206	0.099	0.929	0.211	0.980	0.177	0.644
	CE-XGB	0.222	0.112	0.882	0.216	0.922	0.192	0.680
	LateFusion-noft-MLP	0.126	0.064	0.944	0.300	0.990	0.339	0.836
	LateFusion-noft-Conv	0.255	0.159	0.771	0.080	0.808	0.071	0.414
	LateFusion-noft-LSTM	0.129	0.078	0.915	0.305	0.952	0.280	0.728
	LateFusion-ft-MLP	0.101	0.059	0.944	0.286	0.991	0.341	0.830
	LateFusion-ft-Conv	0.237	0.137	0.857	0.144	0.898	0.103	0.476
	LateFusion-ft-LSTM	0.182	0.086	0.917	0.263	0.959	0.244	0.742
	GNN-SR-GINE	0.348	0.169	0.924	0.367	0.958	0.345	0.781
	GNN-SR-NNConv	0.309	0.144	0.903	0.368	0.929	0.363	0.831
	GNN-SR-Transformer	0.264	0.116	0.946	0.320	0.991	0.382	**0.855**
	GNN-CD-noft-GCN2Conv-same	0.172	0.090	0.898	0.190	0.942	0.162	0.771
	GNN-CD-noft-GCN2Conv-dual	0.181	0.088	0.919	0.325	0.955	0.298	0.821
	GNN-CD-noft-APPNP	0.208	0.101	0.896	0.222	0.936	0.191	0.776
	GNN-CD-noft-TAGConv	0.199	0.143	0.856	0.220	0.888	0.163	0.571
	GNN-CD-ft-GCN2Conv-same	0.184	0.107	0.817	0.288	0.828	0.274	0.822
	GNN-CD-ft-GCN2Conv-dual	0.209	0.097	0.939	0.274	0.985	0.333	0.846
	GNN-CD-ft-APPNP	0.213	0.114	0.849	0.229	0.878	0.199	0.790
	GNN-CD-ft-TAGConv	0.138	0.142	0.856	0.262	0.882	0.246	0.679
	ETTIN	0.046	**0.040**	**0.956**	**0.498**	0.992	0.498	0.851
	ETTIN-HGA	0.085	0.048	0.950	0.357	0.994	0.459	0.851
MedMCQA	YVCE	0.512	0.450	0.302	0.396	0.097	0.398	0.571
	TLCC-MLP	0.470	0.422	0.329	0.390	0.150	0.269	0.551
	TLCC-CONV	0.495	0.448	0.331	0.392	0.152	0.271	0.553
	TLCC-LSTM	0.478	0.430	0.325	0.394	0.140	0.269	0.560
	P(IK)	0.562	0.504	0.258	0.392	0.026	0.256	0.540
	PHSV-half	0.460	0.417	0.358	0.376	0.214	0.255	0.544
	PHSV	0.401	0.353	0.358	0.379	0.211	0.265	0.565
	SFHS-MLP	0.486	0.441	0.335	0.397	0.153	0.283	0.574
	SFHS-Conv	0.500	0.455	0.332	0.397	0.147	0.285	0.574
	SFHS-LSTM	0.502	0.458	0.318	0.396	0.126	0.271	0.552
	GNN-SB-GAT	0.496	0.443	0.298	0.392	0.095	0.285	0.570
	GNN-SB-GCN	0.546	0.497	0.286	0.392	0.073	0.281	0.564
	GNN-SB-GraphSAGE	0.528	0.477	0.298	0.394	0.092	0.286	0.571
	CE-DT	0.535	0.492	0.297	0.382	0.102	0.277	0.536

(Continued on next page)

(Continued from previous page)

Dataset	**Method**	**ECE↓**	**Brier↓**	**Acc↑**	**F1↑**	**Spec↑**	**AUCPR↑**	**AUROC↑**
	CE-KNN	0.530	0.480	0.305	0.380	0.120	0.272	0.551
	CE-LogReg	0.434	0.383	0.361	0.384	0.210	0.256	0.554
	CE-RF	0.450	0.394	0.326	0.379	0.155	0.257	0.555
	CE-XGB	0.535	0.480	0.295	0.381	0.101	0.255	0.554
	LateFusion-noft-MLP	0.513	0.459	0.298	0.395	0.091	0.286	0.573
	LateFusion-noft-Conv	0.377	0.348	0.398	0.391	0.269	0.272	0.552
	LateFusion-noft-LSTM	0.520	0.490	0.337	0.392	0.162	0.276	0.555
	LateFusion-ft-MLP	0.507	0.461	0.315	0.395	0.120	0.283	0.568
	LateFusion-ft-Conv	0.449	0.415	0.369	0.392	0.219	0.275	0.556
	LateFusion-ft-LSTM	0.527	0.480	0.313	0.394	0.118	0.273	0.559
	GNN-SR-GINE	0.447	0.390	0.313	0.392	0.122	0.268	0.550
	GNN-SR-NNConv	0.448	0.385	0.281	0.394	0.062	0.284	0.572
	GNN-SR-Transformer	0.371	0.323	0.337	0.393	0.162	0.283	0.570
	GNN-CD-noft-GCN2Conv-same	0.452	0.397	0.296	0.393	0.089	0.279	0.564
	GNN-CD-noft-GCN2Conv-dual	0.479	0.424	0.308	0.397	0.106	0.284	0.570
	GNN-CD-noft-APPNP	0.429	0.374	0.301	0.395	0.095	0.275	0.556
	GNN-CD-noft-TAGConv	0.433	0.437	**0.439**	0.363	**0.368**	0.333	0.509
	GNN-CD-ft-GCN2Conv-same	0.550	0.498	0.277	0.392	0.058	0.289	0.570
	GNN-CD-ft-GCN2Conv-dual	**0.340**	**0.299**	0.342	**0.398**	0.163	0.290	**0.580**
	GNN-CD-ft-APPNP	0.425	0.392	0.370	0.386	0.226	0.277	0.555
	GNN-CD-ft-TAGConv	0.693	0.695	0.305	0.390	0.108	**0.400**	0.513
	ETTIN	0.620	0.577	0.268	0.392	0.041	0.285	0.576
	ETTIN-HGA	0.554	0.506	0.289	0.395	0.074	0.282	0.569
MMLU-Pro	YVCE	0.398	0.426	0.468	0.624	0.049	0.567	0.467
	TLCC-MLP	0.022	0.176	0.738	0.716	0.734	0.729	0.809
	TLCC-CONV	0.046	0.180	0.729	0.710	0.714	0.728	0.804
	TLCC-LSTM	0.035	0.178	0.732	0.719	0.703	0.726	0.807
	P(IK)	0.078	0.181	0.724	0.731	0.631	0.715	0.807
	PHSV-half	0.090	0.213	0.694	0.636	0.731	0.656	0.731
	PHSV	0.038	0.202	0.693	0.627	0.745	0.675	0.742
	SFHS-MLP	0.041	0.181	0.735	0.710	0.738	0.743	0.801
	SFHS-Conv	0.038	0.182	0.733	0.714	0.720	0.747	0.803
	SFHS-LSTM	0.047	0.190	0.721	0.696	0.723	0.726	0.785
	GNN-SB-GAT	0.045	0.181	0.729	0.720	0.684	0.746	0.803
	GNN-SB-GCN	0.066	0.182	0.729	0.726	0.665	0.747	0.806
	GNN-SB-GraphSAGE	0.049	0.179	0.732	0.723	0.687	0.752	0.809
	CE-DT	0.125	0.227	0.673	0.668	0.591	0.632	0.728
	CE-KNN	0.108	0.213	0.693	0.669	0.655	0.657	0.751
	CE-LogReg	0.062	0.206	0.705	0.642	0.757	0.649	0.736
	CE-RF	0.057	0.199	0.706	0.663	0.717	0.661	0.758
	CE-XGB	0.117	0.211	0.691	0.673	0.641	0.668	0.760
	LateFusion-noft-MLP	0.027	0.180	0.729	0.701	0.744	0.743	0.803
	LateFusion-noft-Conv	0.063	0.219	0.646	0.580	0.723	0.659	0.699
	LateFusion-noft-LSTM	0.073	0.198	0.715	0.681	0.739	0.719	0.776
	LateFusion-ft-MLP	0.043	0.185	0.725	0.679	0.784	0.740	0.795
	LateFusion-ft-Conv	0.082	0.212	0.691	0.646	0.736	0.681	0.731
	LateFusion-ft-LSTM	0.064	0.193	0.717	0.691	0.721	0.721	0.781
	GNN-SR-GINE	0.110	0.205	0.722	0.708	0.694	0.737	0.798
	GNN-SR-NNConv	0.104	0.199	0.713	0.719	0.622	0.737	0.796
	GNN-SR-Transformer	0.104	0.192	0.732	0.686	0.791	0.755	0.808
	GNN-CD-noft-GCN2Conv-same	0.027	0.185	0.719	0.703	0.696	0.729	0.791
	GNN-CD-noft-GCN2Conv-dual	**0.021**	0.185	0.725	0.695	0.741	0.742	0.793
	GNN-CD-noft-APPNP	0.036	0.190	0.712	0.686	0.715	0.720	0.783
	GNN-CD-noft-TAGConv	0.215	0.311	0.642	0.597	0.679	0.594	0.608
	GNN-CD-ft-GCN2Conv-same	0.099	0.203	0.678	0.701	0.543	0.726	0.782
	GNN-CD-ft-GCN2Conv-dual	0.096	0.194	0.723	0.666	**0.804**	0.744	0.802
	GNN-CD-ft-APPNP	0.077	0.208	0.684	0.648	0.707	0.681	0.749
	GNN-CD-ft-TAGConv	0.314	0.321	0.676	0.686	0.580	0.657	0.701
	ETTIN	0.070	**0.165**	**0.748**	**0.751**	0.666	**0.782**	**0.842**
	ETTIN-HGA	0.063	0.174	0.737	0.731	0.683	0.765	0.824

Table 36: Performance metrics for `QwQ-32B` showing results per method within each test dataset. Each metric value represents the performance of the specified method on the specified dataset for this LLM. **Bold** entries mark the best-performing method for each metric within each dataset.

Dataset	Method	ECE↓	Brier↓	Acc↑	F1↑	Spec↑	AUCPR↑	AUROC↑
BBH	YVCE	0.047	0.115	0.855	**0.922**	0.025	**0.940**	0.690
	TLCC-MLP	0.211	0.183	0.696	0.806	0.463	0.918	0.663
	TLCC-CONV	0.173	0.191	0.712	0.820	0.370	0.897	0.607
	TLCC-LSTM	0.179	0.175	0.733	0.834	0.404	0.913	0.653
	P(IK)	0.089	0.136	0.813	0.893	0.221	0.914	0.679
	PHSV-half	0.061	0.169	0.763	0.862	0.096	0.845	0.651
	PHSV	0.048	0.164	0.766	0.864	0.120	0.874	0.677
	SFHS-MLP	0.095	0.133	0.819	0.895	0.314	0.931	0.718
	SFHS-Conv	0.056	0.126	0.828	0.902	0.268	0.927	0.707
	SFHS-LSTM	0.088	0.134	0.820	0.896	0.277	0.919	0.687
	GNN-SB-GAT	0.057	0.120	0.838	0.908	0.264	0.930	0.729
	GNN-SB-GCN	0.055	0.117	0.840	0.909	0.254	0.935	0.743
	GNN-SB-GraphSAGE	0.040	0.114	0.850	0.917	0.162	0.931	0.735
	CE-DT	0.082	0.171	0.774	0.869	0.125	0.863	0.666
	CE-KNN	0.062	0.159	0.790	0.880	0.086	0.871	0.671
	CE-LogReg	0.086	0.165	0.790	0.882	0.000	0.877	0.689
	CE-RF	0.037	0.156	0.787	0.879	0.062	0.874	0.678
	CE-XGB	0.082	0.163	0.784	0.875	0.122	0.872	0.673
	LateFusion-noft-MLP	0.152	0.141	0.827	0.899	0.392	0.913	0.703
	LateFusion-noft-Conv	0.171	0.163	0.758	0.852	0.414	0.922	0.679
	LateFusion-noft-LSTM	0.143	0.169	0.767	0.861	0.316	0.909	0.645
	LateFusion-ft-MLP	0.070	0.122	0.834	0.905	0.278	0.928	0.728
	LateFusion-ft-Conv	0.047	0.121	0.848	0.916	0.146	0.927	0.701
	LateFusion-ft-LSTM	0.078	0.129	0.832	0.905	0.231	0.922	0.696
	GNN-SR-GINE	0.075	0.124	0.851	0.918	0.125	0.909	0.658
	GNN-SR-NNConv	**0.030**	**0.112**	0.851	0.917	0.214	0.939	**0.758**
	GNN-SR-Transformer	0.126	0.127	**0.855**	0.920	0.135	0.931	0.737
	GNN-CD-noft-GCN2Conv-same	0.080	0.129	0.826	0.901	0.259	0.915	0.692
	GNN-CD-noft-GCN2Conv-dual	0.176	0.162	0.752	0.845	0.514	0.914	0.695
	GNN-CD-noft-APPNP	0.179	0.164	0.769	0.860	0.397	0.897	0.651
	GNN-CD-noft-TAGConv	0.120	0.150	0.807	0.888	0.291	0.922	0.681
	GNN-CD-ft-GCN2Conv-same	0.175	0.179	0.769	0.862	0.323	0.879	0.595
	GNN-CD-ft-GCN2Conv-dual	0.068	0.117	0.847	0.913	0.319	0.931	0.746
	GNN-CD-ft-APPNP	0.257	0.238	0.697	0.802	**0.572**	0.895	0.665
	GNN-CD-ft-TAGConv	0.215	0.237	0.736	0.837	0.405	0.921	0.654
	ETTIN	0.043	0.117	0.841	0.911	0.170	0.931	0.724
	ETTIN-HGA	0.046	0.119	0.853	0.920	0.030	0.911	0.662
FinQA	YVCE	0.487	0.475	0.420	0.582	0.028	**0.627**	0.560
	TLCC-MLP	0.264	0.319	0.477	0.558	0.248	0.466	0.577
	TLCC-CONV	0.302	0.339	0.486	0.556	0.275	0.489	0.578
	TLCC-LSTM	0.188	0.275	0.542	0.514	0.503	0.471	0.582
	P(IK)	**0.045**	0.240	0.601	0.403	0.782	0.489	0.574
	PHSV-half	0.219	0.276	0.463	0.524	0.262	0.440	0.576
	PHSV	0.191	0.271	0.515	0.515	0.402	0.413	0.571
	SFHS-MLP	0.183	0.292	0.558	0.409	0.678	0.415	0.539
	SFHS-Conv	0.163	0.274	0.566	0.525	0.545	0.492	0.595
	SFHS-LSTM	0.173	0.281	0.562	0.422	0.675	0.455	0.554
	GNN-SB-GAT	0.115	0.261	0.564	0.443	0.656	0.457	0.563
	GNN-SB-GCN	0.065	**0.239**	0.593	0.363	0.800	0.508	0.587
	GNN-SB-GraphSAGE	0.057	0.242	0.593	0.376	0.789	0.478	0.568
	CE-DT	0.348	0.375	0.424	0.519	0.177	0.370	0.532
	CE-KNN	0.349	0.353	0.397	0.530	0.090	0.453	0.590
	CE-LogReg	0.437	0.421	0.359	0.529	0.000	0.430	0.553
	CE-RF	0.340	0.346	0.408	0.530	0.115	0.440	0.588
	CE-XGB	0.323	0.346	0.440	0.523	0.207	0.417	0.571
	LateFusion-noft-MLP	0.087	0.245	0.604	0.091	0.980	0.483	0.574
	LateFusion-noft-Conv	0.173	0.273	0.561	0.525	0.534	0.472	0.588
	LateFusion-noft-LSTM	0.274	0.337	0.502	0.510	0.407	0.427	0.535
	LateFusion-ft-MLP	0.132	0.253	0.610	0.141	0.969	0.482	0.565
	LateFusion-ft-Conv	0.275	0.327	0.516	0.556	0.357	0.497	0.592
	LateFusion-ft-LSTM	0.224	0.299	0.501	0.468	0.471	0.448	0.547
	GNN-SR-GINE	0.219	0.289	0.469	0.571	0.192	0.464	0.563
	GNN-SR-NNConv	0.137	0.259	0.573	0.383	0.739	0.470	0.560
	GNN-SR-Transformer	0.087	0.249	0.562	0.455	0.636	0.469	0.559
	GNN-CD-noft-GCN2Conv-same	0.091	0.250	0.568	0.417	0.693	0.484	0.568
	GNN-CD-noft-GCN2Conv-dual	0.146	0.258	0.603	0.058	**0.991**	0.480	0.568
	GNN-CD-noft-APPNP	0.049	0.240	0.598	0.349	0.821	0.483	0.574
	GNN-CD-noft-TAGConv	0.184	0.318	0.549	0.440	0.623	0.453	0.552
	GNN-CD-ft-GCN2Conv-same	0.402	0.430	0.488	0.526	0.344	0.441	0.541
	GNN-CD-ft-GCN2Conv-dual	0.135	0.255	**0.620**	0.225	0.946	0.492	0.583
	GNN-CD-ft-APPNP	0.369	0.397	0.539	0.432	0.610	0.452	0.536
	GNN-CD-ft-TAGConv	0.344	0.389	0.521	0.493	0.484	0.441	0.514
	ETTIN	0.218	0.277	0.526	0.578	0.336	0.542	0.642
	ETTIN-HGA	0.157	0.252	0.598	**0.586**	0.525	0.511	**0.644**
LegalBench	YVCE	0.177	0.251	0.631	0.768	0.058	**0.781**	**0.612**
	TLCC-MLP	0.102	0.242	0.625	0.740	0.263	0.669	0.550
	TLCC-CONV	0.185	0.265	0.636	0.759	0.178	0.666	0.545
	TLCC-LSTM	0.139	0.253	0.625	0.740	0.257	0.669	0.550
	P(IK)	0.346	0.362	0.470	0.384	**0.863**	0.720	0.580

(Continued on next page)

(Continued from previous page)

Dataset	Method	ECE↓	Brier↓	Acc↑	F1↑	Spec↑	AUCPR↑	AUROC↑
	PHSV-half	0.134	0.265	0.547	0.649	0.331	0.642	0.528
	PHSV	0.110	0.255	0.558	0.663	0.319	0.645	0.532
	SFHS-MLP	0.243	0.305	0.545	0.605	0.557	0.684	0.564
	SFHS-Conv	0.210	0.281	0.595	0.684	0.446	0.699	0.581
	SFHS-LSTM	0.245	0.308	0.553	0.620	0.536	0.687	0.567
	GNN-SB-GAT	0.247	0.310	0.535	0.576	0.622	0.693	0.573
	GNN-SB-GCN	0.197	0.281	0.557	0.616	0.571	0.701	0.583
	GNN-SB-GraphSAGE	0.176	0.268	0.565	0.634	0.533	0.696	0.573
	CE-DT	0.173	0.274	0.580	0.688	0.302	0.660	0.554
	CE-KNN	0.160	0.269	0.580	0.691	0.284	0.649	0.548
	CE-LogReg	0.181	0.269	0.613	0.760	0.000	0.645	0.561
	CE-RF	0.133	0.259	0.587	0.698	0.278	0.649	0.553
	CE-XGB	0.170	0.272	0.582	0.689	0.309	0.648	0.550
	LateFusion-noft-MLP	0.254	0.312	0.524	0.548	0.669	0.703	0.581
	LateFusion-noft-Conv	0.163	0.275	0.586	0.675	0.442	0.695	0.575
	LateFusion-noft-LSTM	0.268	0.324	0.540	0.595	0.574	0.691	0.567
	LateFusion-ft-MLP	0.250	0.308	0.517	0.533	0.686	0.704	0.584
	LateFusion-ft-Conv	0.210	0.275	0.634	0.750	0.237	0.703	0.580
	LateFusion-ft-LSTM	0.251	0.311	0.571	0.654	0.470	0.691	0.564
	GNN-SR-GINE	0.109	0.240	0.609	0.714	0.343	0.695	0.573
	GNN-SR-NNConv	0.202	0.286	0.552	0.606	0.584	0.696	0.579
	GNN-SR-Transformer	0.172	0.264	0.573	0.649	0.506	0.687	0.567
	GNN-CD-noft-GCN2Conv-same	0.252	0.306	0.544	0.599	0.575	0.694	0.571
	GNN-CD-noft-GCN2Conv-dual	0.278	0.325	0.512	0.519	0.708	0.705	0.584
	GNN-CD-noft-APPNP	0.240	0.300	0.552	0.615	0.553	0.683	0.561
	GNN-CD-noft-TAGConv	0.232	0.308	0.561	0.639	0.488	0.679	0.552
	GNN-CD-ft-GCN2Conv-same	0.338	0.357	0.582	0.680	0.389	0.686	0.551
	GNN-CD-ft-GCN2Conv-dual	0.244	0.312	0.550	0.602	0.596	0.704	0.583
	GNN-CD-ft-APPNP	0.401	0.412	0.548	0.607	0.563	0.686	0.558
	GNN-CD-ft-TAGConv	0.276	0.336	0.571	0.651	0.484	0.691	0.559
	ETTIN	0.132	0.245	0.635	0.754	0.217	0.711	0.592
	ETTIN-HGA	**0.095**	**0.234**	**0.646**	**0.776**	0.097	0.711	0.593
MATH	YVCE	0.035	0.067	0.923	0.959	0.201	**0.977**	0.684
	TLCC-MLP	0.059	0.069	0.913	0.954	0.234	0.957	0.671
	TLCC-CONV	0.102	0.095	0.861	0.923	0.399	0.962	0.713
	TLCC-LSTM	0.102	0.099	0.854	0.919	0.327	0.961	0.694
	P(IK)	0.024	0.063	0.932	0.965	0.000	0.961	0.664
	PHSV-half	0.119	0.182	0.771	0.867	0.093	0.836	0.656
	PHSV	0.048	0.143	0.813	0.893	0.179	0.883	0.706
	SFHS-MLP	0.045	0.061	0.929	0.963	0.123	0.966	0.721
	SFHS-Conv	0.025	0.062	0.921	0.958	0.201	0.970	0.746
	SFHS-LSTM	0.042	0.070	0.913	0.954	0.179	0.969	0.729
	GNN-SB-GAT	0.016	0.057	0.932	0.964	0.112	0.974	0.769
	GNN-SB-GCN	0.011	0.056	**0.935**	**0.966**	0.063	0.970	0.763
	GNN-SB-GraphSAGE	0.014	**0.055**	0.935	0.966	0.052	0.974	0.775
	CE-DT	0.066	0.116	0.864	0.924	0.249	0.865	0.733
	CE-KNN	0.058	0.098	0.864	0.925	0.170	0.956	0.829
	CE-LogReg	0.155	0.131	0.844	0.915	0.000	0.946	0.815
	CE-RF	0.044	0.096	0.864	0.925	0.150	0.954	**0.839**
	CE-XGB	0.078	0.106	0.865	0.925	0.201	0.950	0.821
	LateFusion-noft-MLP	0.133	0.077	0.928	0.962	0.224	0.970	0.757
	LateFusion-noft-Conv	0.136	0.099	0.853	0.918	0.448	0.968	0.735
	LateFusion-noft-LSTM	0.081	0.086	0.890	0.941	0.220	0.962	0.693
	LateFusion-ft-MLP	0.021	0.056	0.933	0.965	0.030	0.968	0.754
	LateFusion-ft-Conv	0.021	0.062	0.923	0.960	0.164	0.966	0.738
	LateFusion-ft-LSTM	0.046	0.068	0.916	0.956	0.101	0.964	0.709
	GNN-SR-GINE	0.101	0.071	0.925	0.961	0.067	0.966	0.716
	GNN-SR-NNConv	**0.011**	0.058	0.933	0.965	0.034	0.969	0.731
	GNN-SR-Transformer	0.148	0.081	0.932	0.965	0.019	0.968	0.726
	GNN-CD-noft-GCN2Conv-same	0.022	0.058	0.931	0.964	0.153	0.970	0.745
	GNN-CD-noft-GCN2Conv-dual	0.122	0.081	0.895	0.943	0.388	0.968	0.749
	GNN-CD-noft-APPNP	0.136	0.080	0.912	0.953	0.302	0.968	0.731
	GNN-CD-noft-TAGConv	0.064	0.069	0.913	0.954	0.201	0.970	0.708
	GNN-CD-ft-GCN2Conv-same	0.095	0.094	0.867	0.927	0.336	0.964	0.705
	GNN-CD-ft-GCN2Conv-dual	0.056	0.058	0.933	0.965	0.063	0.969	0.766
	GNN-CD-ft-APPNP	0.168	0.146	0.803	0.886	**0.493**	0.965	0.719
	GNN-CD-ft-TAGConv	0.155	0.158	0.827	0.903	0.362	0.961	0.638
	ETTIN	0.024	0.062	0.930	0.963	0.030	0.968	0.723
	ETTIN-HGA	0.038	0.064	0.930	0.964	0.004	0.966	0.719
MedMCQA	YVCE	0.519	0.481	0.303	0.458	0.012	0.401	**0.546**
	TLCC-MLP	**0.321**	**0.343**	0.445	0.435	0.330	0.303	0.520
	TLCC-CONV	0.400	0.414	0.428	0.445	0.284	0.304	0.522
	TLCC-LSTM	0.342	0.364	0.438	0.438	0.312	0.303	0.521
	P(IK)	0.535	0.500	0.300	0.458	0.005	0.297	0.506
	PHSV-half	0.511	0.490	0.304	0.438	0.046	0.294	0.495
	PHSV	0.477	0.455	0.318	0.444	0.065	0.300	0.509
	SFHS-MLP	0.509	0.509	0.359	0.449	0.140	0.298	0.507
	SFHS-Conv	0.493	0.487	0.363	0.450	0.146	0.294	0.506
	SFHS-LSTM	0.485	0.487	0.371	0.447	0.166	0.288	0.499
	GNN-SB-GAT	0.517	0.508	0.345	0.452	0.107	0.288	0.502
	GNN-SB-GCN	0.465	0.459	0.366	0.453	0.148	0.289	0.502
	GNN-SB-GraphSAGE	0.508	0.492	0.331	0.455	0.073	0.290	0.503
	CE-DT	0.544	0.533	0.323	0.446	0.071	0.298	0.490

(Continued on next page)

(Continued from previous page)

Dataset	Method	ECE↓	Brier↓	Acc↑	F1↑	Spec↑	AUCPR↑	AUROC↑
	CE-KNN	0.532	0.508	0.312	0.450	0.042	0.330	0.519
	CE-LogReg	0.505	0.461	0.292	0.451	0.000	0.302	0.525
	CE-RF	0.526	0.500	0.302	0.449	0.026	0.291	0.507
	CE-XGB	0.550	0.535	0.315	0.446	0.055	0.285	0.498
	LateFusion-noft-MLP	0.405	0.406	0.405	0.451	0.228	0.313	0.533
	LateFusion-noft-Conv	0.377	0.389	0.425	0.439	0.284	0.306	0.517
	LateFusion-noft-LSTM	0.522	0.521	0.355	0.440	0.144	0.287	0.495
	LateFusion-ft-MLP	0.461	0.457	0.374	0.454	0.163	0.288	0.507
	LateFusion-ft-Conv	0.556	0.534	0.318	0.453	0.051	0.296	0.512
	LateFusion-ft-LSTM	0.556	0.548	0.348	0.457	0.105	0.291	0.503
	GNN-SR-GINE	0.479	0.448	0.323	0.455	0.059	0.312	0.523
	GNN-SR-NNConv	0.561	0.541	0.320	0.455	0.050	0.284	0.487
	GNN-SR-Transformer	0.430	0.400	0.312	**0.458**	0.029	0.290	0.494
	GNN-CD-noft-GCN2Conv-same	0.498	0.486	0.351	0.453	0.117	0.316	0.525
	GNN-CD-noft-GCN2Conv-dual	0.350	0.389	**0.459**	0.433	0.359	0.309	0.525
	GNN-CD-noft-APPNP	0.401	0.399	0.398	0.452	0.212	0.327	0.540
	GNN-CD-noft-TAGConv	0.461	0.468	0.402	0.441	0.236	0.342	0.510
	GNN-CD-ft-GCN2Conv-same	0.582	0.580	0.351	0.452	0.118	0.315	0.525
	GNN-CD-ft-GCN2Conv-dual	0.431	0.433	0.399	0.452	0.215	0.297	0.520
	GNN-CD-ft-APPNP	0.469	0.487	0.455	0.423	**0.362**	0.318	0.524
	GNN-CD-ft-TAGConv	0.541	0.536	0.409	0.438	0.253	**0.405**	0.504
	ETTIN	0.530	0.516	0.340	0.456	0.090	0.309	0.523
	ETTIN-HGA	0.554	0.530	0.321	0.457	0.050	0.323	0.539
MMLU-Pro	YVCE	0.034	**0.113**	**0.855**	**0.920**	0.119	**0.952**	0.766
	TLCC-MLP	0.249	0.194	0.670	0.775	0.682	0.936	0.736
	TLCC-CONV	0.241	0.220	0.652	0.762	0.622	0.924	0.692
	TLCC-LSTM	0.251	0.204	0.658	0.766	0.655	0.931	0.713
	P(IK)	0.030	0.121	0.843	0.914	0.081	0.927	0.711
	PHSV-half	0.060	0.175	0.758	0.856	0.174	0.841	0.653
	PHSV	0.060	0.158	0.781	0.867	0.319	0.879	0.704
	SFHS-MLP	0.101	0.135	0.811	0.887	0.448	0.941	0.758
	SFHS-Conv	0.089	0.130	0.810	0.886	0.457	0.946	0.769
	SFHS-LSTM	0.102	0.145	0.791	0.874	0.438	0.939	0.753
	GNN-SB-GAT	0.062	0.123	0.827	0.899	0.340	0.942	0.765
	GNN-SB-GCN	0.076	0.118	0.830	0.901	0.398	0.951	**0.792**
	GNN-SB-GraphSAGE	0.042	0.114	0.841	0.910	0.257	0.945	0.776
	CE-DT	0.063	0.161	0.796	0.879	0.275	0.873	0.700
	CE-KNN	**0.026**	0.139	0.809	0.889	0.202	0.908	0.747
	CE-LogReg	0.086	0.161	0.795	0.886	0.000	0.908	0.750
	CE-RF	0.039	0.139	0.811	0.891	0.180	0.906	0.747
	CE-XGB	0.063	0.149	0.801	0.883	0.233	0.902	0.728
	LateFusion-noft-MLP	0.199	0.159	0.777	0.860	0.580	0.935	0.758
	LateFusion-noft-Conv	0.240	0.199	0.685	0.789	0.639	0.935	0.734
	LateFusion-noft-LSTM	0.155	0.176	0.758	0.852	0.410	0.917	0.684
	LateFusion-ft-MLP	0.134	0.139	0.798	0.877	0.527	0.946	0.775
	LateFusion-ft-Conv	0.093	0.149	0.796	0.879	0.371	0.935	0.727
	LateFusion-ft-LSTM	0.105	0.145	0.801	0.882	0.372	0.933	0.730
	GNN-SR-GINE	0.106	0.131	0.823	0.899	0.239	0.930	0.714
	GNN-SR-NNConv	0.041	0.122	0.828	0.901	0.287	0.941	0.754
	GNN-SR-Transformer	0.142	0.135	0.839	0.910	0.193	0.939	0.743
	GNN-CD-noft-GCN2Conv-same	0.081	0.124	0.831	0.902	0.384	0.936	0.754
	GNN-CD-noft-GCN2Conv-dual	0.260	0.205	0.647	0.753	**0.728**	0.937	0.757
	GNN-CD-noft-APPNP	0.190	0.160	0.765	0.854	0.513	0.925	0.724
	GNN-CD-noft-TAGConv	0.144	0.161	0.789	0.873	0.411	0.917	0.683
	GNN-CD-ft-GCN2Conv-same	0.162	0.178	0.752	0.849	0.379	0.907	0.661
	GNN-CD-ft-GCN2Conv-dual	0.149	0.142	0.801	0.878	0.560	0.946	0.780
	GNN-CD-ft-APPNP	0.329	0.307	0.611	0.722	0.695	0.918	0.704
	GNN-CD-ft-TAGConv	0.248	0.262	0.706	0.810	0.526	0.914	0.657
	ETTIN	0.078	0.126	0.818	0.894	0.324	0.941	0.752
	ETTIN-HGA	0.041	0.117	0.842	0.912	0.179	0.936	0.742

Table 37: Performance metrics for `EXAONE-Deep-32B` showing results per method within each test dataset. Each metric value represents the performance of the specified method on the specified dataset for this LLM. **Bold** entries mark the best-performing method for each metric within each dataset.

Dataset	Method	ECE↓	Brier↓	Acc↑	F1↑	Spec↑	AUCPR↑	AUROC↑
BBH	YVCE	0.056	0.145	0.813	0.892	0.228	0.928	0.727
	TLCC-MLP	0.089	0.141	0.808	0.887	0.327	0.902	0.685
	TLCC-CONV	0.057	0.134	0.817	0.894	0.259	0.915	0.702
	TLCC-LSTM	0.046	0.125	**0.841**	**0.911**	0.153	0.923	0.729
	P(IK)	0.049	0.140	0.831	0.907	0.019	0.870	0.596
	PHSV-half	0.264	0.240	0.644	0.742	0.526	0.783	0.633
	PHSV	0.264	0.242	0.635	0.731	0.542	0.769	0.623
	SFHS-MLP	0.051	0.133	0.820	0.897	0.201	0.921	0.728
	SFHS-Conv	0.055	0.147	0.810	0.891	0.195	0.881	0.621
	SFHS-LSTM	0.084	0.159	0.808	0.889	0.244	0.862	0.589
	GNN-SB-GAT	0.062	0.142	0.816	0.896	0.162	0.897	0.671
	GNN-SB-GCN	0.021	0.131	0.836	0.908	0.145	0.898	0.684
	GNN-SB-GraphSAGE	0.058	0.145	0.805	0.887	0.212	0.891	0.656
	CE-DT	0.120	0.233	0.731	0.833	0.241	0.814	0.581
	CE-KNN	0.038	0.172	0.769	0.863	0.178	0.833	0.668
	CE-LogReg	**0.015**	0.187	0.749	0.857	0.000	0.849	0.693
	CE-RF	0.052	0.171	0.774	0.864	0.226	0.850	0.681
	CE-XGB	0.044	0.169	0.771	0.864	0.172	0.846	0.683
	LateFusion-noft-MLP	0.087	0.151	0.814	0.893	0.198	0.908	0.727
	LateFusion-noft-Conv	0.207	0.201	0.715	0.818	0.437	0.895	0.672
	LateFusion-noft-LSTM	0.119	0.161	0.811	0.892	0.206	0.889	0.660
	LateFusion-ft-MLP	0.092	0.149	0.811	0.891	0.224	0.924	**0.757**
	LateFusion-ft-Conv	0.628	0.614	0.324	0.369	**0.752**	0.847	0.571
	LateFusion-ft-LSTM	0.124	0.175	0.797	0.882	0.215	0.875	0.634
	GNN-SR-GINE	0.096	0.148	0.808	0.890	0.182	0.888	0.643
	GNN-SR-NNConv	0.217	0.181	0.777	0.866	0.325	0.894	0.655
	GNN-SR-Transformer	0.075	0.142	0.815	0.894	0.196	0.891	0.649
	GNN-CD-noft-GCN2Conv-same	0.070	0.155	0.806	0.888	0.212	0.902	0.685
	GNN-CD-noft-GCN2Conv-dual	0.091	0.162	0.795	0.879	0.283	**0.936**	0.756
	GNN-CD-noft-APPNP	0.072	0.172	0.805	0.888	0.210	0.878	0.647
	GNN-CD-noft-TAGConv	0.335	0.261	0.404	0.495	0.670	0.836	0.526
	GNN-CD-ft-GCN2Conv-same	0.172	0.188	0.803	0.886	0.205	0.851	0.579
	GNN-CD-ft-GCN2Conv-dual	0.062	0.154	0.811	0.892	0.188	0.903	0.708
	GNN-CD-ft-APPNP	0.177	0.194	0.793	0.879	0.236	0.849	0.576
	GNN-CD-ft-TAGConv	0.535	0.560	0.422	0.526	0.607	0.832	0.472
	ETTIN	0.023	**0.124**	0.838	0.910	0.119	0.918	0.725
	ETTIN-HGA	0.034	0.127	0.835	0.908	0.106	0.914	0.714
FinQA	YVCE	0.400	0.404	0.490	0.615	0.146	0.546	0.577
	TLCC-MLP	0.129	0.243	**0.615**	0.588	0.582	0.573	0.673
	TLCC-CONV	0.239	0.292	0.571	0.630	0.353	0.537	0.650
	TLCC-LSTM	0.259	0.304	0.543	0.610	0.318	0.511	0.632
	P(IK)	0.099	0.253	0.517	0.484	0.505	0.479	0.547
	PHSV-half	0.434	0.470	0.407	0.351	0.358	0.257	0.414
	PHSV	0.427	0.471	0.394	0.347	0.344	0.260	0.394
	SFHS-MLP	0.105	0.255	0.574	0.427	0.723	0.506	0.559
	SFHS-Conv	0.146	0.262	0.562	0.581	0.450	0.506	0.614
	SFHS-LSTM	0.230	0.311	0.515	0.541	0.398	0.455	0.546
	GNN-SB-GAT	0.134	0.265	0.557	0.546	0.505	0.487	0.590
	GNN-SB-GCN	0.145	0.268	0.539	0.502	0.535	0.495	0.561
	GNN-SB-GraphSAGE	0.071	0.250	0.569	0.263	0.857	0.465	0.556
	CE-DT	0.252	0.353	0.545	0.563	0.413	0.546	0.598
	CE-KNN	0.246	0.310	0.506	0.546	0.344	0.468	0.591
	CE-LogReg	0.340	0.353	0.391	0.562	0.000	0.464	0.604
	CE-RF	0.186	0.272	0.517	0.534	0.393	0.451	0.594
	CE-XGB	0.185	0.281	0.517	0.526	0.407	0.453	0.582
	LateFusion-noft-MLP	0.348	0.380	0.543	0.580	0.395	0.519	0.602
	LateFusion-noft-Conv	0.367	0.390	0.532	0.571	0.386	0.490	0.585
	LateFusion-noft-LSTM	0.361	0.392	0.515	0.498	0.477	0.439	0.536
	LateFusion-ft-MLP	0.334	0.368	0.569	0.515	0.593	0.499	0.569
	LateFusion-ft-Conv	0.429	0.435	0.518	0.561	0.368	0.454	0.543
	LateFusion-ft-LSTM	0.296	0.350	0.527	0.506	0.495	0.447	0.540
	GNN-SR-GINE	0.124	0.247	0.590	0.604	0.483	0.532	0.638
	GNN-SR-NNConv	**0.064**	**0.242**	0.564	0.094	0.942	0.493	0.598
	GNN-SR-Transformer	0.153	0.268	0.541	0.585	0.380	0.507	0.586
	GNN-CD-noft-GCN2Conv-same	0.202	0.285	0.558	0.421	0.693	0.503	0.569
	GNN-CD-noft-GCN2Conv-dual	0.221	0.297	0.534	0.443	0.608	0.477	0.565
	GNN-CD-noft-APPNP	0.338	0.360	0.508	0.545	0.371	0.513	0.580
	GNN-CD-noft-TAGConv	0.174	0.282	0.583	0.125	**0.964**	0.412	0.438
	GNN-CD-ft-GCN2Conv-same	0.339	0.366	0.517	0.588	0.298	0.488	0.583
	GNN-CD-ft-GCN2Conv-dual	0.384	0.398	0.524	0.599	0.292	0.503	0.593
	GNN-CD-ft-APPNP	0.268	0.327	0.567	0.426	0.708	0.503	0.568
	GNN-CD-ft-TAGConv	0.403	0.448	0.529	0.489	0.529	0.498	0.506
	ETTIN	0.203	0.260	0.518	0.627	0.198	**0.600**	**0.696**
	ETTIN-HGA	0.130	0.244	0.569	**0.631**	0.350	0.564	0.674
LegalBench	YVCE	0.132	0.214	0.686	0.791	0.343	**0.859**	**0.688**
	TLCC-MLP	0.135	0.210	0.678	0.776	0.449	0.788	0.620
	TLCC-CONV	0.081	0.199	0.721	0.826	0.222	0.783	0.602
	TLCC-LSTM	0.136	0.216	0.683	0.787	0.363	0.761	0.592
	P(IK)	0.150	0.209	0.671	0.768	0.463	0.821	0.648

(Continued on next page)

(Continued from previous page)

Dataset	Method	ECE↓	Brier↓	Acc↑	F1↑	Spec↑	AUCPR↑	AUROC↑
	PHSV-half	0.234	0.282	0.512	0.576	0.611	0.703	0.528
	PHSV	0.253	0.288	0.499	0.558	0.618	0.699	0.529
	SFHS-MLP	0.136	0.233	0.695	0.816	0.070	0.721	0.501
	SFHS-Conv	0.146	0.227	0.697	0.818	0.067	0.715	0.486
	SFHS-LSTM	0.230	0.273	0.558	0.697	0.190	0.712	0.457
	GNN-SB-GAT	0.180	0.229	0.710	0.827	0.066	0.734	0.517
	GNN-SB-GCN	0.171	0.233	0.698	0.819	0.062	0.750	0.537
	GNN-SB-GraphSAGE	0.213	0.240	0.686	0.809	0.081	0.714	0.462
	CE-DT	0.162	0.280	0.665	0.789	0.131	0.717	0.517
	CE-KNN	0.156	0.245	0.667	0.797	0.053	0.717	0.539
	CE-LogReg	**0.030**	0.208	0.705	0.827	0.000	0.695	0.543
	CE-RF	0.147	0.237	0.667	0.794	0.088	0.693	0.523
	CE-XGB	0.145	0.240	0.665	0.795	0.055	0.697	0.527
	LateFusion-noft-MLP	0.099	0.227	0.696	0.816	0.072	0.781	0.579
	LateFusion-noft-Conv	0.219	0.241	0.650	0.745	0.517	0.823	0.647
	LateFusion-noft-LSTM	0.107	0.224	0.688	0.805	0.166	0.772	0.576
	LateFusion-ft-MLP	0.209	0.262	0.614	0.714	0.492	0.777	0.573
	LateFusion-ft-Conv	0.629	0.621	0.334	0.196	**0.942**	0.780	0.557
	LateFusion-ft-LSTM	0.112	0.237	0.680	0.804	0.084	0.687	0.397
	GNN-SR-GINE	0.103	0.207	0.732	**0.843**	0.041	0.690	0.445
	GNN-SR-NNConv	0.254	0.305	0.379	0.461	0.420	0.664	0.372
	GNN-SR-Transformer	0.154	0.221	0.690	0.812	0.078	0.730	0.479
	GNN-CD-noft-GCN2Conv-same	0.195	0.245	0.692	0.815	0.061	0.754	0.540
	GNN-CD-noft-GCN2Conv-dual	0.251	0.276	0.523	0.606	0.583	0.774	0.552
	GNN-CD-noft-APPNP	0.161	0.243	0.686	0.810	0.060	0.760	0.535
	GNN-CD-noft-TAGConv	0.129	0.226	0.638	0.747	0.384	0.777	0.580
	GNN-CD-ft-GCN2Conv-same	0.093	0.223	0.700	0.821	0.057	0.773	0.545
	GNN-CD-ft-GCN2Conv-dual	0.095	0.221	0.702	0.821	0.075	0.780	0.581
	GNN-CD-ft-APPNP	0.100	0.241	0.676	0.802	0.069	0.753	0.512
	GNN-CD-ft-TAGConv	0.193	0.289	0.668	0.777	0.328	0.753	0.562
	ETTIN	0.031	**0.186**	0.731	0.840	0.081	0.824	0.653
	ETTIN-HGA	0.113	0.200	**0.733**	0.840	0.122	0.814	0.645
MATH	YVCE	0.045	0.081	**0.904**	**0.947**	0.389	0.959	0.764
	TLCC-MLP	0.079	0.093	0.895	0.944	0.115	0.911	0.614
	TLCC-CONV	0.055	0.085	0.899	0.946	0.119	0.946	0.738
	TLCC-LSTM	0.070	0.089	0.901	0.947	0.088	0.950	0.741
	P(IK)	0.051	0.087	0.895	0.945	0.000	0.950	0.746
	PHSV-half	0.262	0.186	0.841	0.903	0.452	0.762	0.603
	PHSV	0.280	0.205	0.817	0.889	0.355	0.715	0.516
	SFHS-MLP	0.083	0.098	0.837	0.904	0.626	0.952	0.788
	SFHS-Conv	0.168	0.127	0.804	0.883	0.620	0.960	0.780
	SFHS-LSTM	0.213	0.152	0.817	0.891	0.677	0.958	0.782
	GNN-SB-GAT	0.076	0.101	0.853	0.916	0.515	0.952	0.773
	GNN-SB-GCN	0.070	0.087	0.875	0.930	0.437	0.956	0.786
	GNN-SB-GraphSAGE	0.144	0.123	0.802	0.880	0.725	0.956	0.778
	CE-DT	0.095	0.163	0.822	0.887	0.565	0.907	0.728
	CE-KNN	0.040	0.093	0.882	0.925	0.667	0.948	**0.880**
	CE-LogReg	0.045	0.169	0.778	0.875	0.000	0.931	0.865
	CE-RF	0.038	0.094	0.881	0.925	0.685	0.943	0.879
	CE-XGB	0.031	0.093	0.881	0.925	0.647	0.948	0.879
	LateFusion-noft-MLP	0.111	0.127	0.851	0.915	0.536	0.930	0.721
	LateFusion-noft-Conv	0.194	0.186	0.766	0.857	0.608	0.946	0.743
	LateFusion-noft-LSTM	0.174	0.161	0.822	0.895	0.608	0.952	0.755
	LateFusion-ft-MLP	0.118	0.127	0.846	0.911	0.539	0.935	0.737
	LateFusion-ft-Conv	0.478	0.465	0.473	0.623	0.356	0.882	0.453
	LateFusion-ft-LSTM	0.218	0.189	0.799	0.879	0.674	0.951	0.755
	GNN-SR-GINE	0.165	0.121	0.808	0.886	0.620	0.948	0.758
	GNN-SR-NNConv	0.249	0.150	0.812	0.887	0.665	0.952	0.767
	GNN-SR-Transformer	0.167	0.119	0.828	0.899	0.635	0.954	0.775
	GNN-CD-noft-GCN2Conv-same	0.156	0.152	0.801	0.880	0.716	0.957	0.781
	GNN-CD-noft-GCN2Conv-dual	0.143	0.167	0.804	0.883	0.656	0.955	0.781
	GNN-CD-noft-APPNP	0.243	0.197	0.801	0.879	0.719	0.953	0.768
	GNN-CD-noft-TAGConv	0.434	0.277	0.294	0.359	**0.913**	0.937	0.671
	GNN-CD-ft-GCN2Conv-same	0.339	0.231	0.803	0.881	0.704	0.950	0.778
	GNN-CD-ft-GCN2Conv-dual	0.130	0.153	0.825	0.897	0.590	0.944	0.748
	GNN-CD-ft-APPNP	0.344	0.239	0.791	0.872	0.728	0.955	0.783
	GNN-CD-ft-TAGConv	0.691	0.686	0.294	0.373	0.805	0.887	0.442
	ETTIN	0.055	0.081	0.901	0.947	0.123	0.965	0.803
	ETTIN-HGA	**0.017**	**0.079**	0.894	0.944	0.009	**0.965**	0.800
MedMCQA	YVCE	0.415	0.394	0.358	0.427	0.164	0.318	0.537
	TLCC-MLP	0.306	0.330	0.479	0.383	0.437	0.282	0.521
	TLCC-CONV	0.402	0.398	0.388	0.418	0.232	0.283	0.523
	TLCC-LSTM	0.416	0.413	0.392	0.418	0.239	0.282	0.524
	P(IK)	0.564	0.519	0.275	0.431	0.001	0.292	0.536
	PHSV-half	0.244	0.264	0.486	0.374	0.454	0.265	0.511
	PHSV	**0.225**	**0.259**	0.575	0.307	0.655	0.260	0.507
	SFHS-MLP	0.531	0.499	0.307	0.431	0.060	0.283	0.516
	SFHS-Conv	0.522	0.488	0.306	0.430	0.062	0.292	0.527
	SFHS-LSTM	0.511	0.475	0.318	0.431	0.083	0.283	0.520
	GNN-SB-GAT	0.581	0.549	0.302	0.432	0.051	0.290	0.527
	GNN-SB-GCN	0.549	0.513	0.304	0.431	0.055	0.290	0.526
	GNN-SB-GraphSAGE	0.561	0.531	0.311	0.431	0.068	0.285	0.518
	CE-DT	0.553	0.545	0.319	0.421	0.098	**0.348**	0.497

(Continued on next page)

(Continued from previous page)

Dataset	**Method**	**ECE↓**	**Brier↓**	**Acc↑**	**F1↑**	**Spec↑**	**AUCPR↑**	**AUROC↑**
	CE-KNN	0.528	0.490	0.297	0.423	0.054	0.295	0.531
	CE-LogReg	0.466	0.413	0.269	0.424	0.000	0.266	0.510
	CE-RF	0.539	0.502	0.299	0.420	0.062	0.282	0.522
	CE-XGB	0.539	0.500	0.298	0.424	0.054	0.284	0.522
	LateFusion-noft-MLP	0.549	0.513	0.302	0.429	0.056	0.284	0.512
	LateFusion-noft-Conv	0.318	0.336	0.432	0.383	0.353	0.269	0.501
	LateFusion-noft-LSTM	0.495	0.471	0.311	0.429	0.072	0.279	0.510
	LateFusion-ft-MLP	0.533	0.500	0.313	0.429	0.076	0.287	0.518
	LateFusion-ft-Conv	0.296	0.311	**0.655**	0.180	**0.850**	0.269	0.488
	LateFusion-ft-LSTM	0.513	0.477	0.314	0.428	0.078	0.291	0.532
	GNN-SR-GINE	0.460	0.423	0.305	0.429	0.060	0.288	0.522
	GNN-SR-NNConv	0.309	0.304	0.370	0.417	0.200	0.276	0.511
	GNN-SR-Transformer	0.481	0.444	0.310	0.430	0.067	0.286	0.518
	GNN-CD-noft-GCN2Conv-same	0.552	0.514	0.309	0.429	0.067	0.292	0.523
	GNN-CD-noft-GCN2Conv-dual	0.440	0.407	0.320	0.423	0.097	0.286	0.511
	GNN-CD-noft-APPNP	0.544	0.503	0.309	0.429	0.067	0.291	0.515
	GNN-CD-noft-TAGConv	0.258	0.278	0.525	0.353	0.545	0.290	0.512
	GNN-CD-ft-GCN2Conv-same	0.437	0.397	0.314	0.429	0.077	0.281	0.508
	GNN-CD-ft-GCN2Conv-dual	0.512	0.472	0.308	0.429	0.067	0.285	0.512
	GNN-CD-ft-APPNP	0.446	0.405	0.317	0.426	0.087	0.286	0.517
	GNN-CD-ft-TAGConv	0.418	0.412	0.492	0.360	0.482	0.309	0.502
	ETTIN	0.522	0.489	0.312	**0.433**	0.068	0.291	0.533
	ETTIN-HGA	0.528	0.499	0.323	0.429	0.094	0.303	**0.543**
MMLU-Pro	YVCE	0.047	0.164	0.775	0.854	0.416	0.902	0.791
	TLCC-MLP	0.106	0.184	0.725	0.801	0.602	0.834	0.728
	TLCC-CONV	0.034	0.162	0.761	0.837	0.518	0.898	0.786
	TLCC-LSTM	0.062	0.170	0.753	0.840	0.360	0.883	0.763
	P(IK)	0.049	0.183	0.716	0.832	0.046	0.857	0.724
	PHSV-half	0.282	0.270	0.519	0.508	0.738	0.673	0.657
	PHSV	0.308	0.297	0.466	0.456	0.644	0.606	0.565
	SFHS-MLP	0.030	0.156	0.773	0.844	0.547	0.907	0.808
	SFHS-Conv	0.057	0.170	0.761	0.835	0.545	0.881	0.771
	SFHS-LSTM	0.096	0.183	0.758	0.829	0.601	0.871	0.766
	GNN-SB-GAT	0.066	0.167	0.769	0.846	0.475	0.886	0.779
	GNN-SB-GCN	**0.012**	0.160	0.773	0.850	0.442	0.883	0.779
	GNN-SB-GraphSAGE	0.064	0.168	0.764	0.834	0.607	0.888	0.779
	CE-DT	0.116	0.249	0.711	0.787	0.490	0.743	0.650
	CE-KNN	0.050	0.159	0.786	0.846	0.540	0.854	0.817
	CE-LogReg	0.095	0.236	0.635	0.777	0.000	0.849	0.802
	CE-RF	0.038	0.160	**0.789**	0.845	0.586	0.864	0.818
	CE-XGB	0.043	0.159	0.786	0.846	0.535	0.861	0.820
	LateFusion-noft-MLP	0.106	0.181	0.776	0.849	0.501	0.856	0.774
	LateFusion-noft-Conv	0.215	0.243	0.658	0.737	0.624	0.808	0.695
	LateFusion-noft-LSTM	0.150	0.202	0.761	0.835	0.542	0.861	0.750
	LateFusion-ft-MLP	0.119	0.182	0.775	0.845	0.554	0.871	0.789
	LateFusion-ft-Conv	0.537	0.538	0.398	0.445	0.544	0.691	0.474
	LateFusion-ft-LSTM	0.140	0.208	0.753	0.826	0.585	0.845	0.734
	GNN-SR-GINE	0.058	0.169	0.756	0.833	0.503	0.887	0.772
	GNN-SR-NNConv	0.168	0.201	0.703	0.775	0.671	0.868	0.747
	GNN-SR-Transformer	0.057	0.166	0.768	0.841	0.539	0.881	0.777
	GNN-CD-noft-GCN2Conv-same	0.091	0.183	0.767	0.836	0.601	0.887	0.786
	GNN-CD-noft-GCN2Conv-dual	0.144	0.204	0.747	0.817	0.630	0.896	0.792
	GNN-CD-noft-APPNP	0.121	0.206	0.769	0.838	0.597	0.861	0.762
	GNN-CD-noft-TAGConv	0.232	0.242	0.512	0.526	**0.842**	0.825	0.680
	GNN-CD-ft-GCN2Conv-same	0.201	0.222	0.760	0.831	0.588	0.825	0.722
	GNN-CD-ft-GCN2Conv-dual	0.104	0.186	0.776	0.845	0.565	0.873	0.783
	GNN-CD-ft-APPNP	0.200	0.230	0.740	0.813	0.606	0.837	0.732
	GNN-CD-ft-TAGConv	0.443	0.467	0.509	0.567	0.653	0.733	0.511
	ETTIN	0.039	**0.151**	0.781	**0.861**	0.366	**0.913**	**0.824**
	ETTIN-HGA	0.037	0.158	0.752	0.840	0.347	0.909	0.804

Table 38: Dataset-averaged performance metrics for `Phi-4-mini-flash-reasoning`. Each metric is the *unweighted* mean ± standard deviation of a method across all test datasets for this LLM. **Bold** entries mark the best method per metric across datasets.

Method	**ECE↓**	**Brier↓**	**Acc↑**	**F1↑**	**Spec↑**	**AUCPR↑**	**AUROC↑**
CE-DT	0.181 ±0.11	0.252 ±0.08	0.588 ±0.18	0.607 ±0.23	0.355 ±0.11	0.600 ±0.28	0.602 ±0.07
CE-KNN	0.176 ±0.10	0.239 ±0.08	0.626 ±0.15	0.591 ±0.25	0.469 ±0.13	0.586 ±0.30	0.618 ±0.09
CE-LogReg	0.224 ±0.13	0.249 ±0.07	0.513 ±0.26	0.644 ±0.23	0.002 ±0.00	0.601 ±0.30	0.630 ±0.11
CE-RF	0.151 ±0.13	0.222 ±0.08	0.613 ±0.18	0.606 ±0.25	0.388 ±0.16	0.600 ±0.30	0.634 ±0.10
CE-XGB	0.153 ±0.13	0.223 ±0.08	0.616 ±0.17	0.605 ±0.25	0.404 ±0.16	0.601 ±0.30	0.632 ±0.10
ETTIN	0.149 ±0.09	0.207 ±0.08	0.648 ±0.17	0.606 ±0.26	0.485 ±0.20	0.625 ±0.30	0.634 ±0.11
ETTIN-HGA	**0.106** ±0.09	**0.194** ±0.08	**0.685** ±0.16	0.613 ±0.28	0.469 ±0.28	0.630 ±0.31	0.641 ±0.12
GNN-CD-ft-APPNP	0.164 ±0.11	0.223 ±0.07	0.641 ±0.15	0.602 ±0.25	0.482 ±0.18	0.611 ±0.29	0.616 ±0.08
GNN-CD-ft-GCN2Conv-dual	0.389 ±0.19	0.398 ±0.20	0.581 ±0.20	0.640 ±0.23	0.204 ±0.10	**0.687** ±0.22	0.554 ±0.04
GNN-CD-ft-GCN2Conv-same	0.141 ±0.08	0.211 ±0.08	0.678 ±0.13	0.564 ±0.29	0.601 ±0.25	0.621 ±0.30	0.617 ±0.10
GNN-CD-ft-TAGConv	0.199 ±0.08	0.261 ±0.09	0.650 ±0.14	0.601 ±0.26	0.474 ±0.20	0.624 ±0.28	0.579 ±0.08
GNN-CD-noft-APPNP	0.348 ±0.17	0.351 ±0.17	0.635 ±0.17	0.624 ±0.25	0.426 ±0.15	0.635 ±0.28	0.622 ±0.08
GNN-CD-noft-GCN2Conv-dual	0.116 ±0.09	0.203 ±0.08	0.668 ±0.13	0.563 ±0.29	0.593 ±0.22	0.626 ±0.31	0.632 ±0.11
GNN-CD-noft-GCN2Conv-same	0.312 ±0.18	0.329 ±0.18	0.600 ±0.21	**0.650** ±0.23	0.255 ±0.12	0.636 ±0.28	0.626 ±0.08
GNN-CD-noft-TAGConv	0.295 ±0.18	0.330 ±0.18	0.617 ±0.19	0.623 ±0.24	0.378 ±0.16	0.650 ±0.27	0.607 ±0.08
GNN-SR-GINE	0.130 ±0.10	0.202 ±0.07	0.670 ±0.13	0.549 ±0.30	0.603 ±0.26	0.623 ±0.30	0.625 ±0.10
GNN-SR-NNConv	0.205 ±0.13	0.236 ±0.03	0.622 ±0.12	0.291 ±0.33	**0.913** ±0.16	0.626 ±0.30	0.629 ±0.09
GNN-SR-Transformer	0.148 ±0.09	0.202 ±0.06	0.668 ±0.14	0.575 ±0.28	0.575 ±0.24	0.638 ±0.30	**0.650** ±0.11
GNN-SB-GAT	0.124 ±0.10	0.204 ±0.07	0.652 ±0.13	0.572 ±0.28	0.538 ±0.20	0.622 ±0.31	0.627 ±0.11
GNN-SB-GCN	0.134 ±0.10	0.212 ±0.08	0.650 ±0.14	0.575 ±0.27	0.528 ±0.22	0.618 ±0.30	0.619 ±0.10
GNN-SB-GraphSAGE	0.131 ±0.10	0.209 ±0.08	0.646 ±0.14	0.580 ±0.27	0.520 ±0.20	0.628 ±0.31	0.636 ±0.11
LateFusion-ft-Conv	0.239 ±0.12	0.255 ±0.05	0.624 ±0.09	0.309 ±0.33	0.889 ±0.18	0.610 ±0.30	0.610 ±0.10
LateFusion-noft-Conv	0.143 ±0.06	0.202 ±0.05	0.679 ±0.11	0.529 ±0.29	0.727 ±0.19	0.630 ±0.30	0.641 ±0.11
LateFusion-ft-LSTM	0.141 ±0.07	0.213 ±0.07	0.662 ±0.13	0.572 ±0.28	0.559 ±0.21	0.620 ±0.31	0.631 ±0.10
LateFusion-noft-LSTM	0.140 ±0.08	0.213 ±0.08	0.665 ±0.14	0.601 ±0.25	0.538 ±0.22	0.630 ±0.30	0.641 ±0.10
LateFusion-ft-MLP	0.158 ±0.10	0.211 ±0.06	0.671 ±0.13	0.571 ±0.28	0.580 ±0.24	0.629 ±0.30	0.640 ±0.10
LateFusion-noft-MLP	0.136 ±0.08	0.205 ±0.06	0.676 ±0.13	0.543 ±0.30	0.646 ±0.23	0.629 ±0.30	0.634 ±0.10
PHSV	0.207 ±0.13	0.261 ±0.08	0.567 ±0.16	0.596 ±0.22	0.302 ±0.14	0.572 ±0.27	0.603 ±0.08
PHSV-half	0.199 ±0.15	0.260 ±0.09	0.570 ±0.18	0.609 ±0.23	0.280 ±0.13	0.574 ±0.28	0.600 ±0.08
P(IK)	0.133 ±0.12	0.211 ±0.09	0.627 ±0.23	0.617 ±0.30	0.246 ±0.34	0.597 ±0.31	0.592 ±0.10
SFHS-Conv	0.129 ±0.10	0.212 ±0.09	0.657 ±0.15	0.592 ±0.26	0.533 ±0.21	0.627 ±0.30	0.636 ±0.11
SFHS-LSTM	0.142 ±0.10	0.215 ±0.09	0.654 ±0.15	0.591 ±0.26	0.526 ±0.22	0.627 ±0.30	0.639 ±0.10
SFHS-MLP	0.136 ±0.10	0.211 ±0.08	0.655 ±0.13	0.574 ±0.27	0.552 ±0.21	0.626 ±0.30	0.631 ±0.10
TLCC-CONV	0.164 ±0.13	0.226 ±0.09	0.627 ±0.14	0.590 ±0.24	0.554 ±0.12	0.626 ±0.31	0.637 ±0.11
TLCC-LSTM	0.164 ±0.14	0.229 ±0.10	0.627 ±0.17	0.607 ±0.25	0.487 ±0.16	0.625 ±0.31	0.636 ±0.11
TLCC-MLP	0.161 ±0.13	0.225 ±0.09	0.636 ±0.15	0.597 ±0.25	0.523 ±0.18	0.625 ±0.31	0.635 ±0.11
YVCE	0.269 ±0.09	0.296 ±0.07	0.555 ±0.14	0.580 ±0.21	0.372 ±0.11	0.604 ±0.28	0.569 ±0.07

Table 39: Dataset-averaged performance metrics for `Qwen3-8B`. Each metric is the *unweighted* mean ± standard deviation of a method across all test datasets for this LLM. **Bold** entries mark the best method per metric across datasets.

Method	**ECE↓**	**Brier↓**	**Acc↑**	**F1↑**	**Spec↑**	**AUCPR↑**	**AUROC↑**
CE-DT	0.209 ±0.18	0.265 ±0.14	0.634 ±0.20	0.723 ±0.20	0.217 ±0.09	0.686 ±0.26	0.636 ±0.10
CE-KNN	0.204 ±0.19	0.257 ±0.14	0.610 ±0.22	0.725 ±0.20	0.073 ±0.04	0.693 ±0.26	0.661 ±0.12
CE-LogReg	0.241 ±0.21	0.278 ±0.15	0.596 ±0.23	0.723 ±0.20	0.000	0.698 ±0.26	0.684 ±0.12
CE-RF	0.201 ±0.20	0.253 ±0.14	0.600 ±0.23	0.724 ±0.20	0.019 ±0.01	0.699 ±0.27	**0.687** ±0.13
CE-XGB	0.204 ±0.19	0.259 ±0.14	0.616 ±0.21	0.722 ±0.20	0.116 ±0.05	0.683 ±0.28	0.663 ±0.13
ETTIN	0.164 ±0.20	0.224 ±0.16	0.679 ±0.22	0.752 ±0.21	0.191 ±0.13	0.722 ±0.27	0.649 ±0.08
ETTIN-HGA	0.164 ±0.16	0.214 ±0.13	0.670 ±0.24	**0.757** ±0.22	0.085 ±0.07	0.729 ±0.27	0.653 ±0.09
GNN-CD-ft-APPNP	0.385 ±0.11	0.384 ±0.11	0.606 ±0.11	0.552 ±0.25	0.678 ±0.21	0.730 ±0.25	0.591 ±0.08
GNN-CD-ft-GCN2Conv-dual	0.188 ±0.11	0.230 ±0.12	0.667 ±0.18	0.675 ±0.25	0.463 ±0.22	0.699 ±0.29	0.630 ±0.11
GNN-CD-ft-GCN2Conv-same	0.186 ±0.12	0.229 ±0.12	0.655 ±0.17	0.565 ±0.34	0.645 ±0.28	0.702 ±0.29	0.645 ±0.11
GNN-CD-ft-TAGConv	0.265 ±0.13	0.337 ±0.16	0.661 ±0.16	0.582 ±0.33	0.618 ±0.28	0.741 ±0.25	0.604 ±0.09
GNN-CD-noft-APPNP	0.342 ±0.18	0.342 ±0.18	0.655 ±0.18	0.640 ±0.27	0.514 ±0.25	**0.759** ±0.22	0.590 ±0.07
GNN-CD-noft-GCN2Conv-dual	0.165 ±0.16	0.228 ±0.13	0.658 ±0.19	0.682 ±0.25	0.369 ±0.23	0.704 ±0.29	0.632 ±0.09
GNN-CD-noft-GCN2Conv-same	0.369 ±0.15	0.364 ±0.15	0.595 ±0.16	0.502 ±0.31	0.713 ±0.23	0.726 ±0.26	0.621 ±0.09
GNN-CD-noft-TAGConv	0.252 ±0.14	0.435 ±0.05	0.545 ±0.05	0.586 ±0.15	0.522 ±0.08	0.711 ±0.24	0.543 ±0.03
GNN-SR-GINE	0.278 ±0.16	0.267 ±0.07	0.507 ±0.20	0.238 ±0.32	**0.887** ±0.24	0.703 ±0.29	0.629 ±0.11
GNN-SR-NNConv	0.258 ±0.11	0.241 ±0.03	0.671 ±0.19	0.642 ±0.28	0.519 ±0.27	0.709 ±0.29	0.647 ±0.11
GNN-SR-Transformer	0.152 ±0.12	**0.205** ±0.11	0.683 ±0.19	0.714 ±0.23	0.365 ±0.20	0.706 ±0.29	0.649 ±0.09
GNN-SB-GAT	0.155 ±0.15	0.219 ±0.13	0.680 ±0.20	0.718 ±0.23	0.309 ±0.18	0.695 ±0.29	0.628 ±0.10
GNN-SB-GCN	**0.130** ±0.17	0.211 ±0.14	0.681 ±0.22	0.699 ±0.26	0.324 ±0.29	0.704 ±0.29	0.643 ±0.10
GNN-SB-GraphSAGE	0.150 ±0.18	0.215 ±0.15	0.685 ±0.22	0.724 ±0.23	0.309 ±0.26	0.722 ±0.28	0.668 ±0.10
LateFusion-ft-Conv	0.161 ±0.12	0.213 ±0.11	0.676 ±0.19	0.728 ±0.22	0.283 ±0.13	0.703 ±0.28	0.625 ±0.08
LateFusion-noft-Conv	0.227 ±0.13	0.261 ±0.13	0.660 ±0.16	0.693 ±0.23	0.404 ±0.13	0.693 ±0.28	0.616 ±0.08
LateFusion-ft-LSTM	0.237 ±0.16	0.270 ±0.16	0.650 ±0.21	0.706 ±0.22	0.309 ±0.16	0.693 ±0.29	0.609 ±0.08
LateFusion-noft-LSTM	0.278 ±0.18	0.300 ±0.18	0.631 ±0.21	0.711 ±0.21	0.254 ±0.08	0.683 ±0.29	0.584 ±0.08
LateFusion-ft-MLP	0.194 ±0.10	0.222 ±0.11	0.660 ±0.18	0.646 ±0.26	0.548 ±0.25	0.703 ±0.30	0.649 ±0.12
LateFusion-noft-MLP	0.174 ±0.15	0.222 ±0.13	0.670 ±0.21	0.637 ±0.31	0.488 ±0.30	0.702 ±0.30	0.649 ±0.12
PHSV	0.184 ±0.15	0.261 ±0.10	0.596 ±0.18	0.682 ±0.19	0.243 ±0.11	0.634 ±0.26	0.603 ±0.09
PHSV-half	0.225 ±0.16	0.287 ±0.12	0.582 ±0.17	0.669 ±0.19	0.224 ±0.13	0.616 ±0.24	0.591 ±0.08
P(IK)	0.199 ±0.21	0.243 ±0.17	0.625 ±0.27	0.727 ±0.22	0.076 ±0.17	0.695 ±0.28	0.583 ±0.08
SFHS-Conv	0.157 ±0.14	0.219 ±0.13	**0.685** ±0.18	0.711 ±0.22	0.410 ±0.21	0.709 ±0.29	0.651 ±0.10
SFHS-LSTM	0.165 ±0.14	0.222 ±0.13	0.677 ±0.19	0.698 ±0.24	0.379 ±0.22	0.703 ±0.29	0.629 ±0.10
SFHS-MLP	0.177 ±0.13	0.226 ±0.13	0.679 ±0.19	0.688 ±0.25	0.456 ±0.20	0.706 ±0.29	0.652 ±0.11
TLCC-CONV	0.167 ±0.09	0.216 ±0.11	0.668 ±0.17	0.713 ±0.22	0.365 ±0.18	0.696 ±0.29	0.606 ±0.10
TLCC-LSTM	0.174 ±0.17	0.225 ±0.14	0.662 ±0.22	0.744 ±0.21	0.188 ±0.10	0.696 ±0.29	0.619 ±0.10
TLCC-MLP	0.171 ±0.08	0.214 ±0.09	0.654 ±0.17	0.702 ±0.21	0.390 ±0.19	0.690 ±0.29	0.601 ±0.10
YVCE	0.213 ±0.21	0.253 ±0.16	0.645 ±0.26	0.754 ±0.21	0.015 ±0.01	0.694 ±0.28	0.561 ±0.06

Table 40: Dataset-averaged performance metrics for `Qwen3-14B`. Each metric is the *unweighted* mean ± standard deviation of a method across all test datasets for this LLM. **Bold** entries mark the best method per metric across datasets.

Method	**ECE↓**	**Brier↓**	**Acc↑**	**F1↑**	**Spec↑**	**AUCPR↑**	**AUROC↑**
CE-DT	0.198 ±0.20	0.257 ±0.15	0.633 ±0.22	0.726 ±0.21	0.191 ±0.10	0.675 ±0.27	0.616 ±0.08
CE-KNN	0.180 ±0.21	0.243 ±0.15	0.621 ±0.24	0.734 ±0.21	0.047 ±0.02	0.699 ±0.28	0.665 ±0.12
CE-LogReg	0.181 ±0.22	0.244 ±0.16	0.636 ±0.23	0.737 ±0.21	0.116 ±0.08	0.700 ±0.29	**0.673** ±0.13
CE-RF	0.178 ±0.21	0.243 ±0.15	0.626 ±0.24	0.737 ±0.21	0.055 ±0.03	0.694 ±0.28	0.664 ±0.11
CE-XGB	0.181 ±0.21	0.247 ±0.15	0.630 ±0.23	0.735 ±0.21	0.086 ±0.04	0.690 ±0.28	0.654 ±0.11
ETTIN	0.178 ±0.22	0.229 ±0.18	0.660 ±0.26	0.762 ±0.22	0.063 ±0.07	0.721 ±0.29	0.649 ±0.11
ETTIN-HGA	0.157 ±0.20	0.216 ±0.16	0.685 ±0.23	0.762 ±0.21	0.150 ±0.15	0.742 ±0.27	0.658 ±0.09
GNN-CD-ft-APPNP	0.186 ±0.09	0.214 ±0.10	0.669 ±0.18	0.632 ±0.28	0.572 ±0.26	0.711 ±0.28	0.632 ±0.09
GNN-CD-ft-GCN2Conv-dual	0.178 ±0.16	0.214 ±0.14	0.677 ±0.22	0.713 ±0.24	0.316 ±0.28	0.726 ±0.28	0.644 ±0.10
GNN-CD-ft-GCN2Conv-same	0.252 ±0.11	0.254 ±0.12	0.643 ±0.17	0.560 ±0.33	0.679 ±0.24	0.717 ±0.29	0.642 ±0.11
GNN-CD-ft-TAGConv	0.266 ±0.13	0.283 ±0.12	0.641 ±0.18	0.584 ±0.30	0.622 ±0.25	0.725 ±0.27	0.623 ±0.08
GNN-CD-noft-APPNP	0.174 ±0.12	0.204 ±0.12	0.647 ±0.24	0.647 ±0.29	0.407 ±0.33	0.716 ±0.28	0.629 ±0.11
GNN-CD-noft-GCN2Conv-dual	**0.156** ±0.16	0.213 ±0.14	**0.691** ±0.21	0.711 ±0.24	0.371 ±0.29	0.723 ±0.28	0.651 ±0.11
GNN-CD-noft-GCN2Conv-same	0.203 ±0.13	0.226 ±0.12	0.643 ±0.21	0.580 ±0.33	0.632 ±0.26	0.722 ±0.28	0.641 ±0.10
GNN-CD-noft-TAGConv	0.242 ±0.11	0.262 ±0.10	0.618 ±0.18	0.569 ±0.31	0.611 ±0.25	0.718 ±0.27	0.616 ±0.07
GNN-SR-GINE	0.165 ±0.15	0.212 ±0.13	0.671 ±0.22	0.711 ±0.23	0.331 ±0.26	0.723 ±0.29	0.645 ±0.09
GNN-SR-NNConv	0.257 ±0.11	0.237 ±0.03	0.595 ±0.18	0.453 ±0.39	**0.774** ±0.23	0.729 ±0.29	0.650 ±0.10
GNN-SR-Transformer	0.182 ±0.14	0.210 ±0.12	0.660 ±0.24	0.689 ±0.25	0.327 ±0.32	0.726 ±0.28	0.652 ±0.10
GNN-SB-GAT	0.163 ±0.17	0.214 ±0.14	0.658 ±0.24	0.640 ±0.30	0.414 ±0.39	0.725 ±0.29	0.648 ±0.11
GNN-SB-GCN	0.163 ±0.18	0.220 ±0.15	0.680 ±0.21	0.679 ±0.27	0.420 ±0.32	0.723 ±0.29	0.650 ±0.11
GNN-SB-GraphSAGE	0.168 ±0.19	0.222 ±0.15	0.669 ±0.22	0.671 ±0.27	0.417 ±0.33	0.724 ±0.29	0.652 ±0.11
LateFusion-ft-Conv	0.185 ±0.09	**0.203** ±0.09	0.671 ±0.20	0.704 ±0.24	0.379 ±0.24	0.723 ±0.29	0.650 ±0.10
LateFusion-noft-Conv	0.199 ±0.09	0.214 ±0.07	0.654 ±0.15	0.609 ±0.27	0.633 ±0.20	0.714 ±0.29	0.640 ±0.09
LateFusion-ft-LSTM	0.210 ±0.13	0.237 ±0.13	0.652 ±0.20	0.602 ±0.31	0.581 ±0.28	0.718 ±0.29	0.652 ±0.11
LateFusion-noft-LSTM	0.202 ±0.15	0.241 ±0.14	0.648 ±0.21	0.662 ±0.26	0.429 ±0.22	0.705 ±0.29	0.618 ±0.10
LateFusion-ft-MLP	0.189 ±0.11	0.215 ±0.11	0.682 ±0.18	0.667 ±0.27	0.523 ±0.26	0.718 ±0.28	0.643 ±0.10
LateFusion-noft-MLP	0.176 ±0.14	0.217 ±0.13	0.675 ±0.20	0.653 ±0.28	0.509 ±0.30	0.725 ±0.28	0.655 ±0.10
PHSV	0.175 ±0.15	0.245 ±0.10	0.614 ±0.18	0.674 ±0.21	0.308 ±0.21	0.660 ±0.26	0.614 ±0.08
PHSV-half	0.167 ±0.16	0.243 ±0.11	0.617 ±0.18	0.679 ±0.21	0.315 ±0.20	0.667 ±0.27	0.621 ±0.08
P(IK)	0.195 ±0.19	0.232 ±0.16	0.658 ±0.26	0.757 ±0.22	0.030 ±0.05	0.683 ±0.29	0.565 ±0.12
SFHS-Conv	0.197 ±0.18	0.237 ±0.16	0.664 ±0.23	0.664 ±0.28	0.420 ±0.31	0.716 ±0.30	0.649 ±0.11
SFHS-LSTM	0.184 ±0.14	0.230 ±0.14	0.670 ±0.20	0.694 ±0.24	0.385 ±0.23	0.707 ±0.29	0.622 ±0.09
SFHS-MLP	0.203 ±0.19	0.243 ±0.16	0.657 ±0.23	0.643 ±0.29	0.433 ±0.34	0.719 ±0.29	0.644 ±0.09
TLCC-CONV	0.178 ±0.11	0.220 ±0.11	0.661 ±0.19	0.717 ±0.21	0.353 ±0.20	0.707 ±0.30	0.616 ±0.09
TLCC-LSTM	0.182 ±0.12	0.223 ±0.12	0.665 ±0.20	0.711 ±0.22	0.381 ±0.23	0.713 ±0.30	0.629 ±0.11
TLCC-MLP	0.177 ±0.09	0.209 ±0.11	0.655 ±0.21	0.712 ±0.22	0.353 ±0.24	0.704 ±0.31	0.614 ±0.11
YVCE	0.238 ±0.24	0.262 ±0.21	0.662 ±0.26	**0.766** ±0.21	0.017 ±0.01	**0.802** ±0.21	0.645 ±0.09

Table 41: Dataset-averaged performance metrics for `Magistral-Small-2506`. Each metric is the *unweighted* mean ± standard deviation of a method across all test datasets for this LLM. **Bold** entries mark the best method per metric across datasets.

Method	**ECE↓**	**Brier↓**	**Acc↑**	**F1↑**	**Spec↑**	**AUCPR↑**	**AUROC↑**
CE-DT	0.237 ±0.15	0.245 ±0.14	0.642 ±0.21	0.392 ±0.24	0.609 ±0.29	0.353 ±0.23	0.604 ±0.07
CE-KNN	0.226 ±0.16	0.236 ±0.14	0.666 ±0.22	0.394 ±0.22	0.660 ±0.30	0.364 ±0.24	0.621 ±0.09
CE-LogReg	0.199 ±0.15	0.219 ±0.10	0.693 ±0.22	0.362 ±0.21	0.745 ±0.28	0.356 ±0.24	0.599 ±0.08
CE-RF	0.185 ±0.15	0.209 ±0.11	0.695 ±0.24	0.399 ±0.19	0.724 ±0.31	0.367 ±0.24	0.630 ±0.08
CE-XGB	0.228 ±0.16	0.234 ±0.14	0.664 ±0.22	0.399 ±0.22	0.653 ±0.30	0.369 ±0.24	0.628 ±0.10
ETTIN	0.152 ±0.23	0.204 ±0.20	**0.716** ±0.26	**0.540** ±0.20	0.647 ±0.38	0.526 ±0.25	**0.737** ±0.12
ETTIN-HGA	0.160 ±0.20	0.201 ±0.17	0.710 ±0.25	0.493 ±0.21	0.669 ±0.36	0.496 ±0.24	0.724 ±0.12
GNN-CD-ft-APPNP	0.200 ±0.13	0.218 ±0.11	0.665 ±0.20	0.394 ±0.21	0.696 ±0.25	0.396 ±0.26	0.652 ±0.11
GNN-CD-ft-GCN2Conv-dual	0.167 ±0.10	**0.185** ±0.08	0.706 ±0.24	0.422 ±0.19	0.758 ±0.31	0.467 ±0.24	0.700 ±0.12
GNN-CD-ft-GCN2Conv-same	0.203 ±0.17	0.225 ±0.15	0.655 ±0.22	0.457 ±0.24	0.578 ±0.31	0.440 ±0.25	0.691 ±0.11
GNN-CD-ft-TAGConv	0.327 ±0.22	0.335 ±0.23	0.662 ±0.23	0.450 ±0.23	0.572 ±0.34	0.431 ±0.24	0.614 ±0.08
GNN-CD-noft-APPNP	0.174 ±0.15	0.203 ±0.11	0.679 ±0.24	0.411 ±0.22	0.695 ±0.32	0.405 ±0.27	0.664 ±0.11
GNN-CD-noft-GCN2Conv-dual	0.170 ±0.16	0.203 ±0.13	0.697 ±0.24	0.445 ±0.21	0.712 ±0.32	0.451 ±0.25	0.684 ±0.12
GNN-CD-noft-GCN2Conv-same	0.163 ±0.16	0.202 ±0.12	0.682 ±0.23	0.418 ±0.25	0.678 ±0.32	0.410 ±0.28	0.664 ±0.12
GNN-CD-noft-TAGConv	0.250 ±0.09	0.282 ±0.12	0.664 ±0.17	0.398 ±0.18	0.691 ±0.20	0.381 ±0.23	0.556 ±0.04
GNN-SR-GINE	0.242 ±0.16	0.231 ±0.08	0.699 ±0.24	0.492 ±0.20	0.644 ±0.35	0.467 ±0.23	0.692 ±0.11
GNN-SR-NNConv	0.221 ±0.16	0.217 ±0.09	0.686 ±0.24	0.480 ±0.23	0.624 ±0.33	0.473 ±0.24	0.706 ±0.11
GNN-SR-Transformer	0.201 ±0.12	0.195 ±0.08	0.712 ±0.24	0.440 ±0.19	**0.759** ±0.31	0.492 ±0.23	0.718 ±0.12
GNN-SB-GAT	0.157 ±0.17	0.199 ±0.14	0.700 ±0.24	0.463 ±0.23	0.674 ±0.33	0.453 ±0.26	0.689 ±0.11
GNN-SB-GCN	0.164 ±0.19	0.206 ±0.16	0.701 ±0.25	0.477 ±0.22	0.670 ±0.34	0.463 ±0.26	0.699 ±0.12
GNN-SB-GraphSAGE	0.151 ±0.19	0.199 ±0.16	0.709 ±0.25	0.486 ±0.20	0.697 ±0.34	0.488 ±0.23	0.717 ±0.11
LateFusion-ft-Conv	0.209 ±0.13	0.228 ±0.11	0.675 ±0.20	0.397 ±0.22	0.706 ±0.26	0.396 ±0.27	0.597 ±0.09
LateFusion-noft-Conv	0.202 ±0.12	0.228 ±0.08	0.626 ±0.15	0.337 ±0.22	0.686 ±0.21	0.375 ±0.27	0.565 ±0.10
LateFusion-ft-LSTM	0.191 ±0.17	0.215 ±0.15	0.696 ±0.24	0.441 ±0.21	0.704 ±0.32	0.429 ±0.27	0.681 ±0.09
LateFusion-noft-LSTM	0.180 ±0.17	0.218 ±0.16	0.696 ±0.23	0.444 ±0.20	0.703 ±0.30	0.438 ±0.25	0.672 ±0.09
LateFusion-ft-MLP	0.161 ±0.18	0.203 ±0.16	0.702 ±0.25	0.426 ±0.20	0.748 ±0.33	0.462 ±0.24	0.683 ±0.13
LateFusion-noft-MLP	0.158 ±0.18	0.201 ±0.15	0.703 ±0.25	0.442 ±0.20	0.729 ±0.34	0.464 ±0.25	0.700 ±0.12
PHSV	0.168 ±0.14	0.204 ±0.10	0.690 ±0.22	0.393 ±0.19	0.724 ±0.28	0.377 ±0.24	0.628 ±0.07
PHSV-half	0.216 ±0.14	0.227 ±0.11	0.691 ±0.22	0.389 ±0.20	0.721 ±0.28	0.371 ±0.24	0.604 ±0.09
P(IK)	**0.135** ±0.21	0.206 ±0.17	0.683 ±0.26	0.409 ±0.32	0.571 ±0.41	0.410 ±0.22	0.681 ±0.13
SFHS-Conv	0.154 ±0.17	0.199 ±0.15	0.711 ±0.23	0.498 ±0.19	0.689 ±0.32	0.473 ±0.25	0.697 ±0.10
SFHS-LSTM	0.164 ±0.17	0.207 ±0.15	0.696 ±0.24	0.454 ±0.21	0.681 ±0.33	0.435 ±0.25	0.672 ±0.11
SFHS-MLP	0.152 ±0.17	0.198 ±0.14	0.713 ±0.24	0.493 ±0.18	0.705 ±0.32	0.484 ±0.23	0.680 ±0.12
TLCC-CONV	0.178 ±0.17	0.204 ±0.15	0.714 ±0.24	0.525 ±0.21	0.631 ±0.39	0.467 ±0.25	0.698 ±0.13
TLCC-LSTM	0.160 ±0.17	0.196 ±0.14	0.712 ±0.24	0.519 ±0.21	0.623 ±0.39	0.466 ±0.25	0.698 ±0.12
TLCC-MLP	0.162 ±0.16	0.195 ±0.14	0.716 ±0.24	**0.532** ±0.19	0.632 ±0.40	0.466 ±0.24	0.694 ±0.12
YVCE	0.557 ±0.27	0.550 ±0.23	0.320 ±0.21	0.406 ±0.27	0.096 ±0.10	**0.543** ±0.13	0.521 ±0.08

Table 42: Dataset-averaged performance metrics for `QwQ-32B`. Each metric is the *unweighted* mean ± standard deviation of a method across all test datasets for this LLM. **Bold** entries mark the best method per metric across datasets.

Method	**ECE↓**	**Brier↓**	**Acc↑**	**F1↑**	**Spec↑**	**AUCPR↑**	**AUROC↑**
CE-DT	0.213 ±0.20	0.272 ±0.16	0.627 ±0.22	0.721 ±0.20	0.200 ±0.09	0.655 ±0.26	0.612 ±0.10
CE-KNN	0.198 ±0.20	0.254 ±0.16	0.625 ±0.23	0.728 ±0.20	0.145 ±0.09	0.695 ±0.26	0.651 ±0.12
CE-LogReg	0.242 ±0.18	0.268 ±0.14	0.615 ±0.24	0.737 ±0.20	0.000	0.685 ±0.27	0.649 ±0.12
CE-RF	0.187 ±0.20	0.249 ±0.15	0.626 ±0.23	0.729 ±0.20	0.135 ±0.09	0.685 ±0.27	0.652 ±0.13
CE-XGB	0.211 ±0.19	0.262 ±0.16	0.631 ±0.22	0.724 ±0.20	0.188 ±0.09	0.679 ±0.28	0.640 ±0.12
ETTIN	0.171 ±0.19	0.224 ±0.16	0.682 ±0.22	0.760 ±0.20	0.195 ±0.12	0.734 ±0.27	0.659 ±0.09
ETTIN-HGA	0.155 ±0.20	0.220 ±0.17	**0.698** ±0.23	**0.769** ±0.21	0.148 ±0.19	0.726 ±0.26	0.650 ±0.08
GNN-CD-ft-APPNP	0.332 ±0.11	0.331 ±0.13	0.609 ±0.12	0.645 ±0.19	0.549 ±0.11	0.706 ±0.27	0.618 ±0.09
GNN-CD-ft-GCN2Conv-dual	0.180 ±0.14	0.220 ±0.14	0.692 ±0.20	0.672 ±0.30	0.450 ±0.32	0.723 ±0.28	**0.663** ±0.11
GNN-CD-ft-GCN2Conv-same	0.292 ±0.18	0.303 ±0.18	0.635 ±0.20	0.716 ±0.20	0.315 ±0.10	0.699 ±0.27	0.596 ±0.07
GNN-CD-ft-TAGConv	0.296 ±0.14	0.320 ±0.13	0.628 ±0.15	0.689 ±0.19	0.419 ±0.10	0.722 ±0.25	0.587 ±0.07
GNN-CD-noft-APPNP	0.199 ±0.12	0.224 ±0.11	0.666 ±0.18	0.681 ±0.25	0.466 ±0.22	0.714 ±0.26	0.630 ±0.08
GNN-CD-noft-GCN2Conv-dual	0.222 ±0.09	0.237 ±0.11	0.645 ±0.16	0.592 ±0.33	**0.615** ±0.24	0.719 ±0.27	0.647 ±0.10
GNN-CD-noft-GCN2Conv-same	0.171 ±0.18	0.225 ±0.16	0.675 ±0.22	0.706 ±0.25	0.363 ±0.23	0.719 ±0.27	0.642 ±0.10
GNN-CD-noft-TAGConv	0.201 ±0.14	0.246 ±0.15	0.670 ±0.20	0.706 ±0.23	0.375 ±0.16	0.714 ±0.27	0.614 ±0.09
GNN-SR-GINE	0.182 ±0.15	0.217 ±0.14	0.667 ±0.24	0.753 ±0.21	0.171 ±0.11	0.713 ±0.27	0.625 ±0.08
GNN-SR-NNConv	0.164 ±0.21	0.230 ±0.18	0.676 ±0.23	0.705 ±0.26	0.318 ±0.29	0.716 ±0.29	0.645 ±0.12
GNN-SR-Transformer	0.184 ±0.12	**0.209** ±0.12	0.679 ±0.24	0.726 ±0.24	0.253 ±0.26	0.714 ±0.28	0.638 ±0.11
GNN-SB-GAT	0.169 ±0.19	0.230 ±0.17	0.674 ±0.23	0.707 ±0.24	0.350 ±0.24	0.714 ±0.29	0.650 ±0.12
GNN-SB-GCN	0.145 ±0.17	0.212 ±0.15	0.687 ±0.22	0.701 ±0.26	0.373 ±0.28	0.726 ±0.28	0.662 ±0.12
GNN-SB-GraphSAGE	**0.139** ±0.19	0.214 ±0.16	0.686 ±0.23	0.710 ±0.26	0.311 ±0.29	0.719 ±0.28	0.655 ±0.12
LateFusion-ft-Conv	0.200 ±0.20	0.245 ±0.17	0.672 ±0.23	0.752 ±0.21	0.221 ±0.13	0.721 ±0.28	0.642 ±0.09
LateFusion-noft-Conv	0.210 ±0.09	0.233 ±0.10	0.645 ±0.15	0.700 ±0.19	0.460 ±0.12	0.716 ±0.28	0.638 ±0.09
LateFusion-ft-LSTM	0.210 ±0.19	0.250 ±0.18	0.661 ±0.22	0.720 ±0.23	0.292 ±0.17	0.708 ±0.28	0.625 ±0.10
LateFusion-noft-LSTM	0.241 ±0.16	0.269 ±0.16	0.635 ±0.20	0.700 ±0.21	0.345 ±0.15	0.699 ±0.28	0.603 ±0.08
LateFusion-ft-MLP	0.178 ±0.16	0.223 ±0.15	0.678 ±0.21	0.646 ±0.32	0.442 ±0.35	0.719 ±0.28	0.652 ±0.11
LateFusion-noft-MLP	0.205 ±0.11	0.223 ±0.12	0.678 ±0.20	0.635 ±0.34	0.512 ±0.29	0.720 ±0.27	0.651 ±0.10
PHSV	0.156 ±0.17	0.241 ±0.12	0.625 ±0.20	0.708 ±0.20	0.234 ±0.13	0.666 ±0.26	0.616 ±0.09
PHSV-half	0.184 ±0.17	0.260 ±0.12	0.601 ±0.19	0.699 ±0.19	0.167 ±0.11	0.650 ±0.24	0.593 ±0.07
P(IK)	0.178 ±0.21	0.237 ±0.17	0.660 ±0.25	0.669 ±0.28	0.325 ±0.39	0.718 ±0.27	0.619 ±0.08
SFHS-Conv	0.173 ±0.17	0.226 ±0.15	0.680 ±0.21	0.734 ±0.21	0.344 ±0.16	0.721 ±0.28	0.651 ±0.11
SFHS-LSTM	0.189 ±0.16	0.238 ±0.15	0.668 ±0.21	0.702 ±0.24	0.378 ±0.21	0.709 ±0.29	0.631 ±0.11
SFHS-MLP	0.196 ±0.17	0.239 ±0.16	0.670 ±0.22	0.701 ±0.24	0.377 ±0.22	0.706 ±0.29	0.634 ±0.11
TLCC-CONV	0.234 ±0.11	0.254 ±0.11	0.629 ±0.16	0.711 ±0.18	0.355 ±0.15	0.707 ±0.27	0.610 ±0.08
TLCC-LSTM	0.200 ±0.09	0.228 ±0.09	0.641 ±0.15	0.702 ±0.19	0.410 ±0.15	0.708 ±0.27	0.619 ±0.08
TLCC-MLP	0.201 ±0.10	0.225 ±0.10	0.638 ±0.17	0.711 ±0.19	0.370 ±0.17	0.708 ±0.28	0.619 ±0.08
YVCE	0.216 ±0.23	0.250 ±0.19	0.664 ±0.26	0.768 ±0.21	0.074 ±0.07	**0.780** ±0.23	0.643 ±0.09

Table 43: Dataset-averaged performance metrics for `EXAONE-Deep-32B`. Each metric is the *unweighted* mean ± standard deviation of a method across all test datasets for this LLM. **Bold** entries mark the best method per metric across datasets.

Method	**ECE↓**	**Brier↓**	**Acc↑**	**F1↑**	**Spec↑**	**AUCPR↑**	**AUROC↑**
CE-DT	0.217 ±0.17	0.304 ±0.13	0.632 ±0.18	0.713 ±0.18	0.323 ±0.19	0.679 ±0.20	0.595 ±0.09
CE-KNN	0.176 ±0.19	0.245 ±0.14	0.651 ±0.22	0.733 ±0.20	0.306 ±0.26	0.686 ±0.25	0.671 ±0.15
CE-LogReg	0.165 ±0.19	0.261 ±0.10	0.588 ±0.21	0.720 ±0.18	0.000	0.676 ±0.26	0.669 ±0.14
CE-RF	0.167 ±0.19	0.239 ±0.14	0.655 ±0.21	0.730 ±0.20	0.340 ±0.26	0.680 ±0.26	0.670 ±0.15
CE-XGB	0.165 ±0.19	0.240 ±0.14	0.653 ±0.21	0.730 ±0.20	0.312 ±0.25	0.681 ±0.26	0.669 ±0.15
ETTIN	0.146 ±0.20	0.215 ±0.15	0.680 ±0.22	**0.770** ±0.20	0.159 ±0.11	0.752 ±0.26	**0.706** ±0.11
ETTIN-HGA	0.143 ±0.19	0.218 ±0.15	0.684 ±0.21	0.765 ±0.20	0.171 ±0.14	0.745 ±0.26	0.697 ±0.10
GNN-CD-ft-APPNP	0.256 ±0.12	0.273 ±0.08	0.647 ±0.18	0.703 ±0.22	0.406 ±0.31	0.697 ±0.25	0.615 ±0.11
GNN-CD-ft-GCN2Conv-dual	0.214 ±0.19	0.264 ±0.14	0.658 ±0.20	0.747 ±0.19	0.296 ±0.23	0.715 ±0.26	0.654 ±0.11
GNN-CD-ft-GCN2Conv-same	0.263 ±0.13	0.271 ±0.09	0.649 ±0.20	0.739 ±0.19	0.321 ±0.27	0.695 ±0.26	0.619 ±0.11
GNN-CD-ft-TAGConv	0.447 ±0.16	0.477 ±0.13	0.486 ±0.12	0.515 ±0.15	0.567 ±0.16	0.669 ±0.22	0.499 ±0.04
GNN-CD-noft-APPNP	0.247 ±0.17	0.280 ±0.13	0.646 ±0.20	0.732 ±0.19	0.337 ±0.28	0.709 ±0.26	0.635 ±0.11
GNN-CD-noft-GCN2Conv-dual	0.215 ±0.12	0.252 ±0.09	0.620 ±0.19	0.675 ±0.21	0.476 ±0.23	0.721 ±0.28	0.660 ±0.13
GNN-CD-noft-GCN2Conv-same	0.211 ±0.18	0.256 ±0.14	0.656 ±0.19	0.712 ±0.22	0.392 ±0.31	0.716 ±0.26	0.647 ±0.12
GNN-CD-noft-TAGConv	0.260 ±0.11	0.261 ±0.02	0.493 ±0.12	0.434 ±0.21	**0.720** ±0.23	0.679 ±0.26	0.568 ±0.10
GNN-SR-GINE	0.168 ±0.15	0.219 ±0.11	0.666 ±0.19	0.748 ±0.19	0.315 ±0.25	0.705 ±0.26	0.630 ±0.13
GNN-SR-NNConv	0.210 ±0.09	0.230 ±0.06	0.601 ±0.19	0.584 ±0.31	0.537 ±0.27	0.691 ±0.27	0.608 ±0.15
GNN-SR-Transformer	0.181 ±0.15	0.227 ±0.12	0.659 ±0.20	0.743 ±0.19	0.316 ±0.24	0.708 ±0.26	0.631 ±0.13
GNN-SB-GAT	0.183 ±0.20	0.242 ±0.16	0.668 ±0.21	0.744 ±0.20	0.295 ±0.23	0.708 ±0.26	0.643 ±0.12
GNN-SB-GCN	0.161 ±0.20	0.232 ±0.15	0.671 ±0.22	0.740 ±0.22	0.279 ±0.22	0.712 ±0.26	0.645 ±0.12
GNN-SB-GraphSAGE	0.185 ±0.19	0.243 ±0.15	0.656 ±0.19	0.684 ±0.27	0.425 ±0.35	0.700 ±0.27	0.625 ±0.13
LateFusion-ft-Conv	0.500 ±0.13	0.497 ±0.12	0.450 ±0.13	0.395 ±0.18	0.635 ±0.25	0.654 ±0.24	0.514 ±0.05
LateFusion-noft-Conv	0.253 ±0.07	0.266 ±0.08	0.626 ±0.12	0.685 ±0.18	0.487 ±0.11	0.705 ±0.27	0.641 ±0.09
LateFusion-ft-LSTM	0.234 ±0.15	0.273 ±0.12	0.645 ±0.19	0.721 ±0.20	0.355 ±0.26	0.683 ±0.26	0.599 ±0.14
LateFusion-noft-LSTM	0.234 ±0.16	0.269 ±0.13	0.651 ±0.20	0.726 ±0.21	0.345 ±0.22	0.699 ±0.27	0.631 ±0.11
LateFusion-ft-MLP	0.234 ±0.17	0.265 ±0.15	0.655 ±0.20	0.718 ±0.20	0.413 ±0.21	0.716 ±0.26	0.657 ±0.12
LateFusion-noft-MLP	0.217 ±0.19	0.263 ±0.15	0.664 ±0.21	0.747 ±0.20	0.293 ±0.21	0.713 ±0.26	0.652 ±0.10
PHSV	0.293 ±0.07	0.294 ±0.09	0.564 ±0.15	0.548 ±0.23	0.526 ±0.14	0.551 ±0.23	0.522 ±0.08
PHSV-half	0.287 ±0.07	0.285 ±0.10	0.568 ±0.15	0.576 ±0.21	0.523 ±0.13	0.574 ±0.25	0.557 ±0.09
P(IK)	0.160 ±0.20	0.232 ±0.15	0.651 ±0.23	0.728 ±0.22	0.172 ±0.24	0.712 ±0.26	0.633 ±0.09
SFHS-Conv	0.182 ±0.17	0.237 ±0.13	0.657 ±0.19	0.740 ±0.19	0.323 ±0.25	0.706 ±0.26	0.633 ±0.12
SFHS-LSTM	0.227 ±0.15	0.259 ±0.12	0.629 ±0.20	0.713 ±0.19	0.365 ±0.24	0.690 ±0.27	0.610 ±0.13
SFHS-MLP	0.156 ±0.19	0.229 ±0.15	0.668 ±0.20	0.720 ±0.23	0.371 ±0.30	0.715 ±0.27	0.650 ±0.14
TLCC-CONV	0.145 ±0.15	0.212 ±0.11	0.693 ±0.19	0.759 ±0.20	0.284 ±0.14	0.727 ±0.26	0.667 ±0.10
TLCC-LSTM	0.165 ±0.15	0.219 ±0.12	0.685 ±0.19	0.752 ±0.20	0.253 ±0.11	0.718 ±0.27	0.663 ±0.10
TLCC-MLP	**0.141** ±0.08	**0.200** ±0.08	**0.700** ±0.15	0.730 ±0.21	0.419 ±0.18	0.715 ±0.25	0.640 ±0.07
YVCE	0.182 ±0.18	0.234 ±0.14	0.671 ±0.21	0.754 ±0.20	0.281 ±0.12	**0.752** ±0.26	0.681 ±0.10

Table 44: Performance for each dataset–method pair averaged across all LLMs. Metrics are *unweighted* means ± standard deviations across models, showing which methods generalize best on each dataset independent of the underlying LLM. **Bold** entries mark the best method per metric within each dataset.

Dataset	**Method**	**ECE↓**	**Brier↓**	**Acc↑**	**F1↑**	**Spec↑**	**AUCPR↑**	**AUROC↑**
BBH	CE-DT	0.103 ±0.02	0.202 ±0.03	0.729 ±0.06	0.795 ±0.13	0.253 ±0.21	0.792 ±0.15	0.651 ±0.04
	CE-KNN	0.075 ±0.02	0.181 ±0.03	0.746 ±0.05	0.803 ±0.14	0.236 ±0.27	0.803 ±0.14	0.682 ±0.01
	CE-LogReg	0.070 ±0.03	0.183 ±0.02	0.749 ±0.04	0.800 ±0.16	0.144 ±0.34	0.809 ±0.14	0.696 ±0.02
	CE-RF	0.049 ±0.02	0.173 ±0.02	0.756 ±0.04	0.804 ±0.14	0.243 ±0.31	0.811 ±0.14	0.696 ±0.02
	CE-XGB	0.071 ±0.02	0.177 ±0.02	0.748 ±0.05	0.806 ±0.14	0.245 ±0.25	0.809 ±0.14	0.694 ±0.01
	ETTIN	0.061 ±0.05	0.143 ±0.03	0.803 ±0.05	**0.853** ±0.09	0.305 ±0.31	**0.885** ±0.07	**0.742** ±0.04
	ETTIN-HGA	0.046 ±0.01	**0.142** ±0.02	**0.805** ±0.04	0.853 ±0.11	0.232 ±0.32	0.872 ±0.09	0.723 ±0.05
	GNN-CD-ft-APPNP	0.185 ±0.11	0.227 ±0.08	0.709 ±0.07	0.771 ±0.12	0.503 ±0.20	0.829 ±0.13	0.662 ±0.05
	GNN-CD-ft-GCN2Conv-dual	0.107 ±0.08	0.167 ±0.05	0.784 ±0.05	0.825 ±0.14	0.368 ±0.28	0.864 ±0.10	0.710 ±0.06
	GNN-CD-ft-GCN2Conv-same	0.146 ±0.04	0.183 ±0.02	0.740 ±0.04	0.812 ±0.08	0.463 ±0.16	0.844 ±0.10	0.688 ±0.08
	GNN-CD-ft-TAGConv	0.268 ±0.14	0.306 ±0.13	0.669 ±0.12	0.739 ±0.14	0.530 ±0.12	0.846 ±0.13	0.645 ±0.09
	GNN-CD-noft-APPNP	0.146 ±0.10	0.197 ±0.05	0.760 ±0.05	0.813 ±0.12	0.398 ±0.21	0.845 ±0.13	0.681 ±0.04
	GNN-CD-noft-GCN2Conv-dual	0.078 ±0.05	0.158 ±0.02	0.771 ±0.04	0.824 ±0.10	0.432 ±0.23	0.868 ±0.11	0.735 ±0.03
	GNN-CD-noft-GCN2Conv-same	0.145 ±0.12	0.200 ±0.08	0.734 ±0.09	0.790 ±0.13	0.450 ±0.27	0.854 ±0.12	0.701 ±0.03
	GNN-CD-noft-TAGConv	0.201 ±0.07	0.254 ±0.08	0.651 ±0.14	0.713 ±0.17	0.493 ±0.18	0.823 ±0.15	0.615 ±0.07
	GNN-SR-GINE	0.149 ±0.14	0.185 ±0.07	0.704 ±0.20	0.735 ±0.25	0.471 ±0.35	0.864 ±0.10	0.715 ±0.05
	GNN-SR-NNConv	0.219 ±0.12	0.204 ±0.05	0.707 ±0.13	0.755 ±0.16	**0.572** ±0.27	0.873 ±0.10	0.738 ±0.04
	GNN-SR-Transformer	0.100 ±0.03	0.150 ±0.03	0.796 ±0.05	0.835 ±0.13	0.365 ±0.29	0.873 ±0.09	0.734 ±0.05
	GNN-SB-GAT	0.062 ±0.02	0.150 ±0.02	0.784 ±0.05	0.837 ±0.10	0.361 ±0.23	0.864 ±0.11	0.721 ±0.05
	GNN-SB-GCN	**0.043** ±0.02	0.144 ±0.03	0.793 ±0.05	0.844 ±0.10	0.347 ±0.24	0.869 ±0.10	0.733 ±0.03
	GNN-SB-GraphSAGE	0.044 ±0.02	0.143 ±0.02	0.793 ±0.05	0.842 ±0.10	0.361 ±0.25	0.877 ±0.09	0.741 ±0.05
	LateFusion-ft-Conv	0.223 ±0.22	0.255 ±0.19	0.672 ±0.20	0.704 ±0.23	0.517 ±0.33	0.846 ±0.13	0.686 ±0.06
	LateFusion-noft-Conv	0.169 ±0.05	0.189 ±0.02	0.715 ±0.05	0.763 ±0.15	0.541 ±0.16	0.847 ±0.14	0.692 ±0.03
	LateFusion-ft-LSTM	0.100 ±0.04	0.170 ±0.02	0.766 ±0.04	0.819 ±0.11	0.416 ±0.23	0.856 ±0.11	0.703 ±0.05
	LateFusion-noft-LSTM	0.115 ±0.03	0.175 ±0.02	0.761 ±0.04	0.815 ±0.12	0.395 ±0.22	0.851 ±0.11	0.687 ±0.05
	LateFusion-ft-MLP	0.100 ±0.04	0.157 ±0.02	0.778 ±0.04	0.821 ±0.12	0.461 ±0.24	0.867 ±0.11	0.740 ±0.03
	LateFusion-noft-MLP	0.092 ±0.04	0.155 ±0.02	0.782 ±0.05	0.826 ±0.12	0.464 ±0.22	0.865 ±0.10	0.735 ±0.03
	PHSV	0.105 ±0.08	0.195 ±0.03	0.711 ±0.05	0.767 ±0.14	0.331 ±0.28	0.786 ±0.13	0.673 ±0.02
	PHSV-half	0.107 ±0.08	0.196 ±0.03	0.717 ±0.04	0.772 ±0.13	0.325 ±0.29	0.782 ±0.12	0.667 ±0.02
	P(IK)	0.065 ±0.03	0.159 ±0.03	0.784 ±0.07	0.847 ±0.12	0.145 ±0.24	0.822 ±0.15	0.654 ±0.04
	SFHS-Conv	0.061 ±0.01	0.151 ±0.02	0.783 ±0.05	0.836 ±0.10	0.393 ±0.22	0.869 ±0.09	0.716 ±0.05
	SFHS-LSTM	0.078 ±0.01	0.159 ±0.02	0.777 ±0.04	0.827 ±0.11	0.387 ±0.23	0.853 ±0.11	0.695 ±0.06
	SFHS-MLP	0.068 ±0.02	0.148 ±0.02	0.786 ±0.04	0.835 ±0.10	0.400 ±0.23	0.876 ±0.10	0.738 ±0.02
	TLCC-CONV	0.101 ±0.05	0.166 ±0.02	0.756 ±0.04	0.817 ±0.08	0.442 ±0.23	0.867 ±0.09	0.699 ±0.07
	TLCC-LSTM	0.076 ±0.05	0.152 ±0.02	0.782 ±0.04	0.837 ±0.09	0.400 ±0.26	0.876 ±0.09	0.727 ±0.05
	TLCC-MLP	0.126 ±0.06	0.167 ±0.02	0.750 ±0.04	0.811 ±0.08	0.490 ±0.20	0.866 ±0.09	0.708 ±0.06
	YVCE	0.146 ±0.18	0.216 ±0.15	0.729 ±0.18	0.823 ±0.15	0.111 ±0.12	0.857 ±0.13	0.638 ±0.09
FinQA	CE-DT	0.295 ±0.06	0.300 ±0.10	0.539 ±0.17	0.424 ±0.20	0.402 ±0.27	0.340 ±0.16	0.565 ±0.02
	CE-KNN	0.296 ±0.07	0.289 ±0.10	0.520 ±0.21	0.423 ±0.18	0.350 ±0.36	0.346 ±0.16	0.573 ±0.03
	CE-LogReg	0.369 ±0.07	0.329 ±0.11	0.449 ±0.24	0.431 ±0.18	0.190 ±0.39	0.347 ±0.16	0.582 ±0.03
	CE-RF	0.289 ±0.08	0.276 ±0.10	0.521 ±0.23	0.435 ±0.15	0.349 ±0.37	0.348 ±0.15	0.589 ±0.02
	CE-XGB	0.280 ±0.07	0.279 ±0.10	0.535 ±0.19	0.422 ±0.18	0.393 ±0.31	0.333 ±0.15	0.565 ±0.02
	ETTIN	0.204 ±0.07	0.232 ±0.10	0.584 ±0.20	**0.482** ±0.15	0.431 ±0.35	0.434 ±0.15	0.642 ±0.05
	ETTIN-HGA	0.163 ±0.04	0.214 ±0.08	0.633 ±0.18	0.471 ±0.17	0.546 ±0.32	0.442 ±0.15	**0.655** ±0.04
	GNN-CD-ft-APPNP	0.257 ±0.13	0.287 ±0.12	0.631 ±0.13	0.317 ±0.13	0.749 ±0.14	0.372 ±0.17	0.552 ±0.01
	GNN-CD-ft-GCN2Conv-dual	0.268 ±0.15	0.306 ±0.16	0.609 ±0.19	0.368 ±0.16	0.636 ±0.30	0.402 ±0.13	0.563 ±0.04
	GNN-CD-ft-GCN2Conv-same	0.224 ±0.12	0.269 ±0.13	0.643 ±0.16	0.287 ±0.23	0.722 ±0.32	0.360 ±0.14	0.567 ±0.03
	GNN-CD-ft-TAGConv	0.256 ±0.11	0.316 ±0.13	0.638 ±0.15	0.301 ±0.18	0.746 ±0.21	0.366 ±0.16	0.541 ±0.03
	GNN-CD-noft-APPNP	0.251 ±0.14	0.285 ±0.12	0.640 ±0.16	0.344 ±0.14	0.727 ±0.22	0.386 ±0.17	0.562 ±0.02
	GNN-CD-noft-GCN2Conv-dual	0.153 ±0.04	0.225 ±0.08	0.662 ±0.15	0.269 ±0.16	0.825 ±0.15	0.373 ±0.14	0.564 ±0.02
	GNN-CD-noft-GCN2Conv-same	0.229 ±0.15	0.278 ±0.13	0.627 ±0.16	0.292 ±0.16	0.749 ±0.25	0.382 ±0.16	0.566 ±0.02
	GNN-CD-noft-TAGConv	0.247 ±0.11	0.315 ±0.12	0.611 ±0.13	0.312 ±0.14	0.724 ±0.19	0.364 ±0.15	0.528 ±0.05
	GNN-SR-GINE	0.191 ±0.10	0.238 ±0.04	0.651 ±0.16	0.339 ±0.24	0.694 ±0.32	0.389 ±0.12	0.600 ±0.05
	GNN-SR-NNConv	0.121 ±0.10	0.217 ±0.05	**0.675** ±0.14	0.152 ±0.15	**0.913** ±0.10	0.383 ±0.13	0.586 ±0.04
	GNN-SR-Transformer	0.150 ±0.06	0.223 ±0.06	0.651 ±0.16	0.394 ±0.15	0.693 ±0.21	0.399 ±0.11	0.595 ±0.04
	GNN-SB-GAT	0.131 ±0.04	0.221 ±0.08	0.649 ±0.16	0.347 ±0.16	0.731 ±0.19	0.365 ±0.13	0.565 ±0.03
	GNN-SB-GCN	0.111 ±0.05	0.213 ±0.08	0.661 ±0.15	0.328 ±0.11	0.791 ±0.15	0.377 ±0.15	0.572 ±0.03
	GNN-SB-GraphSAGE	**0.098** ±0.04	**0.208** ±0.08	0.658 ±0.15	0.319 ±0.10	0.808 ±0.11	0.391 ±0.11	0.594 ±0.06
	LateFusion-ft-Conv	0.220 ±0.12	0.265 ±0.11	0.633 ±0.15	0.357 ±0.24	0.646 ±0.28	0.372 ±0.15	0.565 ±0.02
	LateFusion-noft-Conv	0.210 ±0.11	0.262 ±0.09	0.624 ±0.09	0.356 ±0.19	0.681 ±0.20	0.368 ±0.15	0.567 ±0.04
	LateFusion-ft-LSTM	0.209 ±0.08	0.259 ±0.11	0.623 ±0.17	0.357 ±0.15	0.667 ±0.25	0.349 ±0.14	0.571 ±0.05
	LateFusion-noft-LSTM	0.258 ±0.13	0.295 ±0.14	0.604 ±0.19	0.401 ±0.14	0.592 ±0.27	0.350 ±0.12	0.566 ±0.06
	LateFusion-ft-MLP	0.164 ±0.09	0.234 ±0.11	0.668 ±0.15	0.292 ±0.14	0.823 ±0.17	0.377 ±0.13	0.560 ±0.02
	LateFusion-noft-MLP	0.153 ±0.10	0.231 ±0.11	0.673 ±0.16	0.256 ±0.17	0.832 ±0.22	0.381 ±0.14	0.575 ±0.03
	PHSV	0.255 ±0.10	0.279 ±0.12	0.554 ±0.19	0.379 ±0.15	0.498 ±0.24	0.294 ±0.13	0.534 ±0.07
	PHSV-half	0.273 ±0.09	0.285 ±0.11	0.549 ±0.19	0.385 ±0.15	0.477 ±0.25	0.302 ±0.14	0.530 ±0.06
	P(IK)	0.140 ±0.11	0.226 ±0.10	0.591 ±0.22	0.354 ±0.22	0.558 ±0.41	0.376 ±0.13	0.577 ±0.07
	SFHS-Conv	0.158 ±0.03	0.228 ±0.09	0.652 ±0.15	0.425 ±0.13	0.686 ±0.19	0.388 ±0.12	0.595 ±0.03
	SFHS-LSTM	0.177 ±0.04	0.240 ±0.10	0.634 ±0.16	0.378 ±0.14	0.685 ±0.19	0.361 ±0.12	0.570 ±0.03
	SFHS-MLP	0.155 ±0.04	0.231 ±0.09	0.653 ±0.16	0.359 ±0.08	0.768 ±0.14	0.378 ±0.11	0.561 ±0.01
	TLCC-CONV	0.250 ±0.07	0.275 ±0.11	0.579 ±0.18	0.461 ±0.16	0.462 ±0.26	0.367 ±0.16	0.578 ±0.05
	TLCC-LSTM	0.270 ±0.11	0.288 ±0.12	0.558 ±0.20	0.458 ±0.15	0.421 ±0.30	0.359 ±0.15	0.585 ±0.05
	TLCC-MLP	0.231 ±0.09	0.265 ±0.11	0.569 ±0.20	0.469 ±0.12	0.452 ±0.30	0.364 ±0.15	0.575 ±0.07
	YVCE	0.504 ±0.19	0.491 ±0.17	0.376 ±0.15	0.450 ±0.21	0.129 ±0.18	**0.488** ±0.13	0.532 ±0.05
LegalBench	CE-DT	0.130 ±0.06	0.265 ±0.03	0.605 ±0.06	0.675 ±0.13	0.338 ±0.17	0.633 ±0.10	0.546 ±0.02
	CE-KNN	0.131 ±0.06	0.256 ±0.03	0.610 ±0.06	0.670 ±0.16	0.286 ±0.28	0.635 ±0.11	0.556 ±0.03
	CE-LogReg	0.106 ±0.06	0.244 ±0.03	0.611 ±0.09	0.705 ±0.15	0.147 ±0.30	0.637 ±0.10	0.564 ±0.02
	CE-RF	0.098 ±0.05	0.243 ±0.02	0.614 ±0.06	0.669 ±0.17	0.291 ±0.30	0.636 ±0.10	0.562 ±0.03
	CE-XGB	0.114 ±0.06	0.249 ±0.03	0.611 ±0.06	0.670 ±0.15	0.316 ±0.25	0.635 ±0.10	0.560 ±0.03

(Continued on next page)

(Continued from previous page)

Dataset	Method	ECE↓	Brier↓	Acc↑	F1↑	Spec↑	AUCPR↑	AUROC↑
	ETTIN	**0.083** ±0.04	**0.228** ±0.03	**0.641** ±0.07	0.720 ±0.14	0.272 ±0.21	0.710 ±0.12	**0.610** ±0.04
	ETTIN-HGA	0.084 ±0.02	0.228 ±0.03	0.640 ±0.08	**0.725** ±0.14	0.234 ±0.21	0.705 ±0.12	0.604 ±0.05
	GNN-CD-ft-APPNP	0.272 ±0.18	0.343 ±0.12	0.535 ±0.08	0.532 ±0.15	0.603 ±0.27	0.679 ±0.10	0.556 ±0.03
	GNN-CD-ft-GCN2Conv-dual	0.229 ±0.13	0.305 ±0.08	0.571 ±0.07	0.626 ±0.11	0.495 ±0.25	0.709 ±0.08	0.584 ±0.03
	GNN-CD-ft-GCN2Conv-same	0.251 ±0.14	0.318 ±0.07	0.536 ±0.11	0.518 ±0.21	0.580 ±0.33	0.683 ±0.10	0.570 ±0.03
	GNN-CD-ft-TAGConv	0.324 ±0.11	0.391 ±0.09	0.527 ±0.09	0.552 ±0.17	0.547 ±0.25	0.688 ±0.10	0.565 ±0.03
	GNN-CD-noft-APPNP	0.286 ±0.15	0.338 ±0.11	0.533 ±0.09	0.544 ±0.16	0.572 ±0.27	0.681 ±0.09	0.552 ±0.01
	GNN-CD-noft-GCN2Conv-dual	0.194 ±0.07	0.278 ±0.02	0.547 ±0.03	0.556 ±0.11	0.645 ±0.10	0.692 ±0.10	0.580 ±0.03
	GNN-CD-noft-GCN2Conv-same	0.305 ±0.16	0.352 ±0.12	0.510 ±0.12	0.496 ±0.23	0.582 ±0.32	0.685 ±0.09	0.568 ±0.02
	GNN-CD-noft-TAGConv	0.257 ±0.09	0.336 ±0.07	0.524 ±0.09	0.528 ±0.19	0.605 ±0.19	0.683 ±0.09	0.563 ±0.03
	GNN-SR-GINE	0.138 ±0.12	0.255 ±0.05	0.555 ±0.13	0.529 ±0.31	0.532 ±0.36	0.678 ±0.09	0.562 ±0.06
	GNN-SR-NNConv	0.195 ±0.10	0.279 ±0.02	0.468 ±0.11	0.369 ±0.29	**0.698** ±0.26	0.679 ±0.09	0.555 ±0.09
	GNN-SR-Transformer	0.144 ±0.08	0.250 ±0.02	0.575 ±0.07	0.586 ±0.16	0.572 ±0.27	0.687 ±0.09	0.572 ±0.05
	GNN-SB-GAT	0.176 ±0.09	0.268 ±0.03	0.567 ±0.09	0.589 ±0.16	0.537 ±0.28	0.684 ±0.09	0.568 ±0.03
	GNN-SB-GCN	0.170 ±0.09	0.267 ±0.03	0.581 ±0.06	0.618 ±0.13	0.527 ±0.25	0.688 ±0.09	0.578 ±0.03
	GNN-SB-GraphSAGE	0.182 ±0.09	0.266 ±0.02	0.579 ±0.06	0.611 ±0.13	0.548 ±0.25	0.688 ±0.09	0.571 ±0.06
	LateFusion-ft-Conv	0.272 ±0.18	0.334 ±0.14	0.545 ±0.12	0.483 ±0.31	0.625 ±0.31	0.687 ±0.11	0.575 ±0.03
	LateFusion-noft-Conv	0.191 ±0.08	0.273 ±0.02	0.569 ±0.06	0.561 ±0.20	0.614 ±0.20	0.688 ±0.11	0.585 ±0.04
	LateFusion-ft-LSTM	0.224 ±0.10	0.300 ±0.05	0.552 ±0.09	0.567 ±0.17	0.548 ±0.27	0.671 ±0.09	0.546 ±0.08
	LateFusion-noft-LSTM	0.210 ±0.08	0.290 ±0.04	0.565 ±0.07	0.597 ±0.14	0.518 ±0.21	0.677 ±0.09	0.566 ±0.02
	LateFusion-ft-MLP	0.213 ±0.11	0.290 ±0.03	0.540 ±0.04	0.528 ±0.12	0.698 ±0.11	0.692 ±0.10	0.586 ±0.02
	LateFusion-noft-MLP	0.207 ±0.10	0.287 ±0.04	0.547 ±0.08	0.527 ±0.17	0.639 ±0.28	0.694 ±0.10	0.589 ±0.02
	PHSV	0.156 ±0.07	0.269 ±0.01	0.545 ±0.02	0.584 ±0.07	0.514 ±0.15	0.633 ±0.10	0.547 ±0.02
	PHSV-half	0.163 ±0.05	0.270 ±0.01	0.538 ±0.02	0.584 ±0.07	0.504 ±0.14	0.631 ±0.10	0.541 ±0.03
	P(IK)	0.163 ±0.10	0.262 ±0.05	0.566 ±0.08	0.643 ±0.15	0.369 ±0.29	0.657 ±0.13	0.526 ±0.09
	SFHS-Conv	0.179 ±0.08	0.270 ±0.03	0.577 ±0.08	0.612 ±0.15	0.517 ±0.26	0.683 ±0.09	0.574 ±0.05
	SFHS-LSTM	0.203 ±0.06	0.284 ±0.01	0.557 ±0.02	0.595 ±0.12	0.525 ±0.20	0.670 ±0.09	0.551 ±0.05
	SFHS-MLP	0.203 ±0.11	0.288 ±0.04	0.548 ±0.09	0.555 ±0.17	0.573 ±0.27	0.679 ±0.09	0.569 ±0.04
	TLCC-CONV	0.144 ±0.05	0.250 ±0.03	0.623 ±0.06	0.701 ±0.13	0.349 ±0.22	0.686 ±0.10	0.581 ±0.03
	TLCC-LSTM	0.144 ±0.05	0.252 ±0.03	0.614 ±0.05	0.694 ±0.11	0.368 ±0.22	0.686 ±0.10	0.583 ±0.03
	TLCC-MLP	0.128 ±0.05	0.244 ±0.03	0.609 ±0.05	0.684 ±0.11	0.409 ±0.21	0.687 ±0.10	0.583 ±0.03
	YVCE	0.183 ±0.05	0.258 ±0.05	0.625 ±0.08	0.723 ±0.13	0.206 ±0.21	**0.748** ±0.13	0.604 ±0.08
MATH	CE-DT	0.114 ±0.06	0.127 ±0.02	0.847 ±0.02	0.793 ±0.29	0.416 ±0.25	0.793 ±0.32	0.698 ±0.05
	CE-KNN	0.086 ±0.07	0.101 ±0.02	0.867 ±0.03	0.802 ±0.30	0.375 ±0.34	0.821 ±0.32	0.791 ±0.07
	CE-LogReg	0.139 ±0.10	0.131 ±0.03	0.855 ±0.06	0.789 ±0.31	0.190 ±0.38	0.814 ±0.33	0.773 ±0.13
	CE-RF	0.083 ±0.06	0.098 ±0.01	0.877 ±0.04	0.806 ±0.29	0.353 ±0.39	0.822 ±0.32	**0.795** ±0.09
	CE-XGB	0.088 ±0.07	0.104 ±0.02	0.870 ±0.03	0.807 ±0.29	0.363 ±0.34	0.823 ±0.31	0.789 ±0.07
	ETTIN	0.047 ±0.03	**0.059** ±0.02	0.933 ±0.02	**0.885** ±0.19	0.214 ±0.38	0.889 ±0.19	0.728 ±0.08
	ETTIN-HGA	0.059 ±0.04	0.063 ±0.01	**0.935** ±0.02	0.863 ±0.25	0.172 ±0.40	0.882 ±0.21	0.728 ±0.08
	GNN-CD-ft-APPNP	0.215 ±0.07	0.144 ±0.06	0.846 ±0.06	0.797 ±0.28	0.468 ±0.28	0.837 ±0.31	0.701 ±0.08
	GNN-CD-ft-GCN2Conv-dual	0.107 ±0.06	0.085 ±0.04	0.915 ±0.05	0.838 ±0.28	0.297 ±0.40	0.859 ±0.26	0.714 ±0.09
	GNN-CD-ft-GCN2Conv-same	0.143 ±0.11	0.107 ±0.06	0.868 ±0.05	0.825 ±0.26	0.430 ±0.27	0.846 ±0.28	0.713 ±0.08
	GNN-CD-ft-TAGConv	0.205 ±0.24	0.206 ±0.24	0.785 ±0.24	0.733 ±0.32	0.425 ±0.34	0.834 ±0.29	0.579 ±0.09
	GNN-CD-noft-APPNP	0.140 ±0.07	0.098 ±0.05	0.900 ±0.05	0.824 ±0.30	0.408 ±0.34	0.839 ±0.32	0.705 ±0.07
	GNN-CD-noft-GCN2Conv-dual	0.105 ±0.06	0.090 ±0.04	0.889 ±0.05	0.833 ±0.25	0.405 ±0.35	0.855 ±0.27	0.733 ±0.06
	GNN-CD-noft-GCN2Conv-same	0.107 ±0.06	0.090 ±0.04	0.891 ±0.05	0.813 ±0.31	0.428 ±0.33	0.835 ±0.33	0.720 ±0.06
	GNN-CD-noft-TAGConv	0.163 ±0.14	0.183 ±0.15	0.735 ±0.25	0.688 ±0.33	**0.497** ±0.33	0.829 ±0.33	0.627 ±0.07
	GNN-SR-GINE	0.187 ±0.11	0.112 ±0.05	0.865 ±0.09	0.824 ±0.23	0.381 ±0.36	0.856 ±0.25	0.684 ±0.08
	GNN-SR-NNConv	0.297 ±0.15	0.166 ±0.06	0.850 ±0.10	0.816 ±0.23	0.455 ±0.35	0.865 ±0.25	0.728 ±0.06
	GNN-SR-Transformer	0.170 ±0.05	0.091 ±0.02	0.912 ±0.05	0.843 ±0.26	0.333 ±0.39	0.870 ±0.24	0.742 ±0.06
	GNN-SB-GAT	0.063 ±0.05	0.068 ±0.02	0.914 ±0.04	0.845 ±0.26	0.307 ±0.37	0.852 ±0.28	0.737 ±0.04
	GNN-SB-GCN	0.056 ±0.05	0.064 ±0.02	0.919 ±0.04	0.851 ±0.26	0.287 ±0.37	0.861 ±0.26	0.738 ±0.07
	GNN-SB-GraphSAGE	0.068 ±0.05	0.068 ±0.03	0.909 ±0.06	0.849 ±0.24	0.335 ±0.41	0.867 ±0.25	0.741 ±0.06
	LateFusion-ft-Conv	0.228 ±0.17	0.175 ±0.16	0.785 ±0.19	0.733 ±0.32	0.379 ±0.30	0.808 ±0.35	0.608 ±0.12
	LateFusion-noft-Conv	0.181 ±0.04	0.128 ±0.04	0.831 ±0.05	0.773 ±0.34	0.480 ±0.19	0.815 ±0.36	0.656 ±0.12
	LateFusion-ft-LSTM	0.120 ±0.07	0.091 ±0.05	0.899 ±0.05	0.829 ±0.28	0.361 ±0.37	0.844 ±0.29	0.709 ±0.04
	LateFusion-noft-LSTM	0.098 ±0.05	0.091 ±0.04	0.889 ±0.04	0.831 ±0.26	0.385 ±0.33	0.848 ±0.28	0.687 ±0.05
	LateFusion-ft-MLP	0.112 ±0.07	0.081 ±0.03	0.912 ±0.04	0.838 ±0.27	0.348 ±0.36	0.859 ±0.25	0.734 ±0.05
	LateFusion-noft-MLP	0.107 ±0.06	0.079 ±0.03	0.915 ±0.04	0.842 ±0.27	0.367 ±0.35	0.859 ±0.26	0.736 ±0.05
	PHSV	0.138 ±0.09	0.157 ±0.03	0.822 ±0.05	0.778 ±0.26	0.349 ±0.30	0.735 ±0.27	0.642 ±0.07
	PHSV-half	0.157 ±0.09	0.166 ±0.04	0.818 ±0.06	0.770 ±0.28	0.340 ±0.33	0.728 ±0.27	0.636 ±0.03
	P(IK)	**0.035** ±0.03	0.060 ±0.02	0.934 ±0.02	0.811 ±0.38	0.167 ±0.41	0.843 ±0.28	0.680 ±0.08
	SFHS-Conv	0.076 ±0.06	0.072 ±0.03	0.905 ±0.05	0.845 ±0.25	0.347 ±0.37	0.856 ±0.27	0.731 ±0.04
	SFHS-LSTM	0.086 ±0.07	0.079 ±0.04	0.903 ±0.05	0.842 ±0.25	0.357 ±0.37	0.847 ±0.29	0.710 ±0.05
	SFHS-MLP	0.070 ±0.04	0.069 ±0.02	0.909 ±0.04	0.853 ±0.23	0.362 ±0.36	0.865 ±0.24	0.719 ±0.05
	TLCC-CONV	0.085 ±0.03	0.076 ±0.02	0.902 ±0.04	0.864 ±0.19	0.348 ±0.35	0.866 ±0.22	0.694 ±0.10
	TLCC-LSTM	0.075 ±0.03	0.073 ±0.02	0.908 ±0.04	0.862 ±0.21	0.298 ±0.35	0.866 ±0.22	0.686 ±0.08
	TLCC-MLP	0.075 ±0.02	0.070 ±0.02	0.917 ±0.03	0.870 ±0.20	0.285 ±0.35	0.857 ±0.23	0.660 ±0.08
	YVCE	0.218 ±0.33	0.212 ±0.31	0.761 ±0.34	0.805 ±0.34	0.170 ±0.19	**0.896** ±0.18	0.632 ±0.09
MedMCQA	CE-DT	0.513 ±0.06	0.493 ±0.07	0.322 ±0.02	0.412 ±0.04	0.111 ±0.06	0.308 ±0.03	0.509 ±0.02
	CE-KNN	0.498 ±0.07	0.468 ±0.06	0.324 ±0.06	0.411 ±0.05	0.108 ±0.14	0.294 ±0.03	0.525 ±0.01
	CE-LogReg	0.481 ±0.06	0.433 ±0.07	0.286 ±0.05	0.417 ±0.04	0.037 ±0.08	0.278 ±0.03	0.526 ±0.02
	CE-RF	0.489 ±0.07	0.452 ±0.07	0.314 ±0.03	0.413 ±0.04	0.091 ±0.10	0.276 ±0.02	0.522 ±0.02
	CE-XGB	0.511 ±0.07	0.477 ±0.07	0.316 ±0.03	0.412 ±0.04	0.095 ±0.10	0.273 ±0.02	0.517 ±0.02
	ETTIN	0.511 ±0.11	0.488 ±0.10	0.334 ±0.06	0.416 ±0.05	0.122 ±0.14	0.281 ±0.03	0.524 ±0.03
	ETTIN-HGA	0.487 ±0.10	0.463 ±0.09	0.348 ±0.08	0.414 ±0.05	0.146 ±0.19	0.288 ±0.03	0.529 ±0.03
	GNN-CD-ft-APPNP	0.397 ±0.06	0.392 ±0.06	0.444 ±0.09	0.379 ±0.05	0.369 ±0.21	0.300 ±0.04	0.522 ±0.02
	GNN-CD-ft-GCN2Conv-dual	0.446 ±0.10	0.438 ±0.11	0.369 ±0.06	0.415 ±0.03	0.196 ±0.12	0.322 ±0.07	0.525 ±0.03
	GNN-CD-ft-GCN2Conv-same	0.407 ±0.14	0.412 ±0.11	0.422 ±0.12	0.390 ±0.05	0.314 ±0.25	0.287 ±0.03	0.524 ±0.02
	GNN-CD-ft-TAGConv	0.467 ±0.13	0.479 ±0.12	0.456 ±0.09	0.378 ±0.04	0.395 ±0.18	**0.364** ±0.06	0.509 ±0.01
	GNN-CD-noft-APPNP	0.486 ±0.09	0.468 ±0.10	0.363 ±0.05	0.415 ±0.04	0.185 ±0.11	0.335 ±0.09	0.526 ±0.02
	GNN-CD-noft-GCN2Conv-dual	0.411 ±0.08	0.402 ±0.06	0.390 ±0.08	0.408 ±0.05	0.238 ±0.17	0.285 ±0.03	0.524 ±0.02
	GNN-CD-noft-GCN2Conv-same	0.473 ±0.09	0.462 ±0.08	0.379 ±0.09	0.401 ±0.04	0.233 ±0.19	0.315 ±0.04	0.526 ±0.02

(Continued on next page)

(Continued from previous page)

Dataset	Method	ECE↓	Brier↓	Acc↑	F1↑	Spec↑	AUCPR↑	AUROC↑
	GNN-CD-noft-TAGConv	0.423 ±0.11	0.447 ±0.11	0.447 ±0.05	0.383 ±0.04	0.376 ±0.11	0.339 ±0.03	0.508 ±0.01
	GNN-SR-GINE	0.366 ±0.16	0.363 ±0.10	0.420 ±0.17	0.344 ±0.17	0.311 ±0.38	0.283 ±0.02	0.522 ±0.02
	GNN-SR-NNConv	**0.312** ±0.16	**0.322** ±0.12	0.454 ±0.18	0.341 ±0.15	0.379 ±0.36	0.278 ±0.01	0.515 ±0.03
	GNN-SR-Transformer	0.396 ±0.07	0.369 ±0.06	0.362 ±0.07	0.416 ±0.05	0.173 ±0.17	0.284 ±0.02	0.524 ±0.03
	GNN-SB-GAT	0.461 ±0.10	0.441 ±0.09	0.363 ±0.08	0.412 ±0.05	0.180 ±0.18	0.281 ±0.02	0.522 ±0.03
	GNN-SB-GCN	0.465 ±0.08	0.445 ±0.07	0.357 ±0.07	0.414 ±0.04	0.171 ±0.14	0.281 ±0.02	0.521 ±0.02
	GNN-SB-GraphSAGE	0.482 ±0.08	0.460 ±0.07	0.350 ±0.07	0.418 ±0.04	0.151 ±0.15	0.283 ±0.02	0.523 ±0.03
	LateFusion-ft-Conv	0.353 ±0.16	0.362 ±0.11	0.467 ±0.18	0.325 ±0.17	0.399 ±0.40	0.276 ±0.02	0.514 ±0.02
	LateFusion-noft-Conv	0.321 ±0.11	0.340 ±0.08	**0.469** ±0.09	0.375 ±0.06	**0.412** ±0.18	0.279 ±0.02	0.516 ±0.02
	LateFusion-ft-LSTM	0.454 ±0.12	0.450 ±0.10	0.389 ±0.08	0.401 ±0.05	0.241 ±0.19	0.279 ±0.02	0.519 ±0.02
	LateFusion-noft-LSTM	0.471 ±0.09	0.465 ±0.09	0.374 ±0.07	0.407 ±0.04	0.209 ±0.15	0.277 ±0.02	0.512 ±0.02
	LateFusion-ft-MLP	0.411 ±0.10	0.402 ±0.09	0.407 ±0.09	0.405 ±0.04	0.274 ±0.19	0.283 ±0.02	0.524 ±0.02
	LateFusion-noft-MLP	0.418 ±0.10	0.411 ±0.08	0.398 ±0.09	0.405 ±0.05	0.255 ±0.19	0.289 ±0.02	0.528 ±0.02
	PHSV	0.406 ±0.09	0.386 ±0.07	0.374 ±0.10	0.387 ±0.05	0.227 ±0.22	0.267 ±0.02	0.513 ±0.03
	PHSV-half	0.446 ±0.10	0.423 ±0.09	0.349 ±0.07	0.396 ±0.04	0.177 ±0.15	0.265 ±0.02	0.508 ±0.02
	P(IK)	0.525 ±0.08	0.482 ±0.08	0.281 ±0.02	**0.425** ±0.03	0.018 ±0.03	0.272 ±0.02	0.508 ±0.02
	SFHS-Conv	0.457 ±0.08	0.446 ±0.07	0.379 ±0.06	0.410 ±0.04	0.216 ±0.15	0.285 ±0.02	0.524 ±0.03
	SFHS-LSTM	0.445 ±0.07	0.434 ±0.06	0.380 ±0.06	0.411 ±0.04	0.218 ±0.13	0.278 ±0.02	0.517 ±0.02
	SFHS-MLP	0.452 ±0.09	0.443 ±0.08	0.389 ±0.08	0.410 ±0.05	0.232 ±0.18	0.284 ±0.02	0.525 ±0.03
	TLCC-CONV	0.378 ±0.07	0.380 ±0.05	0.419 ±0.06	0.400 ±0.04	0.304 ±0.12	0.279 ±0.02	0.521 ±0.02
	TLCC-LSTM	0.378 ±0.07	0.377 ±0.05	0.409 ±0.05	0.401 ±0.04	0.288 ±0.11	0.277 ±0.02	0.519 ±0.02
	TLCC-MLP	0.327 ±0.07	0.338 ±0.04	0.445 ±0.06	0.386 ±0.04	0.369 ±0.12	0.277 ±0.02	0.517 ±0.02
	YVCE	0.490 ±0.06	0.453 ±0.06	0.320 ±0.04	0.423 ±0.04	0.090 ±0.11	0.359 ±0.07	**0.534** ±0.02
MMLU-Pro	CE-DT	0.100 ±0.02	0.208 ±0.04	0.715 ±0.07	0.784 ±0.09	0.376 ±0.14	0.781 ±0.10	0.696 ±0.03
	CE-KNN	0.074 ±0.04	0.180 ±0.04	0.733 ±0.07	0.796 ±0.10	0.344 ±0.25	0.823 ±0.11	0.759 ±0.05
	CE-LogReg	0.088 ±0.02	0.199 ±0.04	0.690 ±0.10	0.781 ±0.10	0.155 ±0.30	0.830 ±0.11	0.765 ±0.04
	CE-RF	0.059 ±0.02	0.174 ±0.03	0.735 ±0.07	0.798 ±0.09	0.335 ±0.28	0.833 ±0.10	0.770 ±0.04
	CE-XGB	0.079 ±0.03	0.180 ±0.03	0.731 ±0.07	0.798 ±0.09	0.347 ±0.21	0.830 ±0.10	0.761 ±0.04
	ETTIN	0.054 ±0.02	**0.154** ±0.03	0.770 ±0.04	**0.833** ±0.07	0.397 ±0.18	**0.881** ±0.07	**0.789** ±0.04
	ETTIN-HGA	**0.046** ±0.01	0.154 ±0.03	**0.772** ±0.05	0.833 ±0.08	0.363 ±0.21	0.879 ±0.07	0.783 ±0.03
	GNN-CD-ft-APPNP	0.197 ±0.13	0.251 ±0.07	0.672 ±0.05	0.731 ±0.06	**0.691** ±0.09	0.834 ±0.10	0.732 ±0.02
	GNN-CD-ft-GCN2Conv-dual	0.159 ±0.14	0.210 ±0.12	0.733 ±0.10	0.798 ±0.10	0.496 ±0.24	0.861 ±0.09	0.749 ±0.09
	GNN-CD-ft-GCN2Conv-same	0.166 ±0.06	0.205 ±0.02	0.707 ±0.04	0.769 ±0.06	0.632 ±0.16	0.853 ±0.08	0.748 ±0.05
	GNN-CD-ft-TAGConv	0.279 ±0.09	0.314 ±0.08	0.653 ±0.07	0.719 ±0.09	0.629 ±0.11	0.814 ±0.12	0.669 ±0.08
	GNN-CD-noft-APPNP	0.175 ±0.11	0.218 ±0.07	0.733 ±0.05	0.795 ±0.08	0.555 ±0.10	0.851 ±0.09	0.744 ±0.04
	GNN-CD-noft-GCN2Conv-dual	0.102 ±0.09	0.183 ±0.03	0.720 ±0.06	0.778 ±0.08	0.591 ±0.15	0.871 ±0.08	0.769 ±0.03
	GNN-CD-noft-GCN2Conv-same	0.168 ±0.13	0.219 ±0.08	0.710 ±0.08	0.774 ±0.08	0.592 ±0.21	0.860 ±0.09	0.761 ±0.03
	GNN-CD-noft-TAGConv	0.210 ±0.05	0.281 ±0.08	0.639 ±0.09	0.694 ±0.12	0.603 ±0.17	0.813 ±0.12	0.663 ±0.05
	GNN-SR-GINE	0.133 ±0.14	0.196 ±0.07	0.684 ±0.17	0.718 ±0.22	0.561 ±0.27	0.865 ±0.08	0.761 ±0.03
	GNN-SR-NNConv	0.171 ±0.10	0.205 ±0.04	0.699 ±0.09	0.720 ±0.18	0.668 ±0.24	0.866 ±0.08	0.763 ±0.03
	GNN-SR-Transformer	0.088 ±0.03	0.165 ±0.03	0.764 ±0.06	0.813 ±0.10	0.459 ±0.24	0.872 ±0.07	0.769 ±0.02
	GNN-SB-GAT	0.057 ±0.01	0.161 ±0.03	0.755 ±0.06	0.814 ±0.09	0.464 ±0.14	0.871 ±0.08	0.772 ±0.03
	GNN-SB-GCN	0.051 ±0.03	0.159 ±0.03	0.760 ±0.07	0.817 ±0.09	0.473 ±0.13	0.869 ±0.09	0.777 ±0.03
	GNN-SB-GraphSAGE	0.050 ±0.01	0.157 ±0.03	0.762 ±0.07	0.817 ±0.09	0.474 ±0.18	0.875 ±0.08	0.783 ±0.03
	LateFusion-ft-Conv	0.197 ±0.18	0.249 ±0.15	0.666 ±0.15	0.684 ±0.21	0.548 ±0.25	0.819 ±0.12	0.690 ±0.11
	LateFusion-noft-Conv	0.163 ±0.08	0.211 ±0.02	0.681 ±0.04	0.725 ±0.10	0.669 ±0.15	0.836 ±0.10	0.724 ±0.03
	LateFusion-ft-LSTM	0.116 ±0.04	0.188 ±0.03	0.735 ±0.05	0.790 ±0.09	0.566 ±0.15	0.853 ±0.09	0.748 ±0.03
	LateFusion-noft-LSTM	0.124 ±0.04	0.193 ±0.02	0.733 ±0.04	0.792 ±0.08	0.514 ±0.14	0.851 ±0.08	0.732 ±0.03
	LateFusion-ft-MLP	0.114 ±0.05	0.175 ±0.02	0.741 ±0.05	0.789 ±0.08	0.651 ±0.11	0.867 ±0.09	0.781 ±0.02
	LateFusion-noft-MLP	0.088 ±0.06	0.169 ±0.02	0.752 ±0.04	0.801 ±0.08	0.620 ±0.09	0.866 ±0.08	0.780 ±0.03
	PHSV	0.123 ±0.10	0.219 ±0.05	0.651 ±0.12	0.704 ±0.15	0.418 ±0.22	0.745 ±0.10	0.679 ±0.06
	PHSV-half	0.131 ±0.09	0.220 ±0.04	0.658 ±0.09	0.714 ±0.13	0.408 ±0.26	0.746 ±0.08	0.685 ±0.03
	P(IK)	0.073 ±0.03	0.172 ±0.03	0.747 ±0.08	0.828 ±0.08	0.163 ±0.24	0.844 ±0.09	0.728 ±0.04
	SFHS-Conv	0.061 ±0.02	0.162 ±0.03	0.759 ±0.05	0.812 ±0.08	0.559 ±0.11	0.871 ±0.08	0.778 ±0.02
	SFHS-LSTM	0.083 ±0.02	0.176 ±0.02	0.743 ±0.05	0.800 ±0.08	0.543 ±0.11	0.862 ±0.08	0.760 ±0.02
	SFHS-MLP	0.070 ±0.03	0.165 ±0.03	0.756 ±0.05	0.808 ±0.09	0.558 ±0.12	0.873 ±0.08	0.779 ±0.03
	TLCC-CONV	0.106 ±0.08	0.184 ±0.02	0.713 ±0.05	0.771 ±0.07	0.636 ±0.07	0.864 ±0.08	0.760 ±0.04
	TLCC-LSTM	0.101 ±0.08	0.179 ±0.02	0.721 ±0.05	0.783 ±0.07	0.567 ±0.16	0.863 ±0.09	0.764 ±0.03
	TLCC-MLP	0.127 ±0.09	0.185 ±0.01	0.706 ±0.03	0.763 ±0.06	0.682 ±0.04	0.856 ±0.09	0.761 ±0.03
	YVCE	0.134 ±0.14	0.214 ±0.12	0.708 ±0.16	0.805 ±0.13	0.150 ±0.17	0.825 ±0.15	0.680 ±0.13

Table 45: Performance for each LLM–method pair averaged across datasets. Metrics are *unweighted* means ± standard deviations across datasets, emphasizing which methods work best for a given LLM after averaging out dataset effects. **Bold** entries mark the best method per metric within each LLM.

Method	**ECE↓**	**Brier↓**	**Acc↑**	**F1↑**	**Spec↑**	**AUCPR↑**	**AUROC↑**
			`Phi-4-mini-flash-reasoning`				
CE-DT	0.181 ±0.11	0.252 ±0.08	0.588 ±0.18	0.607 ±0.23	0.355 ±0.11	0.600 ±0.28	0.602 ±0.07
CE-KNN	0.176 ±0.10	0.239 ±0.08	0.626 ±0.15	0.591 ±0.25	0.469 ±0.13	0.586 ±0.30	0.618 ±0.09
CE-LogReg	0.224 ±0.13	0.249 ±0.07	0.513 ±0.26	0.644 ±0.23	0.002 ±0.00	0.601 ±0.30	0.630 ±0.11
CE-RF	0.151 ±0.13	0.222 ±0.08	0.613 ±0.18	0.606 ±0.25	0.388 ±0.16	0.600 ±0.30	0.634 ±0.10
CE-XGB	0.153 ±0.13	0.223 ±0.08	0.616 ±0.17	0.605 ±0.25	0.404 ±0.16	0.601 ±0.30	0.632 ±0.10
ETTIN	0.149 ±0.09	0.207 ±0.08	0.648 ±0.17	0.606 ±0.26	0.485 ±0.20	0.625 ±0.30	0.634 ±0.11
ETTIN-HGA	**0.106** ±0.09	**0.194** ±0.08	**0.685** ±0.16	0.613 ±0.28	0.469 ±0.28	0.630 ±0.31	0.641 ±0.12
GNN-CD-ft-APPNP	0.164 ±0.11	0.223 ±0.07	0.641 ±0.15	0.602 ±0.25	0.482 ±0.18	0.611 ±0.29	0.616 ±0.08
GNN-CD-ft-GCN2Conv-dual	0.389 ±0.19	0.398 ±0.20	0.581 ±0.20	0.640 ±0.23	0.204 ±0.10	**0.687** ±0.22	0.554 ±0.04
GNN-CD-ft-GCN2Conv-same	0.141 ±0.08	0.211 ±0.08	0.678 ±0.13	0.564 ±0.29	0.601 ±0.25	0.621 ±0.30	0.617 ±0.10
GNN-CD-ft-TAGConv	0.199 ±0.08	0.261 ±0.09	0.650 ±0.14	0.601 ±0.26	0.474 ±0.20	0.624 ±0.28	0.579 ±0.08
GNN-CD-noft-APPNP	0.348 ±0.17	0.351 ±0.17	0.635 ±0.17	0.624 ±0.25	0.426 ±0.15	0.635 ±0.28	0.622 ±0.08
GNN-CD-noft-GCN2Conv-dual	0.116 ±0.09	0.203 ±0.08	0.668 ±0.13	0.563 ±0.29	0.593 ±0.22	0.626 ±0.31	0.632 ±0.11
GNN-CD-noft-GCN2Conv-same	0.312 ±0.18	0.329 ±0.18	0.600 ±0.21	**0.650** ±0.23	0.255 ±0.12	0.636 ±0.28	0.626 ±0.08
GNN-CD-noft-TAGConv	0.295 ±0.18	0.330 ±0.18	0.617 ±0.19	0.623 ±0.24	0.378 ±0.16	0.650 ±0.27	0.607 ±0.08
GNN-SR-GINE	0.130 ±0.10	0.202 ±0.07	0.670 ±0.13	0.549 ±0.30	0.603 ±0.26	0.623 ±0.30	0.625 ±0.10
GNN-SR-NNConv	0.205 ±0.13	0.236 ±0.03	0.622 ±0.12	0.291 ±0.33	**0.913** ±0.16	0.626 ±0.30	0.629 ±0.09
GNN-SR-Transformer	0.148 ±0.09	0.202 ±0.06	0.668 ±0.14	0.575 ±0.28	0.575 ±0.24	0.638 ±0.30	**0.650** ±0.11
GNN-SB-GAT	0.124 ±0.10	0.204 ±0.07	0.652 ±0.13	0.572 ±0.28	0.538 ±0.20	0.622 ±0.31	0.627 ±0.11
GNN-SB-GCN	0.134 ±0.10	0.212 ±0.08	0.650 ±0.14	0.575 ±0.27	0.528 ±0.22	0.618 ±0.30	0.619 ±0.10
GNN-SB-GraphSAGE	0.131 ±0.10	0.209 ±0.08	0.646 ±0.14	0.580 ±0.27	0.520 ±0.20	0.628 ±0.31	0.636 ±0.11
LateFusion-ft-Conv	0.239 ±0.12	0.255 ±0.05	0.624 ±0.09	0.309 ±0.33	0.889 ±0.18	0.610 ±0.30	0.610 ±0.10
LateFusion-noft-Conv	0.143 ±0.06	0.202 ±0.05	0.679 ±0.11	0.529 ±0.29	0.727 ±0.19	0.630 ±0.30	0.641 ±0.11
LateFusion-ft-LSTM	0.141 ±0.07	0.213 ±0.07	0.662 ±0.13	0.572 ±0.28	0.559 ±0.21	0.620 ±0.31	0.631 ±0.10
LateFusion-noft-LSTM	0.140 ±0.08	0.213 ±0.08	0.665 ±0.14	0.601 ±0.25	0.538 ±0.22	0.630 ±0.30	0.641 ±0.10
LateFusion-ft-MLP	0.158 ±0.10	0.211 ±0.06	0.671 ±0.13	0.571 ±0.28	0.580 ±0.24	0.629 ±0.30	0.640 ±0.10
LateFusion-noft-MLP	0.136 ±0.08	0.205 ±0.06	0.676 ±0.13	0.543 ±0.30	0.646 ±0.23	0.629 ±0.30	0.634 ±0.10
PHSV	0.207 ±0.13	0.261 ±0.08	0.567 ±0.16	0.596 ±0.22	0.302 ±0.14	0.572 ±0.27	0.603 ±0.08
PHSV-half	0.199 ±0.15	0.260 ±0.09	0.570 ±0.18	0.609 ±0.23	0.280 ±0.13	0.574 ±0.28	0.600 ±0.08
P(IK)	0.133 ±0.12	0.211 ±0.09	0.627 ±0.23	0.617 ±0.30	0.246 ±0.34	0.597 ±0.31	0.592 ±0.10
SFHS-Conv	0.129 ±0.10	0.212 ±0.09	0.657 ±0.15	0.592 ±0.26	0.533 ±0.21	0.627 ±0.30	0.636 ±0.11
SFHS-LSTM	0.142 ±0.10	0.215 ±0.09	0.654 ±0.15	0.591 ±0.26	0.526 ±0.22	0.627 ±0.30	0.639 ±0.10
SFHS-MLP	0.136 ±0.10	0.211 ±0.08	0.655 ±0.13	0.574 ±0.27	0.552 ±0.21	0.626 ±0.30	0.631 ±0.10
TLCC-CONV	0.164 ±0.13	0.226 ±0.09	0.627 ±0.14	0.590 ±0.24	0.554 ±0.12	0.626 ±0.31	0.637 ±0.11
TLCC-LSTM	0.164 ±0.14	0.229 ±0.10	0.627 ±0.17	0.607 ±0.25	0.487 ±0.16	0.625 ±0.31	0.636 ±0.11
TLCC-MLP	0.161 ±0.13	0.225 ±0.09	0.636 ±0.15	0.597 ±0.25	0.523 ±0.18	0.625 ±0.31	0.635 ±0.11
YVCE	0.269 ±0.09	0.296 ±0.07	0.555 ±0.14	0.580 ±0.21	0.372 ±0.11	0.604 ±0.28	0.569 ±0.07
			`Qwen3-8B`				
CE-DT	0.209 ±0.18	0.265 ±0.14	0.634 ±0.20	0.723 ±0.20	0.217 ±0.09	0.686 ±0.26	0.636 ±0.10
CE-KNN	0.204 ±0.19	0.257 ±0.14	0.610 ±0.22	0.725 ±0.20	0.073 ±0.04	0.693 ±0.26	0.661 ±0.12
CE-LogReg	0.241 ±0.21	0.278 ±0.15	0.596 ±0.23	0.723 ±0.20	0.000	0.698 ±0.26	0.684 ±0.12
CE-RF	0.201 ±0.20	0.253 ±0.14	0.600 ±0.23	0.724 ±0.20	0.019 ±0.01	0.699 ±0.27	**0.687** ±0.13
CE-XGB	0.204 ±0.19	0.259 ±0.14	0.616 ±0.21	0.722 ±0.20	0.116 ±0.05	0.683 ±0.28	0.663 ±0.13
ETTIN	0.164 ±0.20	0.224 ±0.16	0.679 ±0.22	0.752 ±0.21	0.191 ±0.13	0.722 ±0.27	0.649 ±0.08
ETTIN-HGA	0.164 ±0.16	0.214 ±0.13	0.670 ±0.24	**0.757** ±0.22	0.085 ±0.07	0.729 ±0.27	0.653 ±0.09
GNN-CD-ft-APPNP	0.385 ±0.11	0.384 ±0.11	0.606 ±0.11	0.552 ±0.25	0.678 ±0.21	0.730 ±0.25	0.591 ±0.08
GNN-CD-ft-GCN2Conv-dual	0.188 ±0.11	0.230 ±0.12	0.667 ±0.18	0.675 ±0.25	0.463 ±0.22	0.699 ±0.29	0.630 ±0.11
GNN-CD-ft-GCN2Conv-same	0.186 ±0.12	0.229 ±0.12	0.655 ±0.17	0.565 ±0.34	0.645 ±0.28	0.702 ±0.29	0.645 ±0.11
GNN-CD-ft-TAGConv	0.265 ±0.13	0.337 ±0.16	0.661 ±0.16	0.582 ±0.33	0.618 ±0.28	0.741 ±0.25	0.604 ±0.09
GNN-CD-noft-APPNP	0.342 ±0.18	0.342 ±0.18	0.655 ±0.18	0.640 ±0.27	0.514 ±0.25	**0.759** ±0.22	0.590 ±0.07
GNN-CD-noft-GCN2Conv-dual	0.165 ±0.16	0.228 ±0.13	0.658 ±0.19	0.682 ±0.25	0.369 ±0.23	0.704 ±0.29	0.632 ±0.09
GNN-CD-noft-GCN2Conv-same	0.369 ±0.15	0.364 ±0.15	0.595 ±0.16	0.502 ±0.31	0.713 ±0.23	0.726 ±0.26	0.621 ±0.09
GNN-CD-noft-TAGConv	0.252 ±0.14	0.435 ±0.05	0.545 ±0.05	0.586 ±0.15	0.522 ±0.08	0.711 ±0.24	0.543 ±0.03
GNN-SR-GINE	0.278 ±0.16	0.267 ±0.07	0.507 ±0.20	0.238 ±0.32	**0.887** ±0.24	0.703 ±0.29	0.629 ±0.11
GNN-SR-NNConv	0.258 ±0.11	0.241 ±0.03	0.671 ±0.19	0.642 ±0.28	0.519 ±0.27	0.709 ±0.29	0.647 ±0.11
GNN-SR-Transformer	0.152 ±0.12	**0.205** ±0.11	0.683 ±0.19	0.714 ±0.23	0.365 ±0.20	0.706 ±0.29	0.649 ±0.09
GNN-SB-GAT	0.155 ±0.15	0.219 ±0.13	0.680 ±0.20	0.718 ±0.23	0.309 ±0.18	0.695 ±0.29	0.628 ±0.10
GNN-SB-GCN	**0.130** ±0.17	0.211 ±0.14	0.681 ±0.22	0.699 ±0.26	0.324 ±0.29	0.704 ±0.29	0.643 ±0.10
GNN-SB-GraphSAGE	0.150 ±0.18	0.215 ±0.15	0.685 ±0.22	0.724 ±0.23	0.309 ±0.26	0.722 ±0.28	0.668 ±0.10
LateFusion-ft-Conv	0.161 ±0.12	0.213 ±0.11	0.676 ±0.19	0.728 ±0.22	0.283 ±0.13	0.703 ±0.28	0.625 ±0.08
LateFusion-noft-Conv	0.227 ±0.13	0.261 ±0.13	0.660 ±0.16	0.693 ±0.23	0.404 ±0.13	0.693 ±0.28	0.616 ±0.08
LateFusion-ft-LSTM	0.237 ±0.16	0.270 ±0.16	0.650 ±0.21	0.706 ±0.22	0.309 ±0.16	0.693 ±0.29	0.609 ±0.08
LateFusion-noft-LSTM	0.278 ±0.18	0.300 ±0.18	0.631 ±0.21	0.711 ±0.21	0.254 ±0.08	0.683 ±0.29	0.584 ±0.08
LateFusion-ft-MLP	0.194 ±0.10	0.222 ±0.11	0.660 ±0.18	0.646 ±0.26	0.548 ±0.25	0.703 ±0.30	0.649 ±0.12
LateFusion-noft-MLP	0.174 ±0.15	0.222 ±0.13	0.670 ±0.21	0.637 ±0.31	0.488 ±0.30	0.702 ±0.30	0.649 ±0.12
PHSV	0.184 ±0.15	0.261 ±0.10	0.596 ±0.18	0.682 ±0.19	0.243 ±0.11	0.634 ±0.26	0.603 ±0.09
PHSV-half	0.225 ±0.16	0.287 ±0.12	0.582 ±0.17	0.669 ±0.19	0.224 ±0.13	0.616 ±0.24	0.591 ±0.08
P(IK)	0.199 ±0.21	0.243 ±0.17	0.625 ±0.27	0.727 ±0.22	0.076 ±0.17	0.695 ±0.28	0.583 ±0.08
SFHS-Conv	0.157 ±0.14	0.219 ±0.13	**0.685** ±0.18	0.711 ±0.22	0.410 ±0.21	0.709 ±0.29	0.651 ±0.10
SFHS-LSTM	0.165 ±0.14	0.222 ±0.13	0.677 ±0.19	0.698 ±0.24	0.379 ±0.22	0.703 ±0.29	0.629 ±0.10
SFHS-MLP	0.177 ±0.13	0.226 ±0.13	0.679 ±0.19	0.688 ±0.25	0.456 ±0.20	0.706 ±0.29	0.652 ±0.11
TLCC-CONV	0.167 ±0.09	0.216 ±0.11	0.668 ±0.17	0.713 ±0.22	0.365 ±0.18	0.696 ±0.29	0.606 ±0.10
TLCC-LSTM	0.174 ±0.17	0.225 ±0.14	0.662 ±0.22	0.744 ±0.21	0.188 ±0.10	0.696 ±0.29	0.619 ±0.10
TLCC-MLP	0.171 ±0.08	0.214 ±0.09	0.654 ±0.17	0.702 ±0.21	0.390 ±0.19	0.690 ±0.29	0.601 ±0.10
YVCE	0.213 ±0.21	0.253 ±0.16	0.645 ±0.26	0.754 ±0.21	0.015 ±0.01	0.694 ±0.28	0.561 ±0.06

(Continued on next page)

(Continued from previous page)

Method	ECE↓	Brier↓	Acc↑	F1↑	Spec↑	AUCPR↑	AUROC↑
			`Qwen3-14B`				
CE-DT	0.198 ±0.20	0.257 ±0.15	0.633 ±0.22	0.726 ±0.21	0.191 ±0.10	0.675 ±0.27	0.616 ±0.08
CE-KNN	0.180 ±0.21	0.243 ±0.15	0.621 ±0.24	0.734 ±0.21	0.047 ±0.02	0.699 ±0.28	0.665 ±0.12
CE-LogReg	0.181 ±0.22	0.244 ±0.16	0.636 ±0.23	0.737 ±0.21	0.116 ±0.08	0.700 ±0.29	**0.673** ±0.13
CE-RF	0.178 ±0.21	0.243 ±0.15	0.626 ±0.24	0.737 ±0.21	0.055 ±0.03	0.694 ±0.28	0.664 ±0.11
CE-XGB	0.181 ±0.21	0.247 ±0.15	0.630 ±0.23	0.735 ±0.21	0.086 ±0.04	0.690 ±0.28	0.654 ±0.11
ETTIN	0.178 ±0.22	0.229 ±0.18	0.660 ±0.26	0.762 ±0.22	0.063 ±0.07	0.721 ±0.29	0.649 ±0.11
ETTIN-HGA	0.157 ±0.20	0.216 ±0.16	0.685 ±0.23	0.762 ±0.21	0.150 ±0.15	0.742 ±0.27	0.658 ±0.09
GNN-CD-ft-APPNP	0.186 ±0.09	0.214 ±0.10	0.669 ±0.18	0.632 ±0.28	0.572 ±0.26	0.711 ±0.28	0.632 ±0.09
GNN-CD-ft-GCN2Conv-dual	0.178 ±0.16	0.214 ±0.14	0.677 ±0.22	0.713 ±0.24	0.316 ±0.28	0.726 ±0.28	0.644 ±0.10
GNN-CD-ft-GCN2Conv-same	0.252 ±0.11	0.254 ±0.12	0.643 ±0.17	0.560 ±0.33	0.679 ±0.24	0.717 ±0.29	0.642 ±0.11
GNN-CD-ft-TAGConv	0.266 ±0.13	0.283 ±0.12	0.641 ±0.18	0.584 ±0.30	0.622 ±0.25	0.725 ±0.27	0.623 ±0.08
GNN-CD-noft-APPNP	0.174 ±0.12	0.204 ±0.12	0.647 ±0.24	0.647 ±0.29	0.407 ±0.33	0.716 ±0.28	0.629 ±0.11
GNN-CD-noft-GCN2Conv-dual	**0.156** ±0.16	0.213 ±0.14	**0.691** ±0.21	0.711 ±0.24	0.371 ±0.29	0.723 ±0.28	0.651 ±0.11
GNN-CD-noft-GCN2Conv-same	0.203 ±0.13	0.226 ±0.12	0.643 ±0.21	0.580 ±0.33	0.632 ±0.26	0.722 ±0.28	0.641 ±0.10
GNN-CD-noft-TAGConv	0.242 ±0.11	0.262 ±0.10	0.618 ±0.18	0.569 ±0.31	0.611 ±0.25	0.718 ±0.27	0.616 ±0.07
GNN-SR-GINE	0.165 ±0.15	0.212 ±0.13	0.671 ±0.22	0.711 ±0.23	0.331 ±0.26	0.723 ±0.29	0.645 ±0.09
GNN-SR-NNConv	0.257 ±0.11	0.237 ±0.03	0.595 ±0.18	0.453 ±0.39	**0.774** ±0.23	0.729 ±0.29	0.650 ±0.10
GNN-SR-Transformer	0.182 ±0.14	0.210 ±0.12	0.660 ±0.24	0.689 ±0.25	0.327 ±0.32	0.726 ±0.28	0.652 ±0.10
GNN-SB-GAT	0.163 ±0.17	0.214 ±0.14	0.658 ±0.24	0.640 ±0.30	0.414 ±0.39	0.725 ±0.29	0.648 ±0.11
GNN-SB-GCN	0.163 ±0.18	0.220 ±0.15	0.680 ±0.21	0.679 ±0.27	0.420 ±0.32	0.723 ±0.29	0.650 ±0.11
GNN-SB-GraphSAGE	0.168 ±0.19	0.222 ±0.15	0.669 ±0.22	0.671 ±0.27	0.417 ±0.33	0.724 ±0.29	0.652 ±0.11
LateFusion-ft-Conv	0.185 ±0.09	**0.203** ±0.09	0.671 ±0.20	0.704 ±0.24	0.379 ±0.24	0.723 ±0.29	0.650 ±0.10
LateFusion-noft-Conv	0.199 ±0.09	0.214 ±0.07	0.654 ±0.15	0.609 ±0.27	0.633 ±0.20	0.714 ±0.29	0.640 ±0.09
LateFusion-ft-LSTM	0.210 ±0.13	0.237 ±0.13	0.652 ±0.20	0.602 ±0.31	0.581 ±0.28	0.718 ±0.29	0.652 ±0.11
LateFusion-noft-LSTM	0.202 ±0.15	0.241 ±0.14	0.648 ±0.21	0.662 ±0.26	0.429 ±0.22	0.705 ±0.29	0.618 ±0.10
LateFusion-ft-MLP	0.189 ±0.11	0.215 ±0.11	0.682 ±0.18	0.667 ±0.27	0.523 ±0.26	0.718 ±0.28	0.643 ±0.10
LateFusion-noft-MLP	0.176 ±0.14	0.217 ±0.13	0.675 ±0.20	0.653 ±0.28	0.509 ±0.30	0.725 ±0.28	0.655 ±0.10
PHSV	0.175 ±0.15	0.245 ±0.10	0.614 ±0.18	0.674 ±0.21	0.308 ±0.21	0.660 ±0.26	0.614 ±0.08
PHSV-half	0.167 ±0.16	0.243 ±0.11	0.617 ±0.18	0.679 ±0.21	0.315 ±0.20	0.667 ±0.27	0.621 ±0.08
P(IK)	0.195 ±0.19	0.232 ±0.16	0.658 ±0.26	0.757 ±0.22	0.030 ±0.05	0.683 ±0.29	0.565 ±0.12
SFHS-Conv	0.197 ±0.18	0.237 ±0.16	0.664 ±0.23	0.664 ±0.28	0.420 ±0.31	0.716 ±0.30	0.649 ±0.11
SFHS-LSTM	0.184 ±0.14	0.230 ±0.14	0.670 ±0.20	0.694 ±0.24	0.385 ±0.23	0.707 ±0.29	0.622 ±0.09
SFHS-MLP	0.203 ±0.19	0.243 ±0.16	0.657 ±0.23	0.643 ±0.29	0.433 ±0.34	0.719 ±0.29	0.644 ±0.09
TLCC-CONV	0.178 ±0.11	0.220 ±0.11	0.661 ±0.19	0.717 ±0.21	0.353 ±0.20	0.707 ±0.30	0.616 ±0.09
TLCC-LSTM	0.182 ±0.12	0.223 ±0.12	0.665 ±0.20	0.711 ±0.22	0.381 ±0.23	0.713 ±0.30	0.629 ±0.11
TLCC-MLP	0.177 ±0.09	0.209 ±0.11	0.655 ±0.21	0.712 ±0.22	0.353 ±0.24	0.704 ±0.31	0.614 ±0.11
YVCE	0.238 ±0.24	0.262 ±0.21	0.662 ±0.26	**0.766** ±0.21	0.017 ±0.01	**0.802** ±0.21	0.645 ±0.09
			`Magistral-Small-2506`				
CE-DT	0.237 ±0.15	0.245 ±0.14	0.642 ±0.21	0.392 ±0.24	0.609 ±0.29	0.353 ±0.23	0.604 ±0.07
CE-KNN	0.226 ±0.16	0.236 ±0.14	0.666 ±0.22	0.394 ±0.22	0.660 ±0.30	0.364 ±0.24	0.621 ±0.09
CE-LogReg	0.199 ±0.15	0.219 ±0.10	0.693 ±0.22	0.362 ±0.21	0.745 ±0.28	0.356 ±0.24	0.599 ±0.08
CE-RF	0.185 ±0.15	0.209 ±0.11	0.695 ±0.24	0.399 ±0.19	0.724 ±0.31	0.367 ±0.24	0.630 ±0.08
CE-XGB	0.228 ±0.16	0.234 ±0.14	0.664 ±0.22	0.399 ±0.22	0.653 ±0.30	0.369 ±0.24	0.628 ±0.10
ETTIN	0.152 ±0.23	0.204 ±0.20	**0.716** ±0.26	**0.540** ±0.20	0.647 ±0.38	0.526 ±0.25	**0.737** ±0.12
ETTIN-HGA	0.160 ±0.20	0.201 ±0.17	0.710 ±0.25	0.493 ±0.21	0.669 ±0.36	0.496 ±0.24	0.724 ±0.12
GNN-CD-ft-APPNP	0.200 ±0.13	0.218 ±0.11	0.665 ±0.20	0.394 ±0.21	0.696 ±0.25	0.396 ±0.26	0.652 ±0.11
GNN-CD-ft-GCN2Conv-dual	0.167 ±0.10	**0.185** ±0.08	0.706 ±0.24	0.422 ±0.19	0.758 ±0.31	0.467 ±0.24	0.700 ±0.12
GNN-CD-ft-GCN2Conv-same	0.203 ±0.17	0.225 ±0.15	0.655 ±0.22	0.457 ±0.24	0.578 ±0.31	0.440 ±0.25	0.691 ±0.11
GNN-CD-ft-TAGConv	0.327 ±0.22	0.335 ±0.23	0.662 ±0.23	0.450 ±0.23	0.572 ±0.34	0.431 ±0.24	0.614 ±0.08
GNN-CD-noft-APPNP	0.174 ±0.15	0.203 ±0.11	0.679 ±0.24	0.411 ±0.22	0.695 ±0.32	0.405 ±0.27	0.664 ±0.11
GNN-CD-noft-GCN2Conv-dual	0.170 ±0.16	0.203 ±0.13	0.697 ±0.24	0.445 ±0.21	0.712 ±0.32	0.451 ±0.25	0.684 ±0.12
GNN-CD-noft-GCN2Conv-same	0.163 ±0.16	0.202 ±0.12	0.682 ±0.23	0.418 ±0.25	0.678 ±0.32	0.410 ±0.28	0.664 ±0.12
GNN-CD-noft-TAGConv	0.250 ±0.09	0.282 ±0.12	0.664 ±0.17	0.398 ±0.18	0.691 ±0.20	0.381 ±0.23	0.556 ±0.04
GNN-SR-GINE	0.242 ±0.16	0.231 ±0.08	0.699 ±0.24	0.492 ±0.20	0.644 ±0.35	0.467 ±0.23	0.692 ±0.11
GNN-SR-NNConv	0.221 ±0.16	0.217 ±0.09	0.686 ±0.24	0.480 ±0.23	0.624 ±0.33	0.473 ±0.24	0.706 ±0.11
GNN-SR-Transformer	0.201 ±0.12	0.195 ±0.08	0.712 ±0.24	0.440 ±0.19	**0.759** ±0.31	0.492 ±0.23	0.718 ±0.12
GNN-SB-GAT	0.157 ±0.17	0.199 ±0.14	0.700 ±0.24	0.463 ±0.23	0.674 ±0.33	0.453 ±0.26	0.689 ±0.11
GNN-SB-GCN	0.164 ±0.19	0.206 ±0.16	0.701 ±0.25	0.477 ±0.22	0.670 ±0.34	0.463 ±0.26	0.699 ±0.12
GNN-SB-GraphSAGE	0.151 ±0.19	0.199 ±0.16	0.709 ±0.25	0.486 ±0.20	0.697 ±0.34	0.488 ±0.23	0.717 ±0.11
LateFusion-ft-Conv	0.209 ±0.13	0.228 ±0.11	0.675 ±0.20	0.397 ±0.22	0.706 ±0.26	0.396 ±0.27	0.597 ±0.09
LateFusion-noft-Conv	0.202 ±0.12	0.228 ±0.08	0.626 ±0.15	0.337 ±0.22	0.686 ±0.21	0.375 ±0.27	0.565 ±0.10
LateFusion-ft-LSTM	0.191 ±0.17	0.215 ±0.15	0.696 ±0.24	0.441 ±0.21	0.704 ±0.32	0.429 ±0.27	0.681 ±0.09
LateFusion-noft-LSTM	0.180 ±0.17	0.218 ±0.16	0.696 ±0.23	0.444 ±0.20	0.703 ±0.30	0.438 ±0.25	0.672 ±0.09
LateFusion-ft-MLP	0.161 ±0.18	0.203 ±0.16	0.702 ±0.25	0.426 ±0.20	0.748 ±0.33	0.462 ±0.24	0.683 ±0.13
LateFusion-noft-MLP	0.158 ±0.18	0.201 ±0.15	0.703 ±0.25	0.442 ±0.20	0.729 ±0.34	0.464 ±0.25	0.700 ±0.12
PHSV	0.168 ±0.14	0.204 ±0.10	0.690 ±0.22	0.393 ±0.19	0.724 ±0.28	0.377 ±0.24	0.628 ±0.07
PHSV-half	0.216 ±0.14	0.227 ±0.11	0.691 ±0.22	0.389 ±0.20	0.721 ±0.28	0.371 ±0.24	0.604 ±0.09
P(IK)	**0.135** ±0.21	0.206 ±0.17	0.683 ±0.26	0.409 ±0.32	0.571 ±0.41	0.410 ±0.22	0.681 ±0.13
SFHS-Conv	0.154 ±0.17	0.199 ±0.15	0.711 ±0.23	0.498 ±0.19	0.689 ±0.32	0.473 ±0.25	0.697 ±0.10
SFHS-LSTM	0.164 ±0.17	0.207 ±0.15	0.696 ±0.24	0.454 ±0.21	0.681 ±0.33	0.435 ±0.25	0.672 ±0.11
SFHS-MLP	0.152 ±0.17	0.198 ±0.14	0.713 ±0.24	0.493 ±0.18	0.705 ±0.32	0.484 ±0.23	0.680 ±0.12
TLCC-CONV	0.178 ±0.17	0.204 ±0.15	0.714 ±0.24	0.525 ±0.21	0.631 ±0.39	0.467 ±0.25	0.698 ±0.13
TLCC-LSTM	0.160 ±0.17	0.196 ±0.14	0.712 ±0.24	0.519 ±0.21	0.623 ±0.39	0.466 ±0.25	0.698 ±0.12
TLCC-MLP	0.162 ±0.16	0.195 ±0.14	0.716 ±0.24	0.532 ±0.19	0.632 ±0.40	0.466 ±0.24	0.694 ±0.12
YVCE	0.557 ±0.27	0.550 ±0.23	0.320 ±0.21	0.406 ±0.27	0.096 ±0.10	**0.543** ±0.13	0.521 ±0.08
			`QwQ-32B`				
CE-DT	0.213 ±0.20	0.272 ±0.16	0.627 ±0.22	0.721 ±0.20	0.200 ±0.09	0.655 ±0.26	0.612 ±0.10
CE-KNN	0.198 ±0.20	0.254 ±0.16	0.625 ±0.23	0.728 ±0.20	0.145 ±0.09	0.695 ±0.26	0.651 ±0.12
CE-LogReg	0.242 ±0.18	0.268 ±0.14	0.615 ±0.24	0.737 ±0.20	0.000	0.685 ±0.27	0.649 ±0.12

(Continued on next page)

(Continued from previous page)

Method	**ECE↓**	**Brier↓**	**Acc↑**	**F1↑**	**Spec↑**	**AUCPR↑**	**AUROC↑**
CE-RF	0.187 ±0.20	0.249 ±0.15	0.626 ±0.23	0.729 ±0.20	0.135 ±0.09	0.685 ±0.27	0.652 ±0.13
CE-XGB	0.211 ±0.19	0.262 ±0.16	0.631 ±0.22	0.724 ±0.20	0.188 ±0.09	0.679 ±0.28	0.640 ±0.12
ETTIN	0.171 ±0.19	0.224 ±0.16	0.682 ±0.22	0.760 ±0.20	0.195 ±0.12	0.734 ±0.27	0.659 ±0.09
ETTIN-HGA	0.155 ±0.20	0.220 ±0.17	**0.698** ±0.23	**0.769** ±0.21	0.148 ±0.19	0.726 ±0.26	0.650 ±0.08
GNN-CD-ft-APPNP	0.332 ±0.11	0.331 ±0.13	0.609 ±0.12	0.645 ±0.19	0.549 ±0.11	0.706 ±0.27	0.618 ±0.09
GNN-CD-ft-GCN2Conv-dual	0.180 ±0.14	0.220 ±0.14	0.692 ±0.20	0.672 ±0.30	0.450 ±0.32	0.723 ±0.28	**0.663** ±0.11
GNN-CD-ft-GCN2Conv-same	0.292 ±0.18	0.303 ±0.18	0.635 ±0.20	0.716 ±0.20	0.315 ±0.10	0.699 ±0.27	0.596 ±0.07
GNN-CD-ft-TAGConv	0.296 ±0.14	0.320 ±0.13	0.628 ±0.15	0.689 ±0.19	0.419 ±0.10	0.722 ±0.25	0.587 ±0.07
GNN-CD-noft-APPNP	0.199 ±0.12	0.224 ±0.11	0.666 ±0.18	0.681 ±0.25	0.466 ±0.22	0.714 ±0.26	0.630 ±0.08
GNN-CD-noft-GCN2Conv-dual	0.222 ±0.09	0.237 ±0.11	0.645 ±0.16	0.592 ±0.33	**0.615** ±0.24	0.719 ±0.27	0.647 ±0.10
GNN-CD-noft-GCN2Conv-same	0.171 ±0.18	0.225 ±0.16	0.675 ±0.22	0.706 ±0.25	0.363 ±0.23	0.719 ±0.27	0.642 ±0.10
GNN-CD-noft-TAGConv	0.201 ±0.14	0.246 ±0.15	0.670 ±0.20	0.706 ±0.23	0.375 ±0.16	0.714 ±0.27	0.614 ±0.09
GNN-SR-GINE	0.182 ±0.15	0.217 ±0.14	0.667 ±0.24	0.753 ±0.21	0.171 ±0.11	0.713 ±0.27	0.625 ±0.08
GNN-SR-NNConv	0.164 ±0.21	0.230 ±0.18	0.676 ±0.23	0.705 ±0.26	0.318 ±0.29	0.716 ±0.29	0.645 ±0.12
GNN-SR-Transformer	0.184 ±0.12	**0.209** ±0.12	0.679 ±0.24	0.726 ±0.24	0.253 ±0.26	0.714 ±0.28	0.638 ±0.11
GNN-SB-GAT	0.169 ±0.19	0.230 ±0.17	0.674 ±0.23	0.707 ±0.24	0.350 ±0.24	0.714 ±0.29	0.650 ±0.12
GNN-SB-GCN	0.145 ±0.17	0.212 ±0.15	0.687 ±0.22	0.701 ±0.26	0.373 ±0.28	0.726 ±0.28	0.662 ±0.12
GNN-SB-GraphSAGE	**0.139** ±0.19	0.214 ±0.16	0.686 ±0.23	0.710 ±0.26	0.311 ±0.29	0.719 ±0.28	0.655 ±0.12
LateFusion-ft-Conv	0.200 ±0.20	0.245 ±0.17	0.672 ±0.23	0.752 ±0.21	0.221 ±0.13	0.721 ±0.28	0.642 ±0.09
LateFusion-noft-Conv	0.210 ±0.09	0.233 ±0.10	0.645 ±0.15	0.700 ±0.19	0.460 ±0.12	0.716 ±0.28	0.638 ±0.09
LateFusion-ft-LSTM	0.210 ±0.19	0.250 ±0.18	0.661 ±0.22	0.720 ±0.23	0.292 ±0.17	0.708 ±0.28	0.625 ±0.10
LateFusion-noft-LSTM	0.241 ±0.16	0.269 ±0.16	0.635 ±0.20	0.700 ±0.21	0.345 ±0.15	0.699 ±0.28	0.603 ±0.08
LateFusion-ft-MLP	0.178 ±0.16	0.223 ±0.15	0.678 ±0.21	0.646 ±0.32	0.442 ±0.35	0.719 ±0.28	0.652 ±0.11
LateFusion-noft-MLP	0.205 ±0.11	0.223 ±0.12	0.678 ±0.20	0.635 ±0.34	0.512 ±0.29	0.720 ±0.27	0.651 ±0.10
PHSV	0.156 ±0.17	0.241 ±0.12	0.625 ±0.20	0.708 ±0.20	0.234 ±0.13	0.666 ±0.26	0.616 ±0.09
PHSV-half	0.184 ±0.17	0.260 ±0.12	0.601 ±0.19	0.699 ±0.19	0.167 ±0.11	0.650 ±0.24	0.593 ±0.07
P(IK)	0.178 ±0.21	0.237 ±0.17	0.660 ±0.25	0.669 ±0.28	0.325 ±0.39	0.718 ±0.27	0.619 ±0.08
SFHS-Conv	0.173 ±0.17	0.226 ±0.15	0.680 ±0.21	0.734 ±0.21	0.344 ±0.16	0.721 ±0.28	0.651 ±0.11
SFHS-LSTM	0.189 ±0.16	0.238 ±0.15	0.668 ±0.21	0.702 ±0.24	0.378 ±0.21	0.709 ±0.29	0.631 ±0.11
SFHS-MLP	0.196 ±0.17	0.239 ±0.16	0.670 ±0.22	0.701 ±0.24	0.377 ±0.22	0.706 ±0.29	0.634 ±0.11
TLCC-CONV	0.234 ±0.11	0.254 ±0.11	0.629 ±0.16	0.711 ±0.18	0.355 ±0.15	0.707 ±0.27	0.610 ±0.08
TLCC-LSTM	0.200 ±0.09	0.228 ±0.09	0.641 ±0.15	0.702 ±0.19	0.410 ±0.15	0.708 ±0.27	0.619 ±0.08
TLCC-MLP	0.201 ±0.10	0.225 ±0.10	0.638 ±0.17	0.711 ±0.19	0.370 ±0.17	0.708 ±0.28	0.619 ±0.08
YVCE	0.216 ±0.23	0.250 ±0.19	0.664 ±0.26	0.768 ±0.21	0.074 ±0.07	**0.780** ±0.23	0.643 ±0.09
EXAONE-Deep-32B							
CE-DT	0.217 ±0.17	0.304 ±0.13	0.632 ±0.18	0.713 ±0.18	0.323 ±0.19	0.679 ±0.20	0.595 ±0.09
CE-KNN	0.176 ±0.19	0.245 ±0.14	0.651 ±0.22	0.733 ±0.20	0.306 ±0.26	0.686 ±0.25	0.671 ±0.15
CE-LogReg	0.165 ±0.19	0.261 ±0.10	0.588 ±0.21	0.720 ±0.18	0.000	0.676 ±0.26	0.669 ±0.14
CE-RF	0.167 ±0.19	0.239 ±0.14	0.655 ±0.21	0.730 ±0.20	0.340 ±0.26	0.680 ±0.26	0.670 ±0.15
CE-XGB	0.165 ±0.19	0.240 ±0.14	0.653 ±0.21	0.730 ±0.20	0.312 ±0.25	0.681 ±0.26	0.669 ±0.15
ETTIN	0.146 ±0.20	0.215 ±0.15	0.680 ±0.22	**0.770** ±0.20	0.159 ±0.11	0.752 ±0.26	**0.706** ±0.11
ETTIN-HGA	0.143 ±0.19	0.218 ±0.15	0.684 ±0.21	0.765 ±0.20	0.171 ±0.14	0.745 ±0.26	0.697 ±0.10
GNN-CD-ft-APPNP	0.256 ±0.12	0.273 ±0.08	0.647 ±0.18	0.703 ±0.22	0.406 ±0.31	0.697 ±0.25	0.615 ±0.11
GNN-CD-ft-GCN2Conv-dual	0.214 ±0.19	0.264 ±0.14	0.658 ±0.20	0.747 ±0.19	0.296 ±0.23	0.715 ±0.26	0.654 ±0.11
GNN-CD-ft-GCN2Conv-same	0.263 ±0.13	0.271 ±0.09	0.649 ±0.20	0.739 ±0.19	0.321 ±0.27	0.695 ±0.26	0.619 ±0.11
GNN-CD-ft-TAGConv	0.447 ±0.16	0.477 ±0.13	0.486 ±0.12	0.515 ±0.15	0.567 ±0.16	0.669 ±0.22	0.499 ±0.04
GNN-CD-noft-APPNP	0.247 ±0.17	0.280 ±0.13	0.646 ±0.20	0.732 ±0.19	0.337 ±0.28	0.709 ±0.26	0.635 ±0.11
GNN-CD-noft-GCN2Conv-dual	0.215 ±0.12	0.252 ±0.09	0.620 ±0.19	0.675 ±0.21	0.476 ±0.23	0.721 ±0.28	0.660 ±0.13
GNN-CD-noft-GCN2Conv-same	0.211 ±0.18	0.256 ±0.14	0.656 ±0.19	0.712 ±0.22	0.392 ±0.31	0.716 ±0.26	0.647 ±0.12
GNN-CD-noft-TAGConv	0.260 ±0.11	0.261 ±0.02	0.493 ±0.12	0.434 ±0.21	**0.720** ±0.23	0.679 ±0.26	0.568 ±0.10
GNN-SR-GINE	0.168 ±0.15	0.219 ±0.11	0.666 ±0.19	0.748 ±0.19	0.315 ±0.25	0.705 ±0.26	0.630 ±0.13
GNN-SR-NNConv	0.210 ±0.09	0.230 ±0.06	0.601 ±0.19	0.584 ±0.31	0.537 ±0.27	0.691 ±0.27	0.608 ±0.15
GNN-SR-Transformer	0.181 ±0.15	0.227 ±0.12	0.659 ±0.20	0.743 ±0.19	0.316 ±0.24	0.708 ±0.26	0.631 ±0.13
GNN-SB-GAT	0.183 ±0.20	0.242 ±0.16	0.668 ±0.21	0.744 ±0.20	0.295 ±0.23	0.708 ±0.26	0.643 ±0.12
GNN-SB-GCN	0.161 ±0.20	0.232 ±0.15	0.671 ±0.22	0.740 ±0.22	0.279 ±0.22	0.712 ±0.26	0.645 ±0.12
GNN-SB-GraphSAGE	0.185 ±0.19	0.243 ±0.15	0.656 ±0.19	0.684 ±0.27	0.425 ±0.35	0.700 ±0.27	0.625 ±0.13
LateFusion-ft-Conv	0.500 ±0.13	0.497 ±0.12	0.450 ±0.13	0.395 ±0.18	0.635 ±0.25	0.654 ±0.24	0.514 ±0.05
LateFusion-noft-Conv	0.253 ±0.07	0.266 ±0.08	0.626 ±0.12	0.685 ±0.18	0.487 ±0.11	0.705 ±0.27	0.641 ±0.09
LateFusion-ft-LSTM	0.234 ±0.15	0.273 ±0.12	0.645 ±0.19	0.721 ±0.20	0.355 ±0.26	0.683 ±0.26	0.599 ±0.14
LateFusion-noft-LSTM	0.234 ±0.16	0.269 ±0.13	0.651 ±0.20	0.726 ±0.21	0.345 ±0.22	0.699 ±0.27	0.631 ±0.11
LateFusion-ft-MLP	0.234 ±0.17	0.265 ±0.15	0.655 ±0.20	0.718 ±0.20	0.413 ±0.21	0.716 ±0.26	0.657 ±0.12
LateFusion-noft-MLP	0.217 ±0.19	0.263 ±0.15	0.664 ±0.21	0.747 ±0.20	0.293 ±0.21	0.713 ±0.26	0.652 ±0.10
PHSV	0.293 ±0.07	0.294 ±0.09	0.564 ±0.15	0.548 ±0.23	0.526 ±0.14	0.551 ±0.23	0.522 ±0.08
PHSV-half	0.287 ±0.07	0.285 ±0.10	0.568 ±0.15	0.576 ±0.21	0.523 ±0.13	0.574 ±0.25	0.557 ±0.09
P(IK)	0.160 ±0.20	0.232 ±0.15	0.651 ±0.23	0.728 ±0.22	0.172 ±0.24	0.712 ±0.26	0.633 ±0.09
SFHS-Conv	0.182 ±0.17	0.237 ±0.13	0.657 ±0.19	0.740 ±0.19	0.323 ±0.25	0.706 ±0.26	0.633 ±0.12
SFHS-LSTM	0.227 ±0.15	0.259 ±0.12	0.629 ±0.20	0.713 ±0.19	0.365 ±0.24	0.690 ±0.27	0.610 ±0.13
SFHS-MLP	0.156 ±0.19	0.229 ±0.15	0.668 ±0.20	0.720 ±0.23	0.371 ±0.30	0.715 ±0.27	0.650 ±0.14
TLCC-CONV	0.145 ±0.15	0.212 ±0.11	0.693 ±0.19	0.759 ±0.20	0.284 ±0.14	0.727 ±0.26	0.667 ±0.10
TLCC-LSTM	0.165 ±0.15	0.219 ±0.12	0.685 ±0.19	0.752 ±0.20	0.253 ±0.11	0.718 ±0.27	0.663 ±0.10
TLCC-MLP	**0.141** ±0.08	**0.200** ±0.08	**0.700** ±0.15	0.730 ±0.21	0.419 ±0.18	0.715 ±0.25	0.640 ±0.07
YVCE	0.182 ±0.18	0.234 ±0.14	0.671 ±0.21	0.754 ±0.20	0.281 ±0.12	**0.752** ±0.26	0.681 ±0.10

Table 46: Overall method performance aggregated across LLMs and datasets using a two-stage *unweighted* average: first average a method across datasets within each LLM, then average those LLM-level means across LLMs. We report mean ± standard deviation; **bold** marks the best method per metric.

Method	**ECE↓**	**Brier↓**	**Acc↑**	**F1↑**	**Spec↑**	**AUCPR↑**	**AUROC↑**
CE-DT	0.209 ±0.02	0.266 ±0.02	0.626 ±0.02	0.647 ±0.13	0.316 ±0.16	0.608 ±0.13	0.611 ±0.01
CE-KNN	0.193 ±0.02	0.246 ±0.01	0.633 ±0.02	0.651 ±0.14	0.283 ±0.24	0.620 ±0.13	0.648 ±0.02
CE-LogReg	0.209 ±0.03	0.253 ±0.02	0.607 ±0.06	0.654 ±0.15	0.144 ±0.30	0.619 ±0.13	0.651 ±0.03
CE-RF	0.178 ±0.02	0.236 ±0.02	0.636 ±0.03	0.654 ±0.13	0.277 ±0.27	0.621 ±0.13	0.656 ±0.02
CE-XGB	0.190 ±0.03	0.244 ±0.01	0.635 ±0.02	0.652 ±0.13	0.293 ±0.21	0.617 ±0.13	0.648 ±0.02
ETTIN	0.160 ±0.01	0.217 ±0.01	0.677 ±0.02	**0.698** ±0.10	0.290 ±0.22	0.680 ±0.09	**0.672** ±0.04
ETTIN-HGA	**0.148** ±0.02	0.211 ±0.01	**0.689** ±0.01	0.693 ±0.12	0.282 ±0.23	0.678 ±0.10	0.670 ±0.03
GNN-CD-ft-APPNP	0.254 ±0.09	0.274 ±0.07	0.640 ±0.03	0.588 ±0.11	0.564 ±0.11	0.642 ±0.13	0.620 ±0.02
GNN-CD-ft-GCN2Conv-dual	0.219 ±0.08	0.252 ±0.08	0.663 ±0.04	0.645 ±0.12	0.415 ±0.19	0.669 ±0.10	0.641 ±0.05
GNN-CD-ft-GCN2Conv-same	0.223 ±0.06	0.249 ±0.03	0.653 ±0.01	0.600 ±0.11	0.523 ±0.16	0.646 ±0.11	0.635 ±0.03
GNN-CD-ft-TAGConv	0.300 ±0.08	0.336 ±0.08	0.621 ±0.07	0.570 ±0.08	0.545 ±0.08	0.652 ±0.12	0.584 ±0.04
GNN-CD-noft-APPNP	0.247 ±0.08	0.267 ±0.07	0.655 ±0.02	0.622 ±0.11	0.474 ±0.12	0.656 ±0.13	0.628 ±0.02
GNN-CD-noft-GCN2Conv-dual	0.174 ±0.04	0.223 ±0.02	0.663 ±0.03	0.611 ±0.10	0.523 ±0.14	0.657 ±0.11	0.651 ±0.02
GNN-CD-noft-GCN2Conv-same	0.238 ±0.08	0.267 ±0.06	0.642 ±0.04	0.594 ±0.12	0.506 ±0.19	0.655 ±0.12	0.640 ±0.02
GNN-CD-noft-TAGConv	0.250 ±0.03	0.303 ±0.07	0.601 ±0.07	0.553 ±0.12	0.550 ±0.15	0.642 ±0.13	0.584 ±0.03
GNN-SR-GINE	0.194 ±0.06	0.225 ±0.02	0.646 ±0.07	0.582 ±0.20	0.492 ±0.27	0.656 ±0.10	0.641 ±0.03
GNN-SR-NNConv	0.219 ±0.04	0.232 ±0.01	0.642 ±0.04	0.526 ±0.15	**0.614** ±0.21	0.657 ±0.10	0.648 ±0.03
GNN-SR-Transformer	0.175 ±0.02	**0.208** ±0.01	0.677 ±0.02	0.648 ±0.12	0.432 ±0.19	0.664 ±0.09	0.656 ±0.03
GNN-SB-GAT	0.158 ±0.02	0.218 ±0.02	0.672 ±0.02	0.641 ±0.11	0.430 ±0.15	0.653 ±0.10	0.648 ±0.02
GNN-SB-GCN	0.150 ±0.02	0.216 ±0.01	0.678 ±0.02	0.645 ±0.10	0.432 ±0.14	0.657 ±0.10	0.653 ±0.03
GNN-SB-GraphSAGE	0.154 ±0.02	0.217 ±0.01	0.675 ±0.02	0.643 ±0.09	0.446 ±0.15	0.664 ±0.09	0.659 ±0.03
LateFusion-ft-Conv	0.249 ±0.13	0.273 ±0.11	0.628 ±0.09	0.548 ±0.20	0.519 ±0.26	0.635 ±0.12	0.606 ±0.05
LateFusion-noft-Conv	0.206 ±0.04	0.234 ±0.03	0.648 ±0.02	0.592 ±0.14	0.566 ±0.13	0.639 ±0.13	0.623 ±0.03
LateFusion-ft-LSTM	0.204 ±0.04	0.243 ±0.03	0.661 ±0.02	0.627 ±0.11	0.467 ±0.17	0.642 ±0.11	0.633 ±0.03
LateFusion-noft-LSTM	0.213 ±0.05	0.252 ±0.03	0.654 ±0.02	0.641 ±0.11	0.436 ±0.16	0.642 ±0.10	0.625 ±0.03
LateFusion-ft-MLP	0.186 ±0.03	0.223 ±0.02	0.674 ±0.02	0.612 ±0.10	0.542 ±0.12	0.658 ±0.10	0.654 ±0.02
LateFusion-noft-MLP	0.178 ±0.03	0.222 ±0.02	0.678 ±0.01	0.610 ±0.10	0.530 ±0.15	0.659 ±0.10	0.657 ±0.02
PHSV	0.197 ±0.05	0.251 ±0.03	0.609 ±0.05	0.600 ±0.12	0.389 ±0.20	0.577 ±0.11	0.598 ±0.04
PHSV-half	0.213 ±0.04	0.260 ±0.02	0.605 ±0.05	0.603 ±0.11	0.372 ±0.21	0.575 ±0.11	0.594 ±0.02
P(IK)	0.167 ±0.03	0.227 ±0.01	0.651 ±0.02	0.651 ±0.13	0.237 ±0.20	0.636 ±0.12	0.612 ±0.04
SFHS-Conv	0.165 ±0.02	0.222 ±0.01	0.676 ±0.02	0.656 ±0.10	0.453 ±0.14	0.659 ±0.10	0.653 ±0.02
SFHS-LSTM	0.178 ±0.03	0.229 ±0.02	0.666 ±0.02	0.642 ±0.10	0.452 ±0.13	0.645 ±0.11	0.634 ±0.02
SFHS-MLP	0.170 ±0.03	0.224 ±0.02	0.674 ±0.02	0.637 ±0.09	0.482 ±0.13	0.659 ±0.09	0.648 ±0.02
TLCC-CONV	0.178 ±0.03	0.222 ±0.02	0.665 ±0.03	0.669 ±0.09	0.424 ±0.14	0.655 ±0.10	0.639 ±0.04
TLCC-LSTM	0.174 ±0.02	0.220 ±0.01	0.665 ±0.03	0.673 ±0.09	0.390 ±0.16	0.655 ±0.10	0.644 ±0.03
TLCC-MLP	0.169 ±0.02	0.212 ±0.01	0.666 ±0.03	0.664 ±0.08	0.448 ±0.11	0.651 ±0.10	0.634 ±0.03
YVCE	0.279 ±0.14	0.307 ±0.12	0.586 ±0.14	0.671 ±0.15	0.143 ±0.15	**0.696** ±0.10	0.603 ±0.06

G Supplementary Performance Visualization with Ellipse Plots

This section provides supplementary ellipse plots to complement the main 1-ECE vs. AUROC analysis presented in Figure 1. Together, Figures 5, 6, 7, 8, 9, and 10 visualize performance trade-offs across a broad set of calibration, discrimination, and threshold-dependent evaluation metrics.

The data for each ellipse is aggregated in a two-stage process to ensure a balanced comparison. First, for each confidence estimation method, its performance on a given metric is computed separately for each of the six LRMs by taking an unweighted average across all test datasets. This yields an LRM-specific mean score. The center of each ellipse shown in Figures 5 through 10 represents the final mean performance, obtained by averaging these six LRM-specific means. The width and height of each ellipse correspond to the standard deviation of the LRM-specific means, illustrating the consistency of each method across different model architectures. For metrics where lower values indicate better performance, including ECE and Brier Score, we plot their inverse quantities such as 1-ECE and 1-Brier so that the optimal region is consistently located in the top-right corner of each plot.

H Detailed Calibration Analysis by Dataset

This section provides a more granular breakdown of the calibration performance discussed in the main paper, complementing the overall reliability diagrams shown in Figure 2. Figures 11, 12, 13, 14, 15, and 16 present dataset-specific reliability diagrams for the best-performing variant within each method family, where "best" is defined as the variant achieving the lowest average ECE on the corresponding dataset.

Each plot visualizes calibration quality by plotting the average predicted confidence for a given bin on the x-axis against the empirical accuracy of the predictions within that bin on the y-axis. The dashed diagonal line represents perfect calibration, where the predicted confidence exactly matches the observed accuracy. A method's curve deviating from this diagonal indicates miscalibration:

- **Below the diagonal:** Indicates **overconfidence**. For example, if predictions in the 50% confidence bin are correct only 40% of the time, the corresponding point falls below the line.
- **Above the diagonal:** Indicates **underconfidence**. For example, if predictions in the 60% confidence bin are correct 70% of the time, the corresponding point rises above the line.

Taken together, Figures 11 through 16 enable a direct comparison of calibration behavior across domains, illustrating how confidence reliability varies with the underlying reasoning task.

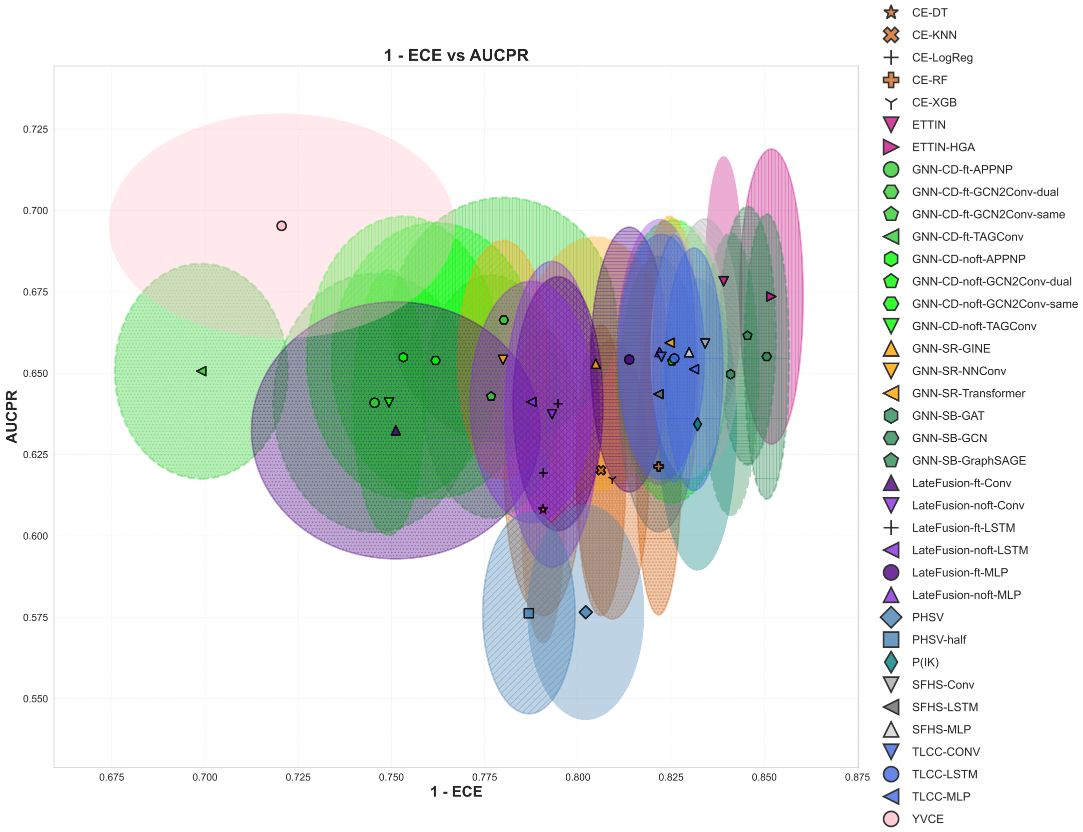

Figure 5: Performance trade-off between calibration (1-ECE) and discrimination focused on the positive class (AUCPR). This plot confirms that methods with the best calibration are not necessarily the best at identifying correct answers.

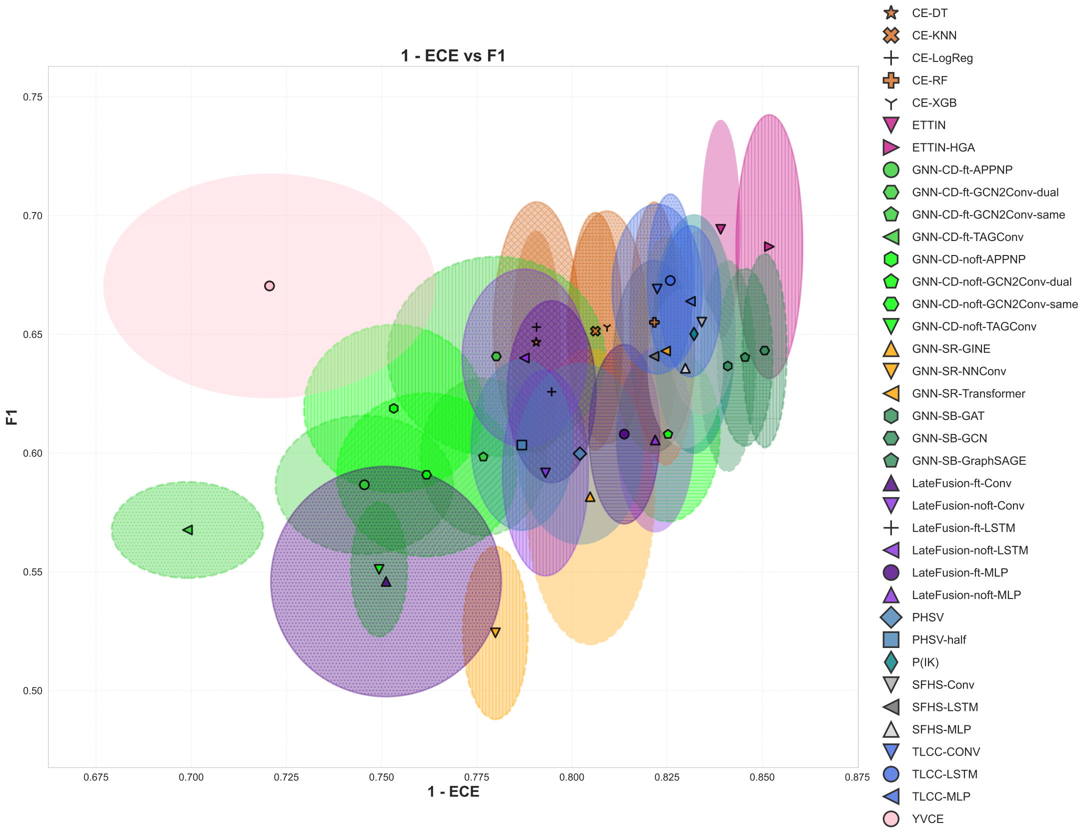

Figure 6: Performance trade-off between calibration (1-ECE) and F1 Score. This view highlights the relationship between probabilistic accuracy and the harmonic mean of precision and recall.

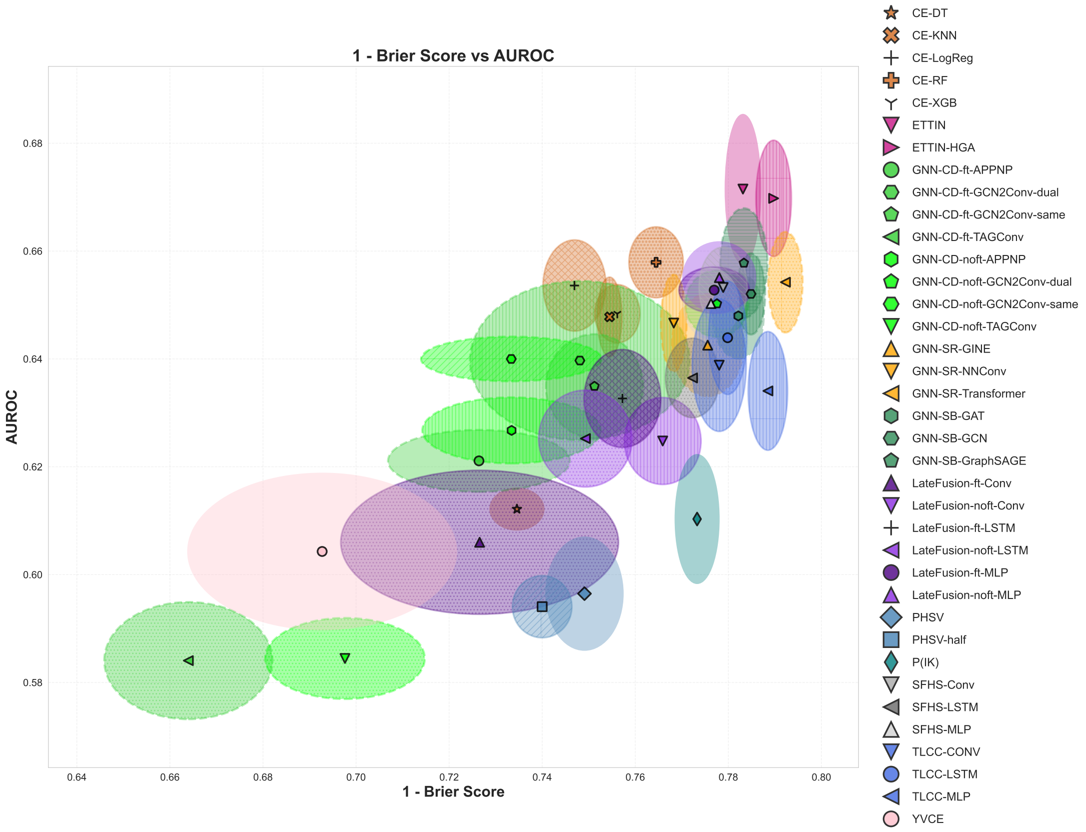

Figure 7: Performance trade-off between the Brier Score (plotted as 1-Brier), which combines calibration and discrimination, and pure discrimination (AUROC).

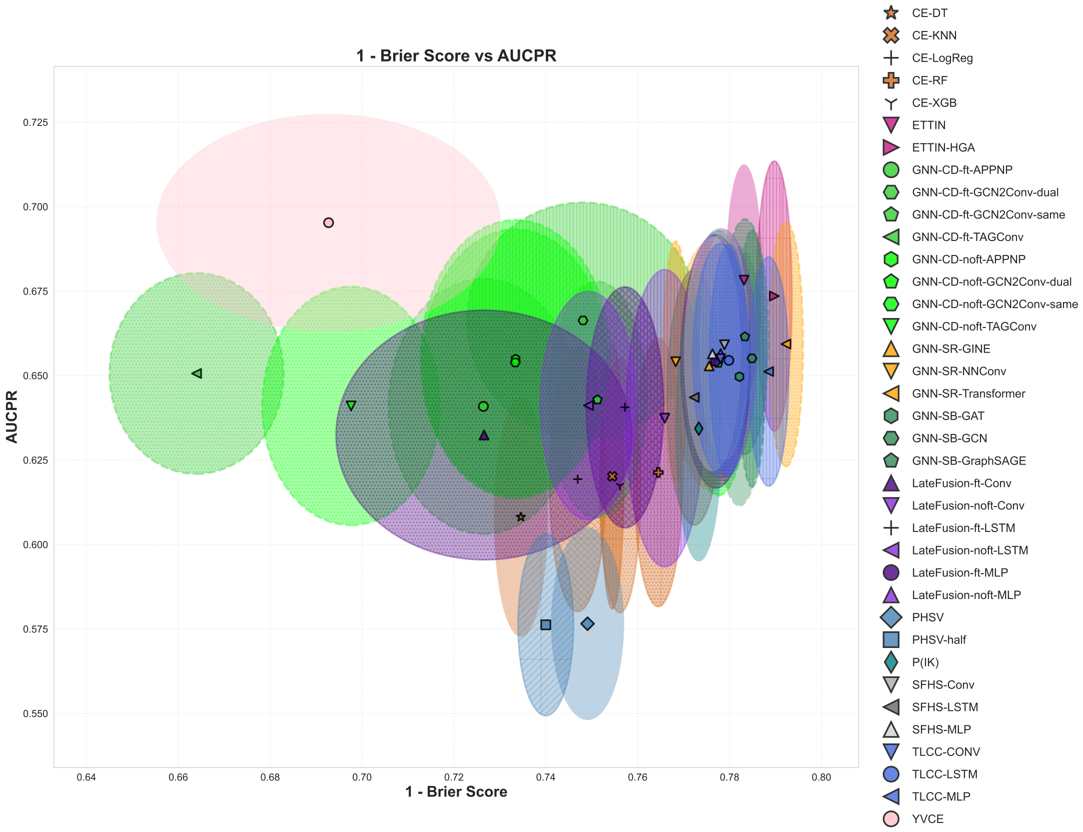

Figure 8: Performance trade-off between the Brier Score (1-Brier) and discrimination focused on the positive class (AUCPR).

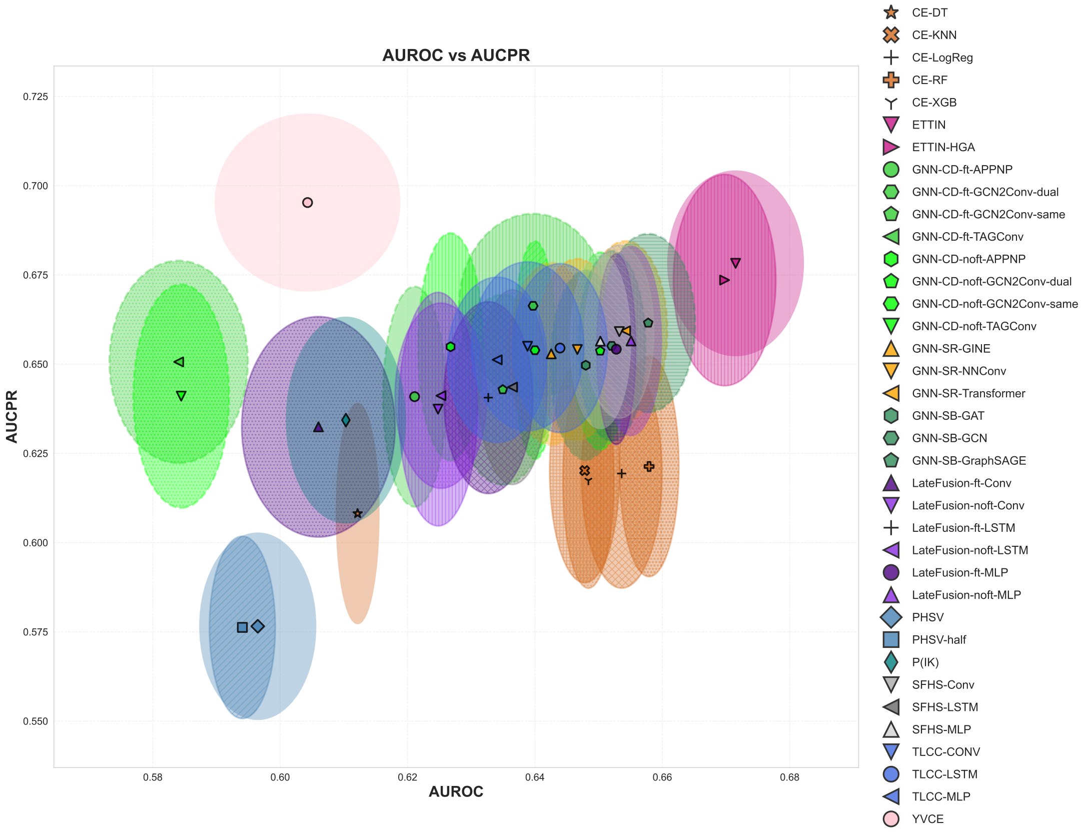

Figure 9: A comparison of the two primary discrimination metrics, AUROC and AUCPR. The strong positive correlation indicates that most methods that are good at general ranking are also good at ranking the positive class specifically.

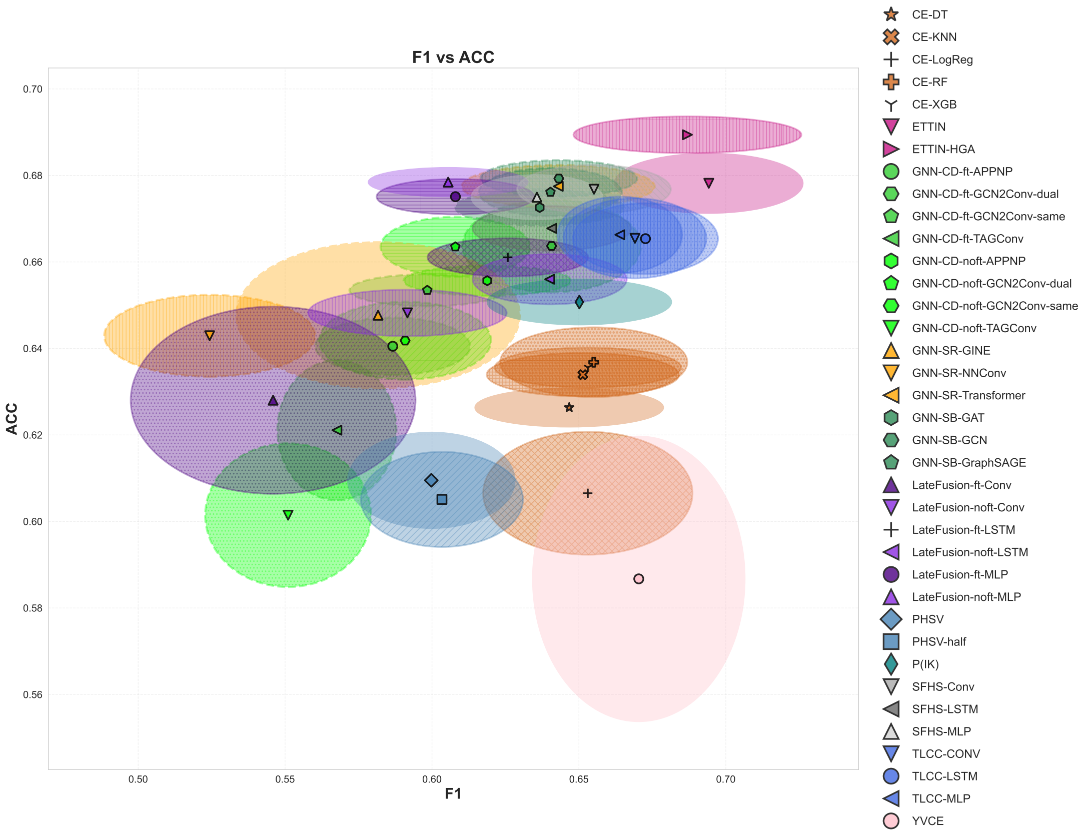

Figure 10: A comparison of two threshold-dependent classification metrics, F1 Score and Accuracy. This plot shows the relationship between balanced performance (F1) and overall correctness (Accuracy) at each method's optimal threshold.

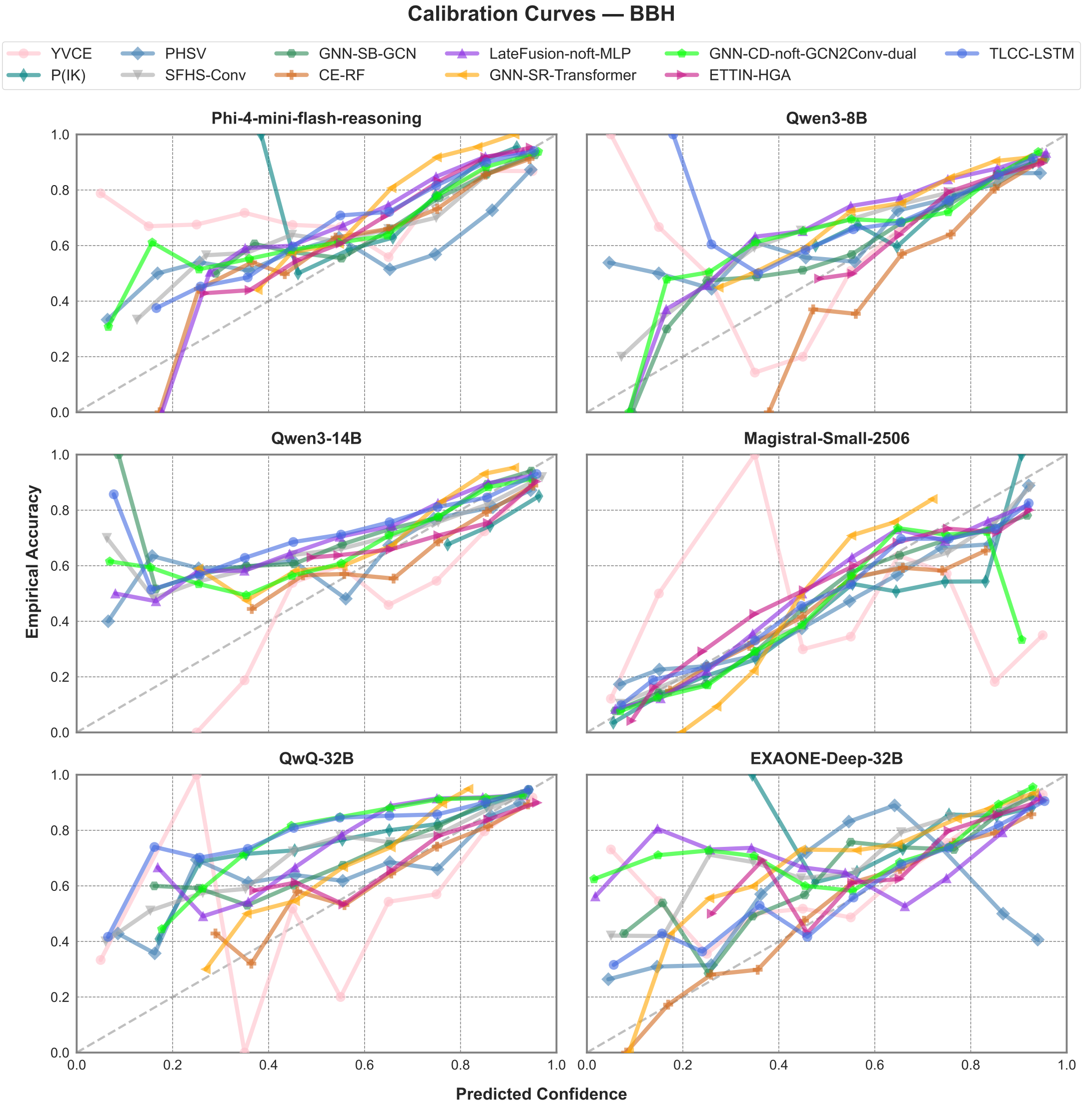

Figure 11: Reliability diagrams for the top-performing method variants on the BBH dataset, aggregated across all LRMs.

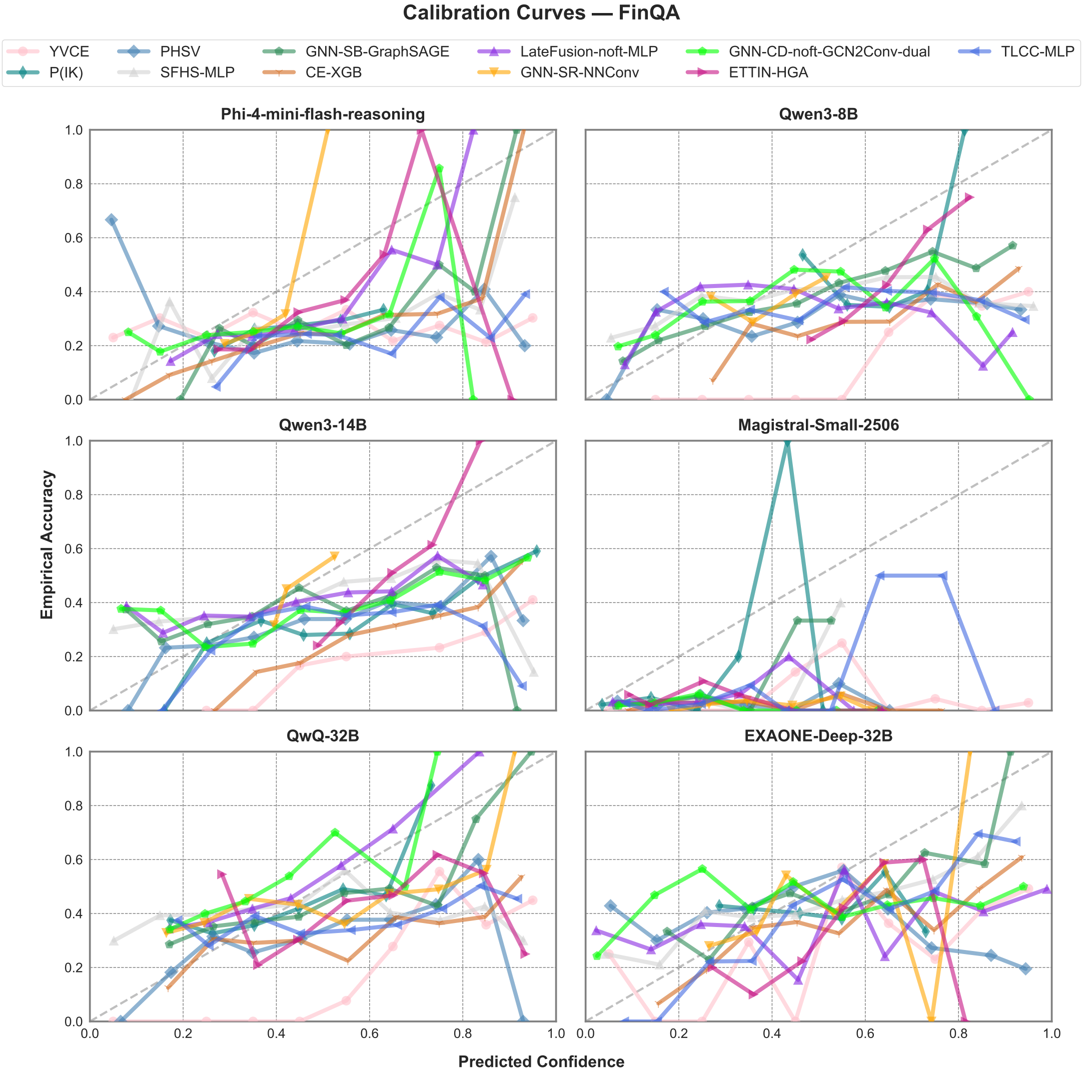

Figure 12: Reliability diagrams for the top-performing method variants on the FinQA dataset, aggregated across all LRMs.

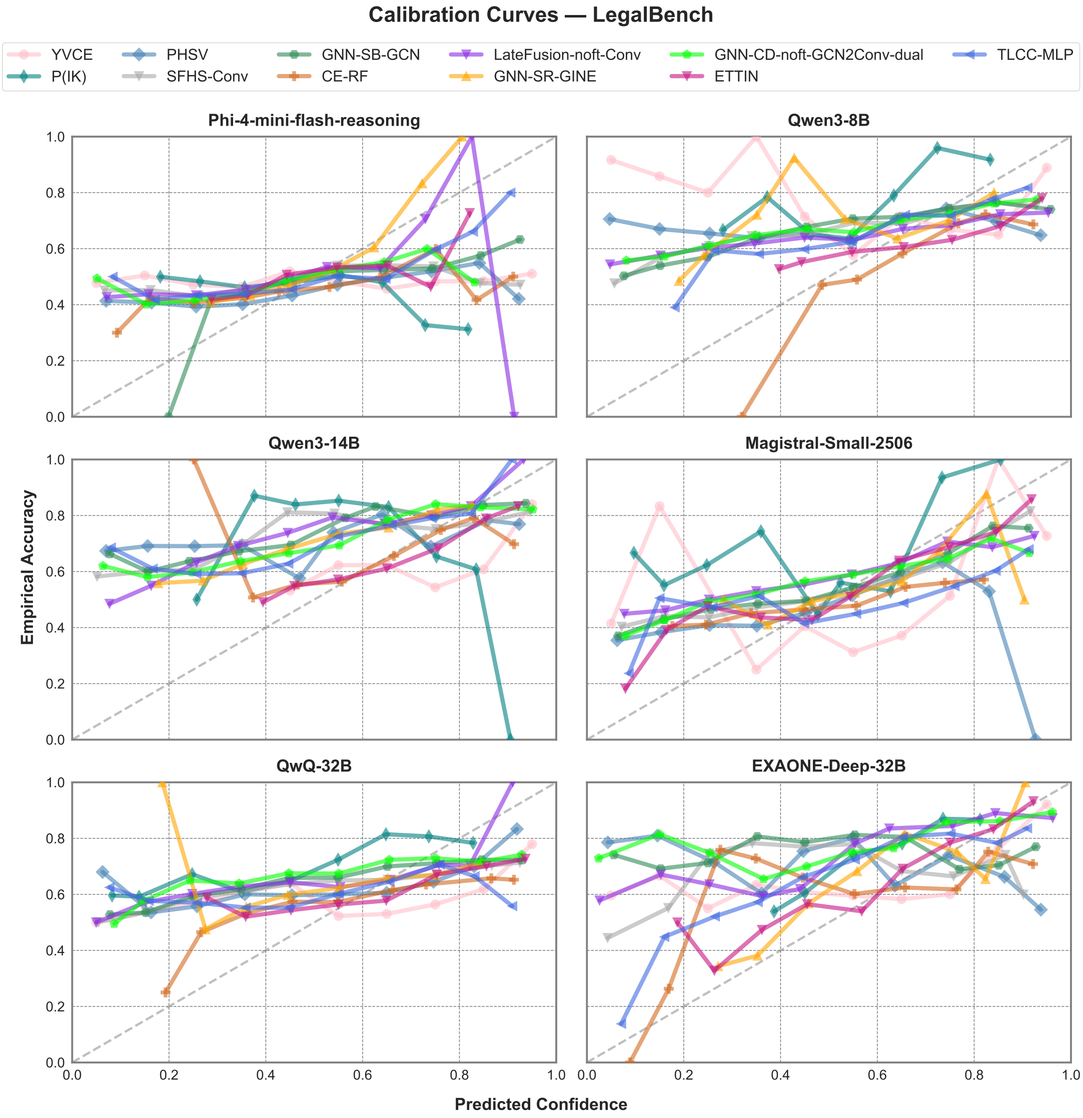

Figure 13: Reliability diagrams for the top-performing method variants on the LegalBench dataset, aggregated across all LRMs.

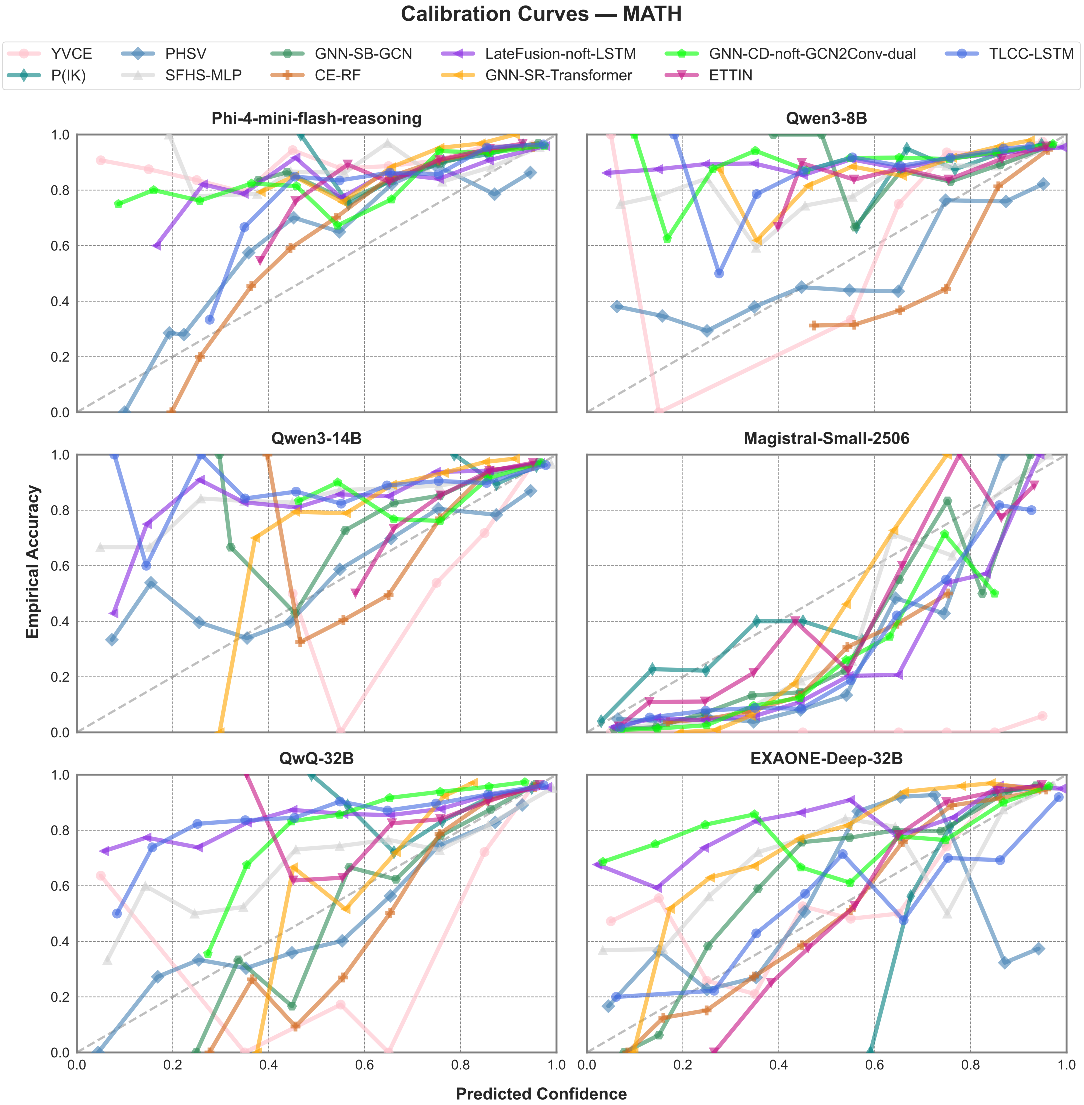

Figure 14: Reliability diagrams for the top-performing method variants on the MATH dataset, aggregated across all LRMs.

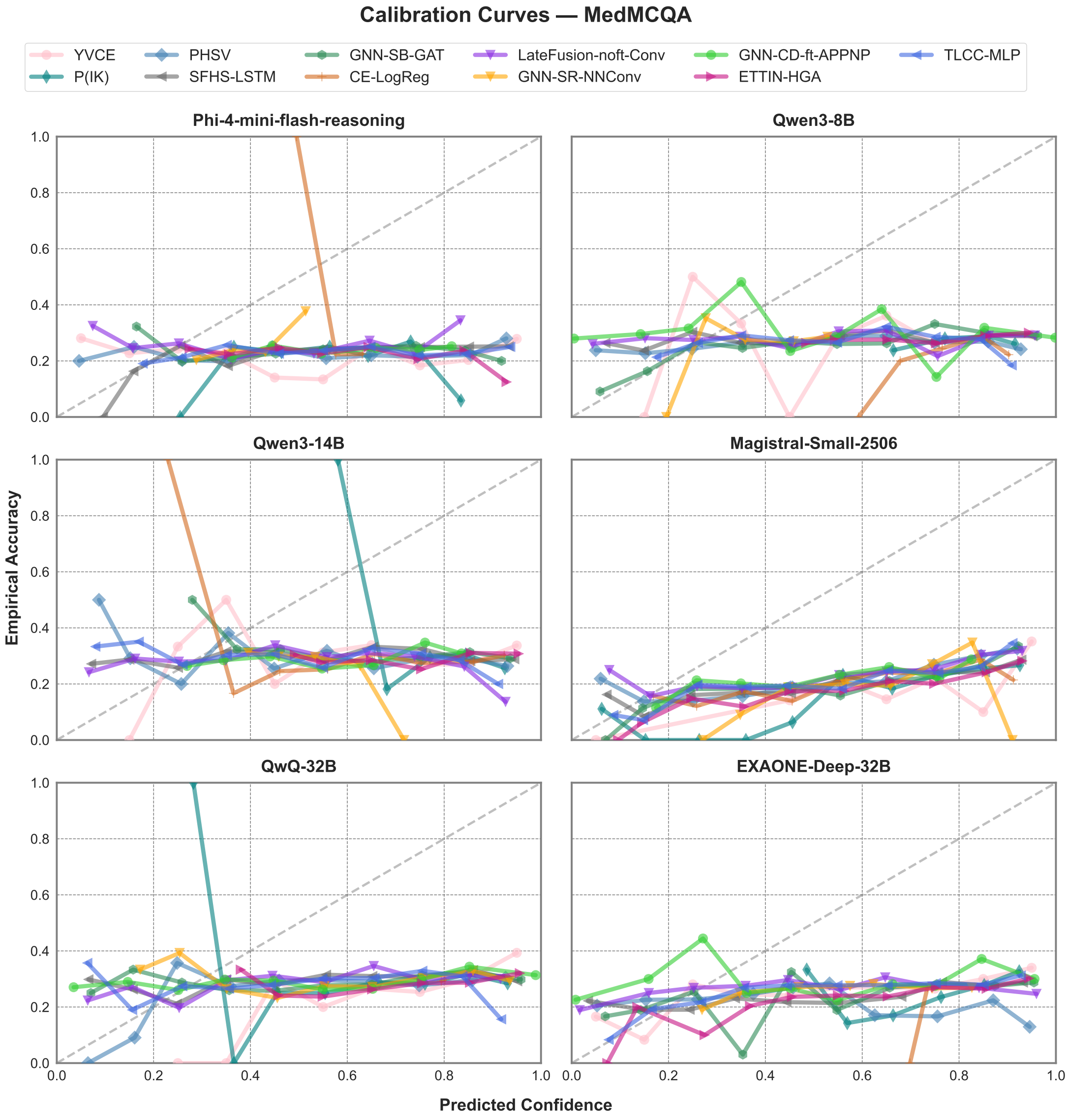

Figure 15: Reliability diagrams for the top-performing method variants on the MedMCQA dataset, aggregated across all LRMs.

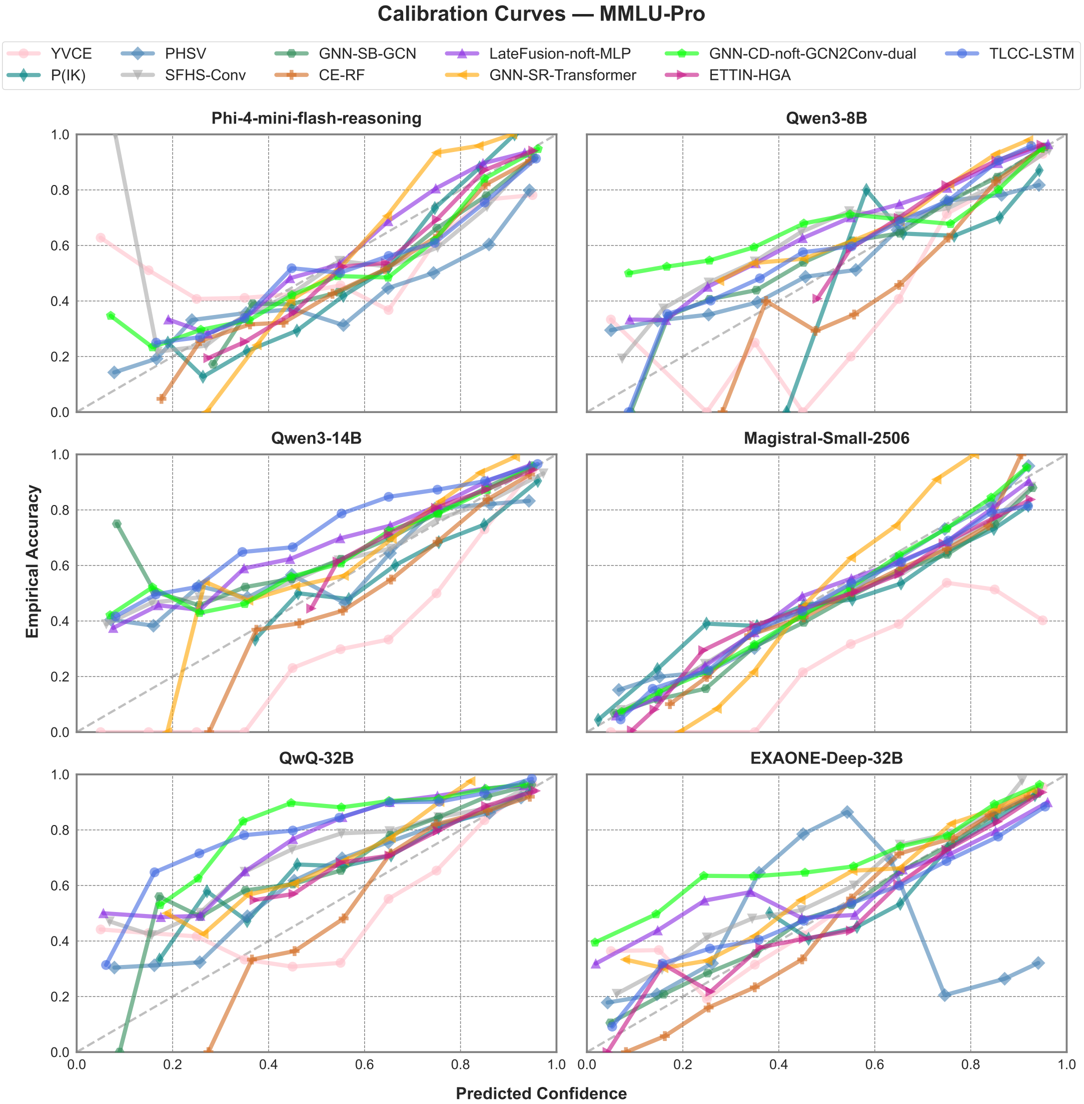

Figure 16: Reliability diagrams for the top-performing method variants on the MMLU-Pro dataset, aggregated across all LRMs.