From Causal Plausibility to Causal Reliability: Evaluating LLMs as Calibrated Direct Causal-Edge Classifiers

Amit Kumar[1] **Elnur Adl Zarabi**[1] **Suranjana Trivedy**[2] **Zhiqian Chen**[3]
Lei Zhang[4] Kaiqun Fu[5] Taoran Ji[1]
[1]Texas A&M University-Corpus Christi, USA
[2]BITS Pilani Goa, India
[3]Mississippi State University, USA
[4]Northern Illinois University, USA
[5]Texas Christian University, USA

Abstract

Large language models (LLMs) are increasingly used to provide prior causal knowledge for structural causal discovery, yet whether their direct-edge judgments and associated confidence can be trusted remains unclear. We systematically evaluate 12 instruction-tuned open-weight models across six benchmark causal graphs, five prompting strategies, and four confidence sources: verbalized, logit-based, cross-prompt agreement, and cross-model agreement. Our evaluation yields three key findings. (i) LLM-based causal judgments are strongly recall-dominant. Models tend to predict overly dense graphs with many false-positive edges, while prompting primarily shifts the precision–recall trade-off rather than consistently resolving overprediction. Improvements with model scale also diminish on the largest graphs and do not eliminate miscalibration. (ii) LLMs often capture causal relatedness without reliably identifying directness or orientation. Relative to the published reference graphs, models incorrectly classify 40.0% of indirect and 36.0% of reversed non-edges as direct causal edges, compared with 28.2% of other non-edges. Moreover, 80.8% and 84.6% of these false positives receive verbalized confidence of at least 80%, revealing substantial overconfidence in structurally incorrect predictions. (iii) Conventional confidence estimates are unreliable, whereas agreement provides a more promising signal. Logit-based confidence frequently collapses near 1.0 regardless of correctness, while cross-prompt and cross-model agreement achieve better mean calibration and discrimination, although their advantages are not statistically significant after Holm correction. A benchmark-familiarity audit additionally identifies potential familiarity in five model–dataset pairs, all involving AsiaM. Overall, our results suggest that LLMs are better viewed as sources of externally validated soft causal priors than as direct evidence of causal structure. Replication materials are available on GitHub.

1 Introduction

Causal discovery (CD) aims to infer directed causal relationships among variables, typically represented as a causal graph, to support explanation, intervention analysis, and scientific understanding (Pearl, 2009). For example, consider a simplified medical graph *Viral Exposure* → *Infection* → *Positive Test*. Within this graph, adjacent variables are connected by direct edges, whereas *Viral Exposure* influences *Positive Test* only indirectly through *Infection*. Traditional approaches, including constraint-based methods such as PC and FCI (Spirtes et al., 2000), score-based methods such as GES (Chickering, 2002) and NOTEARS (Zheng et al., 2018), and neural formulations such as GraN-DAG (Lachapelle et al., 2020), operate on observational data under assumptions about the data-generating process. However, observational data may be limited or insufficient to distinguish Markov-equivalent structures (Glymour et al., 2019), motivating the use of domain knowledge to guide or orient candidate graphs (Vashishtha et al., 2023). Such knowledge is often costly to elicit or unavailable in novel domains.

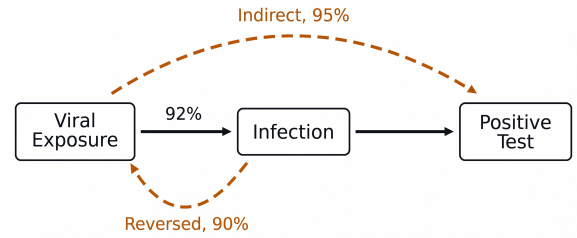

Figure 1: Illustration of unreliable pairwise direct-edge judgments. Solid arrows form the reference chain; dashed arrows show an indirect relation predicted as direct and a reversed relation. Confidence values are illustrative.

Large language models (LLMs) offer a complementary source of causal knowledge for CD, with recent work exploring direct causal reasoning (Kıcıman et al., 2024), prior-guided structure learning

(Vashishtha et al., 2023; Ban et al., 2023b; Kampani et al., 2024), and graph orientation and refinement (Long et al., 2023; Ban et al., 2023a). Although LLMs demonstrate non-trivial performance on pairwise causal tasks using textual metadata (Kıcıman et al., 2024), their predictions may depend on the frequency of causal relations in pre-training corpora and vary under contextual changes (Feng et al., 2025). Prompt style and response agreement can also affect confidence calibration (Xia et al., 2025). Existing work shows that pairwise LLM-based edge classification can yield poor graph recovery (Babakov et al., 2025), but does not systematically identify the structural sources of these errors or whether confidence reflects their correctness. Because such judgments are used as structural constraints, optimization priors, and causal-order information in CD pipelines (Ban et al., 2023b; Kampani et al., 2024; Vashishtha et al., 2023), their reliability remains important even when pairwise classification is insufficient as a standalone CD method.

This gap reflects a distinction between causal plausibility and causal reliability. For queried variables A and B, a reliable direct-edge judgment requires distinguishing $A \rightarrow B$ from reverse directionality, indirect influence, and no direct edge. A model may recognize causal relatedness while misidentifying directness or orientation, producing many spurious edges in sparse graphs. Confidence reliability is therefore as important as classification performance when LLM judgments inform graph learning. Figure 1 illustrates this reliability gap.

Accordingly, our aim is to characterize the reliability of edge judgments on which LLM-assisted CD pipelines may depend, rather than to propose a new CD method. Our contributions are as follows:

- We formulate pairwise direct causal-edge classification as a calibrated evaluation task, jointly assessing edge prediction quality, graph reconstruction, and confidence reliability under structural ambiguity and severe class imbalance.
- We develop an evaluation framework spanning six benchmark causal graphs, five prompting strategies, 12 instruction-tuned models at two scales, and four confidence sources: verbalized, logit-based, cross-prompt agreement, and cross-model agreement. We also evaluate potential benchmark familiarity across all 72 model–dataset combinations. We will release the code and evaluation artifacts to support reproducibility.
- We show that LLM edge judgments exhibit recall-dominant behavior and prompt sensitivity. Larger models achieve higher mean F1 on five of six datasets, but provide only marginal gains on the largest graphs and do not resolve miscalibration. Relative to the published reference graphs, false-positive rates are higher for indirect (40.0%) and reversed (36.0%) non-edges than for other non-edges (28.2%); over 80% of false positives on indirect and reversed relations receive verbalized confidence of at least 80%. A three-way task-formulation ablation further shows that these errors are not solely induced by evaluating the two directions through separate binary queries.
- We find that agreement-based confidence achieves better mean calibration and discrimination than verbalized and logit-based confidence, although the differences are not statistically significant after Holm correction and remain model- and prompt-dependent.

2 Evaluation Design

The framework in Fig. 2 treats LLMs as calibrated pairwise direct causal-edge classifiers. Across two model-scale groups, five prompting strategies, and four confidence sources, we ask whether models avoid edge overprediction, distinguish direct edges from reversed, indirect, and other non-edges, and assign confidence that reflects correctness. Pairwise predictions are compared with published reference graphs and aggregated into reconstructed graphs to assess edge classification and graph recovery. Structural error analysis, confidence thresholding, a three-way label-space ablation, and a benchmark-familiarity audit further test these reliability concerns.

2.1 Prompt Styles

Prompt formulation affects LLM performance (White et al., 2023; Chen et al., 2023; Qiao et al., 2023). We adapt pairwise causal prompting from prior work on causal reasoning and LLM-guided structure discovery (Kıcıman et al., 2024; Vashishtha et al., 2023) to direct-edge classification with explicit direct-causation instructions and verbal confidence elicitation (Lin et al., 2022).

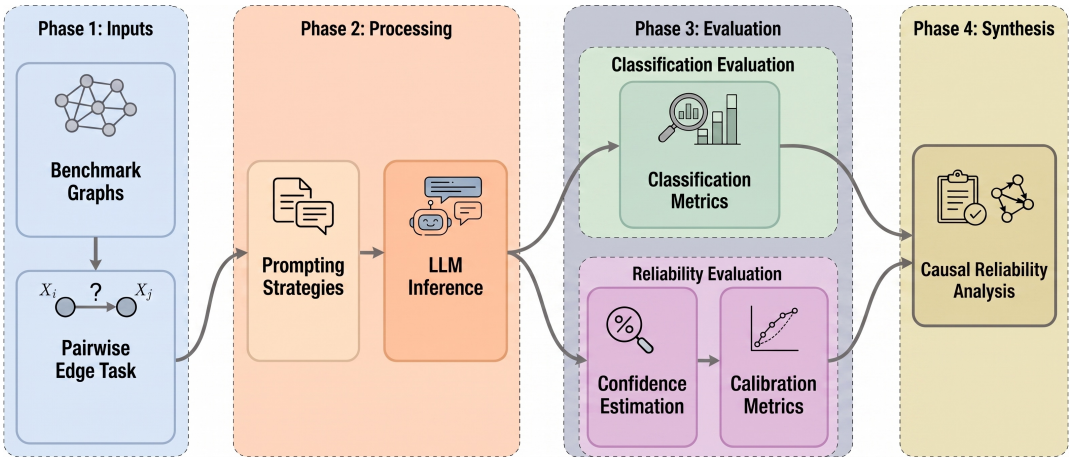

Figure 2: Evaluation pipeline for pairwise direct-edge classification, graph reconstruction, confidence calibration, and structural error analysis.

We evaluate five prompt styles that provide progressively richer guidance. **Name-only** supplies the dataset context and variable names, testing whether the model can infer a direct edge from names alone. **Metadata** additionally provides variable definitions to clarify their domain-specific meanings. **Chain-of-Thought (CoT)** requests a brief rationale before the prediction (Wei et al., 2022). **Few-shot (FS)** provides five labeled direct-edge examples that demonstrate the required task and output format (Brown et al., 2020). **Few-shot + CoT (FS+CoT)** combines these demonstrations with brief reasoning. All styles ask whether A directly causes B and require a binary "Yes"/"No" judgment with a self-reported confidence score from 0–100. Full templates appear in Appendix A.1.

2.2 Direct Edge Classification and Graph Reconstruction

The primary task is binary direct-edge classification. Given an ordered pair (A, B) and dataset context, the model predicts whether A directly causes B, with YES as the positive class and NO as the negative class. Responses follow a standardized answer–confidence format and are parsed deterministically, prioritizing labeled fields and then answer-first forms while ignoring prompt echoes. Missing fields trigger one retry; fewer than 0.2% of responses remain without a valid label and are excluded from classification.

For each model–prompt–dataset setting, all positively predicted ordered pairs are combined to form a reconstructed directed graph. We compare this reconstructed graph with the published reference graph using normalized Structural Hamming Distance (nSHD), defined in Section 3.3. Graph reconstruction is induced directly from pairwise predictions and does not enforce global structural constraints such as acyclicity.

2.3 Confidence Estimation Task

This task assesses whether confidence reflects correctness in binary edge classification. Let x denote an input query, $\mathcal{Y} = \{\text{YES}, \text{NO}\}$ the label set, and $\hat{y} \in \mathcal{Y}$ the predicted label. We consider four confidence sources, all normalized to $[0, 1]$.

Verbalized confidence. The model reports a confidence score $s_{\text{verb}} \in [0, 100]$ with its binary prediction (Lin et al., 2022; Tian et al., 2023). We define

$$c_{\text{verb}}(\hat{y}) = \frac{s_{\text{verb}}}{100}. \tag{1}$$

Logit-based confidence. We aggregate valid surface forms for each label and normalize their decoder logits over $\mathcal{Y}$ (Jiang et al., 2021; Tian et al., 2023):

$$p(y \mid x) = \frac{\exp(z_y)}{\sum_{y' \in \mathcal{Y}} \exp(z_{y'})}. \tag{2}$$

The predicted label and its confidence are

$$\hat{y} = \operatorname*{argmax}_{y \in \mathcal{Y}} p(y \mid x), \qquad c_{\text{logit}}(\hat{y}) = \max_{y \in \mathcal{Y}} p(y \mid x). \tag{3}$$

Cross-prompt agreement. For $K = 5$ prompt-specific predictions $y^{(1)}, \ldots, y^{(K)}$, let $\hat{y}$ be the majority label. Confidence is the fraction of prompts supporting it (Wang et al., 2023; Xiong et al., 2024):

$$c_{\text{prompt}}(\hat{y}) = \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}\left[y^{(k)} = \hat{y}\right]. \tag{4}$$

Cross-model agreement. For M predictions $y_1, \ldots, y_M$ under a fixed prompt, let $\hat{y}$ be the majority label. Computed separately within each model-scale group, confidence is (Xia et al., 2025; Zhang et al., 2024)

$$c_{\text{model}}(\hat{y}) = \frac{1}{M} \sum_{m=1}^{M} \mathbf{1}[y_m = \hat{y}]\,. \tag{5}$$

3 Experiment

3.1 Models

We evaluate 12 instruction-tuned open-weight models from five families: Qwen, Gemma, Llama, Mistral, and Phi (Yang et al., 2025, 2024; Google DeepMind, 2026; Grattafiori et al., 2024; Meta AI, 2024; Mistral AI, 2024; Abouelenin et al., 2025; Abdin et al., 2024), accessed through Hugging Face.[1] To examine scale effects, we group models by parameter count.

Small models (4–14B): Qwen3-4B-Instruct, Qwen3-8B-Instruct, Gemma-4-E4B-IT, Llama-3.1-8B-Instruct, Ministral-8B-Instruct-2410, Phi-4-Mini-Instruct, and Phi-4.

Large models (31–72B): Qwen3-32B-Instruct, Qwen2.5-72B-Instruct, Gemma-4-31B-IT, Llama-3.1-70B-Instruct, and Llama-3.3-70B-Instruct.

Figures abbreviate Gemma-4-31B-IT as Gemma-31B, Gemma-4-E4B-IT as Gemma-E4B, and Llama-3.1-8B-Instruct as Llama-8B.

3.2 Datasets

We use six graphs from CausalGraphBench (Babakov et al., 2025): AsiaM, River Status, COVID, Coal Gasifier, Hepar2, and Munin1. They span four size categories and medical, ecological, public-health, and industrial domains. Dataset statistics, benchmark IDs, and preprocessing details appear in Appendix A.3, Table A1.

This subset provides a controlled size- and domain-stratified evaluation. Pairwise classification scales as $n(n-1)$, and the six graphs already require approximately 2.5M queries, making full 35-graph evaluation prohibitive. We acknowledge in the Limitations that results may not capture full benchmark diversity.

Each dataset provides a domain description, variable definitions, and a published reference graph. The pairwise setting is highly imbalanced—only a small fraction of ordered pairs are true edges, worsening with graph size. This motivates our calibration focus, as aggregate accuracy can mask overconfidence on sparse edge classes.

3.3 Evaluation Metrics

We evaluate performance along two dimensions: direct-edge classification and confidence calibration. For classification, we report Precision, Recall, and F1 over ordered variable pairs (Davis and Goadrich, 2006). To compare structural recovery across datasets of different sizes, we also report normalized Structural Hamming Distance (nSHD), based on SHD (Tsamardinos et al., 2006). For dataset d, nSHD is defined as:

$$\text{nSHD}_d = \frac{\text{SHD}_d}{n_d(n_d - 1)}, \tag{6}$$

where n_d is the number of variables in dataset d.

For calibration, we report 10-bin squared-gap ECE, based on the binning framework of Guo et al. (2017), Brier Score (Brier, 1950), and AUROC (Ulmer et al., 2024). Lower ECE and Brier indicate better confidence–correctness alignment; higher AUROC indicates stronger discrimination between correct and incorrect predictions.

3.4 Experimental Settings

Few-shot prompts use five fixed demonstrations: three positive and two negative examples, held constant across models. We use temperature 1.0, repetition penalty 1.1, and maximum generation lengths of 128 tokens for direct-answer prompts and 384 tokens for reasoning prompts.

4 Results and Analysis

4.1 Primary Edge Classification

We analyze how prompt design, dataset characteristics, model scale, and potential benchmark familiarity affect direct-edge classification. We first examine the general precision–recall behavior and prompt sensitivity, then assess whether performance differences are associated with dataset characteristics and model scale. We finally evaluate potential benchmark familiarity and whether a three-way label space better distinguishes edge direction from the absence of a direct edge.

4.1.1 Recall-Dominant Behavior and Prompt Sensitivity

The results reveal a consistent recall-dominant pattern across both model groups. Recall exceeds

[1] Hugging Face models: https://huggingface.co/models.

precision in 68 of 72 model–dataset combinations (94.4%) after averaging across prompts, and in 305 of 360 model–dataset–prompt combinations (84.7%). Figure 3 summarizes this pattern at the prompt and model levels, macro-averaged across the remaining dimensions. Nearly all aggregates lie above the equal-precision–recall diagonal.

Although the magnitude of this behavior varies across settings, its direction remains stable. In Table 1, mean recall exceeds mean precision for every dataset and model-scale group, with recall standard deviations of 0.173–0.316 across model–prompt combinations. Models therefore predict edges for a large fraction of ordered pairs, producing dense graphs with many false positives. This pattern suggests that LLMs often treat causal relatedness as evidence of direct causation.

We further observe that this behavior persists across prompt styles. FS, CoT, and FS+CoT often shift models toward higher recall without comparable precision gains. Appendix Figure A1 also shows that the best-performing prompt varies across model–dataset combinations. Thus, prompt design affects the precision–recall tradeoff, but no prompt style consistently achieves a better balance across models and datasets.

4.1.2 Effect of Graph Size and Semantic Accessibility

Having established that prompting does not eliminate recall-dominant behavior, we next examine why the balance between precision and recall varies across datasets. We use F1 to summarize this balance and compare edge-classification performance across graphs. Table 1 shows that F1 generally declines as graph size and class imbalance increase: mean F1 falls from 0.281 and 0.299 on RIVER STATUS to 0.031 for both model groups on MUNIN1. Although the SDs indicate variation across model–prompt combinations, the largest graphs consistently yield low F1. Appendix Figure A3 and Table A5 report the complete results.

Graph size alone does not fully explain these differences. We therefore examine semantic accessibility using description length, code-like variable names, and acronym density (Appendix Table A2). Considering all 12 models, even the lowest model-level mean F1 on RIVER STATUS (0.220) exceeds the highest achieved on COAL GASIFIER (0.121) and MUNIN1 (0.050), with each model averaged across five prompts. This contrast is consistent with the natural-language descriptions in RIVER STATUS and the more technical metadata in the latter datasets. MUNIN1 is particularly challenging, combining 186 variables with 100% code-like names and 27.4% acronym-containing descriptions.

However, metadata complexity alone is insufficient. COVID achieves comparatively higher F1 despite having 100% code-like names and 45% acronym-containing descriptions, as its metadata refers to broadly recognizable public-health concepts. Overall, these exploratory indicators suggest that performance is jointly associated with graph size, class imbalance, and the linguistic accessibility of domain concepts; they do not establish causal effects. Overall, these exploratory indicators suggest that performance is jointly associated with graph size, class imbalance, and the linguistic accessibility of domain concepts; they do not establish causal effects. This variation also motivates assessing whether greater model scale mitigates these constraints.

4.1.3 Effect of Model Scale

Model scale improves mean F1 on five of six datasets, but the gains are uneven. Figure 4 compares small and large LLMs across datasets. The largest improvements occur on ASIAM (0.478 to 0.680) and COVID (0.280 to 0.403). In contrast, improvements are modest on RIVER STATUS (+0.018), COAL GASIFIER (+0.017), and HEPAR2 (+0.037), while performance is effectively unchanged on MUNIN1.

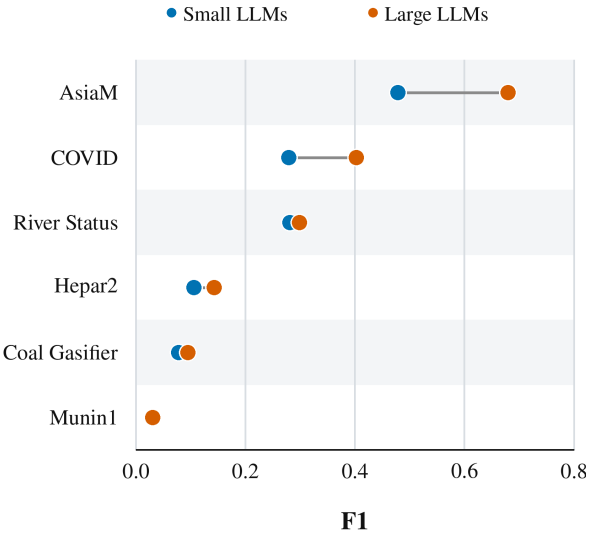

Figure 4: Dataset-level mean F1 for small and large LLMs, averaged across models and prompts within each group. Lines connect group means; rightward shifts favor large LLMs.

Scale also does not eliminate recall-dominant behavior. Figure 3 shows that nearly all models remain above the equal-precision–recall diagonal,

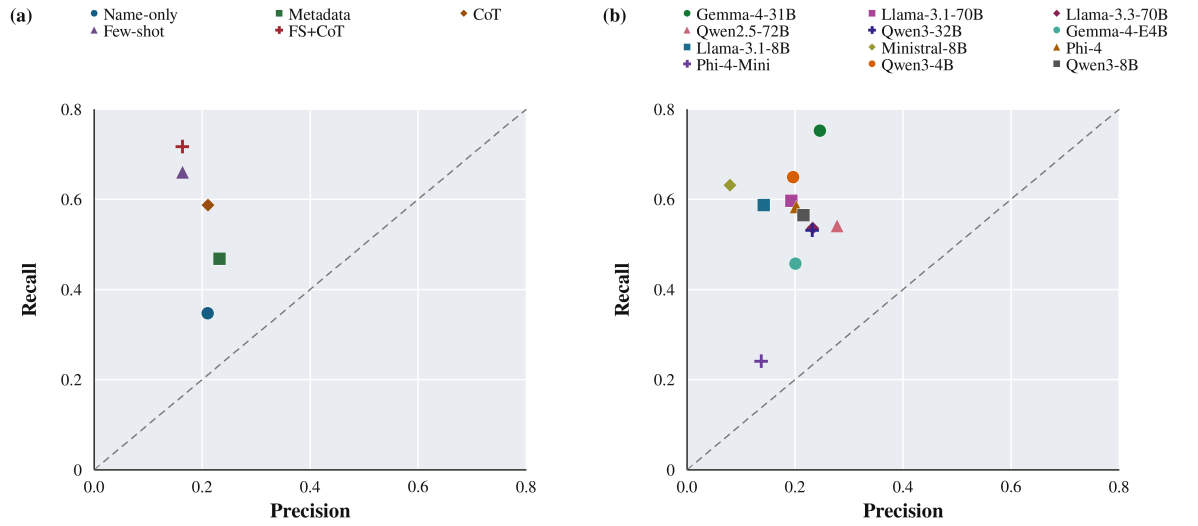

Figure 3: Precision–recall behavior by (a) prompt, averaged across models and datasets, and (b) model, averaged across prompts and datasets. The dashed diagonal denotes equal precision and recall; points above it are recall-dominant.

Table 1: Dataset-level edge-classification and graph-reconstruction performance. Values are mean ± SD across model–prompt combinations within each scale group (S: 7 models; L: 5 models). Higher precision, recall, and F1 and lower nSHD indicate better performance.

Dataset	Precision ↑		Recall ↑		F1 ↑		nSHD ↓	
	S	L	S	L	S	L	S	L
AsiaM	0.415 ± 0.175	$\mathbf{0.608 \pm 0.139}$	0.707 ± 0.264	$\mathbf{0.830 \pm 0.210}$	0.478 ± 0.149	$\mathbf{0.680 \pm 0.126}$	0.281 ± 0.145	$\mathbf{0.141 \pm 0.062}$
River Status	0.198 ± 0.048	$\mathbf{0.216 \pm 0.062}$	0.592 ± 0.213	$\mathbf{0.597 \pm 0.292}$	0.281 ± 0.050	$\mathbf{0.299 \pm 0.107}$	0.357 ± 0.139	$\mathbf{0.292 \pm 0.090}$
COVID	0.261 ± 0.181	$\mathbf{0.412 \pm 0.171}$	0.443 ± 0.221	$\mathbf{0.545 \pm 0.252}$	0.280 ± 0.144	$\mathbf{0.403 \pm 0.080}$	0.191 ± 0.158	$\mathbf{0.104 \pm 0.039}$
Coal Gasifier	0.049 ± 0.029	$\mathbf{0.069 \pm 0.057}$	0.428 ± 0.247	$\mathbf{0.505 \pm 0.316}$	0.078 ± 0.045	$\mathbf{0.095 \pm 0.053}$	0.288 ± 0.204	$\mathbf{0.225 \pm 0.159}$
Hepar2	0.062 ± 0.021	$\mathbf{0.086 \pm 0.024}$	0.546 ± 0.173	$\mathbf{0.593 \pm 0.218}$	0.106 ± 0.030	$\mathbf{0.143 \pm 0.029}$	0.265 ± 0.144	$\mathbf{0.191 \pm 0.097}$
Munin1	$\mathbf{0.018 \pm 0.016}$	$\mathbf{0.018 \pm 0.008}$	0.468 ± 0.259	$\mathbf{0.480 \pm 0.295}$	$\mathbf{0.031 \pm 0.019}$	$\mathbf{0.031 \pm 0.014}$	0.326 ± 0.241	$\mathbf{0.269 \pm 0.219}$

while Table 1 shows recall exceeding precision for both scale groups on every dataset. Thus, larger models can improve edge recovery, particularly on semantically accessible graphs, but provide limited gains on the most difficult datasets and do not resolve systematic edge overprediction. The especially large improvement on ASIAM warrants qualification through the benchmark-familiarity audit presented next.

4.1.4 Potential Benchmark Familiarity Audit

The benchmark-familiarity audit flags five model–dataset pairs as high risk, all involving ASIAM; no model is flagged on the other five datasets. Recall-dominant overprediction and low performance on larger graphs persist across these unflagged datasets, indicating that the principal classification patterns are not driven by ASIAM alone.

Following Babakov et al. (2025), we apply a two-stage node- and structure-recall test to all 72 model–dataset pairs. Each model first reproduces a graph's nodes using only its source reference and domain description. Generated and reference nodes are aligned through semantic matching using Mixtral-8x7B-Instruct, an independent judge not included among the evaluated models. A pair is flagged when node-count deviation is below 15% and node recall exceeds 0.85; only flagged pairs proceed to structure recall.

Table 2: Structure recall for model–dataset pairs passing the benchmark-familiarity screen, all on AsiaM (7 nodes, 8 edges). Nodes reports generated/true counts, Dev. is node-count deviation, and SHD is the number of edge edits.

Model	Nodes	Recall	Dev.	Edge F1	SHD↓
Phi-4	8/7	0.857	14.3%	0.500	6
Qwen3-32B	7/7	0.857	0.0%	0.714	4
Qwen2.5-72B	8/7	1.000	14.3%	0.857	2
Gemma-4-31B	7/7	1.000	0.0%	0.800	3
Llama-3.3-70B	8/7	1.000	14.3%	0.941	1

The flagged pairs show varying degrees of structure recovery (Table 2), indicating potential familiarity rather than uniform graph memorization. Notably, Gemma-4-31B-IT is flagged, suggesting that familiarity may partly contribute to the strong model-scale gain observed on ASIAM. We therefore interpret results on this dataset cautiously, while the principal classification conclusions remain supported by the five unflagged datasets. Full results appear in Appendix Table A15.

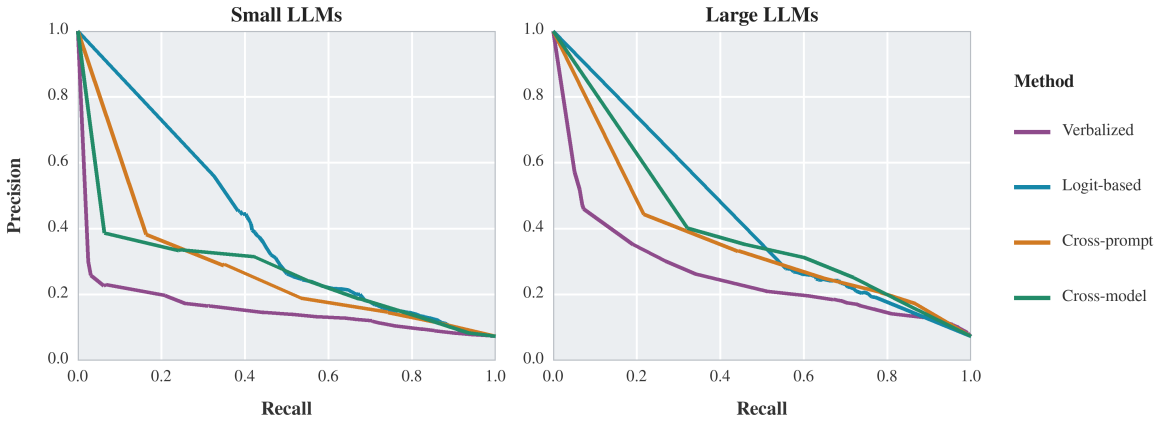

Figure 5: Edge precision–recall curves obtained by thresholding positive-edge confidence for small and large model groups, macro-averaged across six datasets.

4.2 Confidence Calibration

Having established that LLMs frequently overpredict causal edges, we examine whether confidence can identify which predictions are reliable. This is important when LLM judgments are used as causal priors: informative confidence can support edge weighting, thresholding, and graph sparsification, whereas overconfident errors can reinforce incorrect structure. We compare four confidence sources in terms of calibration and discrimination, examine their sensitivity to prompt style, and test whether post-hoc temperature scaling improves logit-based calibration.

4.2.1 Agreement Supports Edge Ranking but Does Not Guarantee Instance-Level Calibration

Agreement-based confidence provides the strongest aggregate reliability signal, although its advantage varies across datasets and model groups. Table 3 shows that cross-model agreement achieves the highest mean AUROC for small models (0.692 ± 0.043), while cross-prompt agreement performs best for large models (0.772 ± 0.047). Both methods also achieve lower mean ECE and Brier scores than verbalized and logit-based confidence. However, pairwise two-sided Wilcoxon tests over the six datasets show no significant differences after Holm adjustment (minimum adjusted $p = 0.1875$; Appendix Table A11). Agreement therefore performs better descriptively, but its statistical superiority across datasets is not established.

Its practical value is clearest when confidence is used to rank or filter candidate edges. Correctness AUROC evaluates whether confidence separates correct from incorrect predictions, whereas edge AUPRC treats the presence of a reference edge as the positive class. We assign positive-edge confidence c to YES predictions and $1 - c$ to NO predictions; logit-based confidence uses the normalized YES probability. Figure 5 shows that cross-model agreement achieves the highest edge AUPRC for both small (0.228) and large (0.269) LLMs. For large models, a cross-prompt threshold of 0.8 reduces graph density by 44.7% and increases precision from 0.250 to 0.331, while F1 remains nearly unchanged (0.314 to 0.310). Agreement can therefore support controllable graph sparsification, although increasing precision generally reduces recall.

Table 3: Calibration by model group and confidence source, reported as mean ± SD across six datasets. ECE denotes squared-gap ECE. Bold and underlining indicate the best and second-best results within each group.

Group	Method	ECE↓	Brier↓	AUROC↑	Acc.↑
Small	Cross-model	**0.019 ±0.013**	**0.149 ±0.043**	**0.692 ±0.043**	0.792 ±0.071
	Cross-prompt	<u>0.038 ±0.012</u>	<u>0.176 ±0.043</u>	<u>0.661 ±0.047</u>	0.755 ±0.060
	Logit-based	0.226 ±0.054	0.354 ±0.046	0.611 ±0.037	0.549 ±0.064
	Verbalized	0.229 ±0.034	0.364 ±0.032	0.358 ±0.059	0.709 ±0.057
Large	Cross-prompt	**0.016 ±0.012**	**0.127 ±0.050**	**0.772 ±0.047**	0.820 ±0.070
	Cross-model	<u>0.024 ±0.018</u>	<u>0.139 ±0.056</u>	<u>0.704 ±0.055</u>	0.822 ±0.070
	Verbalized	0.205 ±0.061	0.320 ±0.063	0.448 ±0.099	0.794 ±0.072
	Logit-based	0.217 ±0.053	0.346 ±0.057	0.582 ±0.021	0.620 ±0.061

These aggregate gains do not guarantee informative instance-level uncertainty. When confidence is concentrated near aggregate accuracy, ECE can be low without separating correct from incorrect

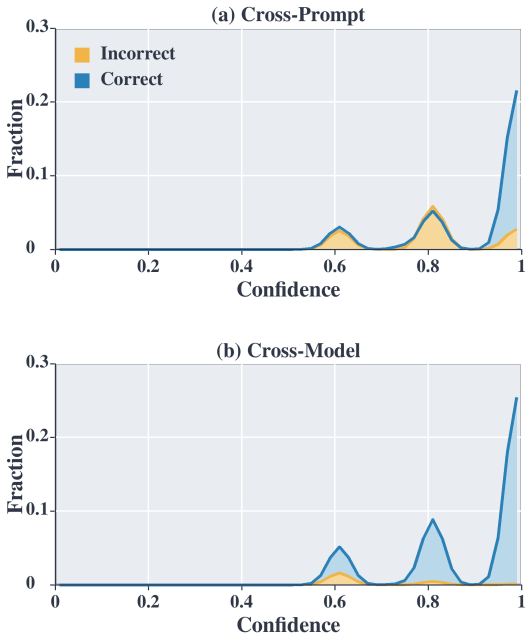

Figure 6: Agreement-confidence distributions pooled across six datasets: (a) cross-prompt agreement for Gemma-4-31B-IT and (b) cross-model agreement under Metadata prompting. Correct predictions concentrate more strongly near full agreement.

predictions. Appendix Figures A4–A7 show that verbalized and logit-based confidence can cluster within narrow high-confidence ranges, partly reflecting dominant non-edge behavior. Logit-based confidence is particularly concentrated near 1.0 for both correct and incorrect predictions (Appendix Figure A10).

Agreement is less affected because it measures stability across prompts or models. Figure 6 shows that correct predictions concentrate more strongly near full agreement than incorrect predictions. Nevertheless, agreement should be interpreted as an aggregate stability signal rather than a calibrated probability of correctness. Because prompt formulation is one such inference condition, the following subsection examines how prompt style affects calibration.

4.2.2 Effects of Prompt Style on Calibration

Prompt formulation affects both causal-edge predictions and how confidence is expressed, but no prompt style consistently provides the best calibration across models and datasets. Appendix Tables A7 and A9 show substantial variation across prompt styles and confidence sources. For example, CoT reduces verbalized ECE for Gemma-4-E4B-IT on COVID, HEPAR2, and MUNIN1, but the corresponding AUROC does not consistently improve. Prompting can therefore shift confidence closer to aggregate accuracy without making it more informative for distinguishing correct from incorrect predictions.

This instability provides the rationale for cross-prompt agreement, which measures whether an edge judgment remains stable across all five prompt formulations rather than relying on confidence from one prompt. Its reliability nevertheless remains model- and dataset-dependent, so prompt agreement should not be treated as uniformly calibrated. The instability of direct confidence across prompt settings also raises whether its calibration can be improved after inference, which we examine next through post-hoc temperature scaling.

4.2.3 Post-hoc Temperature Scaling

Post-hoc temperature scaling changes logit-based calibration in a model-specific way but cannot correct errors in the underlying causal-edge decision. We evaluate this without rerunning the LLMs by rescaling the saved Yes/No logits and recomputing normalized binary confidence. The predicted label remains unchanged because positive temperature scaling preserves the ordering of the Yes and No logits. This analysis therefore isolates changes in confidence calibration from changes in edge classification.

As shown in Appendix Fig. A11, Gemma-31B benefits from higher-temperature smoothing: its ECE and Brier score decrease as temperature increases, suggesting that the original Yes/No logits are overconfident. Qwen2.5-72B changes less across temperatures and maintains lower ECE, indicating comparatively stable logit-based confidence. AUROC remains nearly unchanged for both models because temperature scaling preserves the ranking induced by the original logit differences.

These results show that temperature scaling can adjust confidence sharpness but cannot correct cases where the model assigns the higher logit to the wrong causal label. To identify the structural distinctions underlying these persistent errors, we next examine whether overconfident false positives concentrate on indirect and reversed relations.

4.3 Overconfident False Positives on Structurally Difficult Non-Edges

The analysis reveals a clear structural pattern: false positives occur more frequently on indirect and reversed relations and often carry high verbal confidence. We analyze valid model-query instances where the reference graph contains no direct edge

$A \to B$. Non-edges are divided into three mutually exclusive categories: *reversed direct*, where $B \to A$ exists; *indirect*, where B is reachable from A through a directed path of length at least two; and *other* non-edges. This separation distinguishes orientation errors from cases where a mediated causal relation is incorrectly classified as a direct edge.

Table 4: False-positive behavior by non-edge type. FP Rate is the fraction of valid queries producing false positives; High-Conf. FP Rate is the fraction producing false positives with verbal confidence at least 80.

Non-edge Type	Valid Pairs	FP Count	FP Rate	High-Conf. FP Rate
Reversed direct	29,559	10,644	36.0%	30.4%
Indirect	161,520	64,568	**40.0%**	**32.3%**
Other	2,256,303	637,205	28.2%	22.1%

Table 4 shows that models predict a direct edge for 40.0% of indirect non-edges and 36.0% of reversed direct non-edges, compared with 28.2% of other non-edges. High-confidence false positives follow the same pattern: 32.3% of indirect and 30.4% of reversed direct queries produce false positives with verbal confidence of at least 80, compared with 22.1% of other non-edges. Overprediction is therefore especially pronounced when the queried variables are causally related in the reference graph but not through the proposed direct edge. The indirect-versus-other pattern holds across all six datasets, while reversed direct non-edges have a higher false-positive rate than other non-edges on five of six datasets.

Figure 7 further shows that these errors remain concentrated at high confidence. Among false positives with parseable verbal confidence, 84.6% of reversed direct, 80.8% of indirect, and 78.4% of other false positives fall in the 80–100 confidence range (Appendix Table A12). Appendix Table A13 reports the corresponding worst-case model-level behavior.

These findings should be interpreted relative to the published reference graphs. Because CausalGraphBench provides a single reference structure for each dataset, some absent edges may reflect graph-construction choices, and our analysis cannot establish that every false positive is causally invalid. Nevertheless, the higher false-positive rates on reversed and indirect relations show that the errors are structurally concentrated rather than uniform. Reversed predictions contradict the encoded orientation, while indirect predictions collapse a represented mediated path into a direct edge. Thus, relative to the benchmark specification, LLMs struggle to preserve causal direction and distinguish direct from mediated influence, although some predicted relations absent from the reference graph may remain causally plausible.

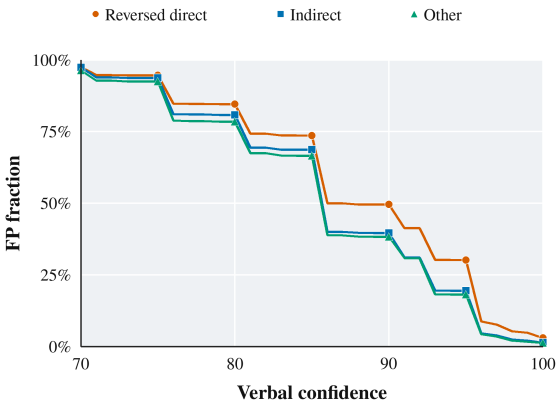

Figure 7: Verbal-confidence survival among false positives with parseable scores. Each curve shows the fraction at or above a given threshold; reversed-direct errors remain most concentrated at high confidence.

Prompting strategy is also associated with this behavior. Appendix Table A12 shows that FS+CoT produces the highest high-confidence false-positive rate across all three categories, indicating that reasoning-oriented demonstrations are associated with a greater tendency to infer direct edges between causally related variables. Overall, LLMs appear to capture coarse causal relatedness while frequently misstating directness or orientation with high confidence. This finding motivates the following ablation, which tests whether jointly representing both edge directions and the no-edge class reduces these errors.

4.4 Three-Way Label-Space Ablation

Three-way labeling does not consistently improve graph recovery, indicating that independent binary queries are not the sole cause of edge overprediction. In the original formulation, each unordered variable pair is evaluated through two separate questions: whether $A \to B$ exists and whether $B \to A$ exists. Because these decisions are made independently, a model can predict both directions or fail to compare a directed edge directly against its reverse.

The ablation replaces these two queries with one three-way decision over $A \to B$, $B \to A$, and no direct edge. This formulation forces the two directions to compete and prevents reciprocal predictions for the same pair. It also makes directness

explicit: if A influences B only through an intermediate variable, the correct label is no direct edge. The ablation therefore tests whether jointly deciding edge existence and orientation reduces the indirect and reversed errors identified in Section 4.3.

We evaluate Qwen3-4B-Instruct, Phi-4, Gemma-4-31B-IT, and Llama-3.3-70B-Instruct on RIVER STATUS, COVID, HEPAR2, and MUNIN1 using Metadata prompting. The selected datasets represent the small, medium, large, and very-large categories through River Status, COVID, Hepar2, and Munin1, respectively, while the models represent four families across two scale groups. Binary results are taken from the corresponding Metadata runs in the main experiment. Three-way predictions are converted into directed graphs by adding the selected $A \to B$ or $B \to A$ edge and adding no edge for the third label. Both formulations are then evaluated using the same edge precision, recall, F1, and nSHD metrics, macro-averaged across the four datasets. All three-way runs cover every unordered pair and produce validly parsed labels. The complete prompt appears in Appendix A.1.6.

Table 5: Binary versus three-way direct-edge classification under Metadata prompting, macro-averaged across four datasets. Bold indicates the better formulation for each model and metric.

Model	Formulation	P↑	R↑	F1↑	nSHD↓
Qwen3-4B	Binary	**0.229**	0.570	**0.261**	**0.200**
	Three-way	0.124	**0.678**	0.194	0.276
Phi-4	Binary	0.222	0.416	0.234	**0.132**
	Three-way	**0.224**	**0.627**	**0.303**	0.145
Gemma-4-31B	Binary	0.177	**0.816**	0.270	0.242
	Three-way	**0.187**	0.714	**0.275**	**0.205**
Llama-3.3-70B	Binary	**0.162**	0.076	0.098	**0.068**
	Three-way	0.152	**0.795**	**0.236**	0.294

Table 5 shows that three-way labeling improves F1 for Phi-4, Gemma-4-31B-IT, and Llama-3.3-70B-Instruct, but decreases it for Qwen3-4B-Instruct. These F1 changes do not translate into consistent structural improvement. Only Gemma-4-31B-IT improves both F1 and nSHD, and its F1 increase is marginal. For Llama-3.3-70B-Instruct, recall increases from 0.076 to 0.795, but nSHD worsens from 0.068 to 0.294. The model therefore predicts substantially more true edges while adding enough false edges to produce a less accurate reconstructed graph.

The explicit no-edge label also does not eliminate directness errors. We define the three-way non-edge false-positive rate as the fraction of reference no-edge pairs assigned either directed label. On MUNIN1, this rate ranges from 13.2% for Phi-4 to 91.0% for Llama-3.3-70B-Instruct, reaching 62.9% for Qwen3-4B-Instruct and 45.2% for Gemma-4-31B-IT. Direction-specific recall is also asymmetric under the fixed unordered-pair ordering, with Qwen3-4B-Instruct and Llama-3.3-70B-Instruct recovering no $B \to A$ edges on MUNIN1. Because the directional classes are imbalanced, this asymmetry indicates sensitivity to pair ordering rather than establishing that one causal direction is intrinsically harder.

Overall, forcing edge existence and orientation into a single mutually exclusive decision does not reliably reduce overprediction or improve orientation recovery. The errors observed in the binary experiment are therefore not solely artifacts of asking about each direction separately; models continue to select directed relations when the reference graph specifies no direct edge. Full dataset-level results are reported in Appendix Table A14.

5 Conclusion

We evaluated whether LLMs provide reliable direct causal-edge judgments and confidence estimates across 12 open-weight models, six benchmark graphs, and five prompting strategies. Models were recall-dominant, producing dense graphs, while prompting shifted the precision–recall tradeoff and gains from model scale diminished on larger graphs. False positives concentrated on indirect and reversed relations and were frequently assigned high confidence, indicating that LLMs often capture causal relatedness without preserving directness or orientation. A mutually exclusive three-way formulation did not consistently improve graph recovery, showing that these errors are not solely artifacts of separate binary queries.

Verbalized and logit-based confidence were unreliable, whereas agreement-based methods achieved better mean calibration and discrimination without establishing statistical superiority across datasets. These patterns persisted on the five graphs not flagged for potential benchmark familiarity. Overall, LLM judgments are better treated as externally validated soft priors than as direct evidence of causal structure.

Future Work. Future work should develop graph-aware uncertainty methods that incorporate sparsity, acyclicity, and directional consistency, and evaluate calibrated LLM judgments as soft priors within data-driven CD pipelines. Experimenting

with Small Language Models (SLMs) can further assess whether lightweight models offer competitive edge-level reliability at lower computational cost. Robustness should also be tested across paraphrased prompts and broader graph collections.

Limitations

Our evaluation covers six of the 35 CausalGraphBench graphs. Although stratified by size and domain, they may not represent the benchmark's full diversity or noisier real-world settings. We treat the published structures as fixed reference graphs, although they may reflect expert choices about causal granularity; alternative annotations and inter-annotator agreement estimates are unavailable. Consequently, not every prediction labeled as a false positive can be established as causally invalid.

Our language-only, pairwise formulation uses neither observational nor interventional data and does not enforce global constraints such as acyclicity. The semantic-accessibility analysis relies on heuristic metadata indicators and is exploratory rather than causal. We compare prompting strategies but do not test robustness to paraphrased wording within each strategy. Results may also depend on the selected open-weight models and decoding settings. Finally, the benchmark-familiarity audit flags five model–dataset pairs involving AsiaM, but its single detection protocol and judge model cannot establish memorization or rule out familiarity with unflagged datasets.

References

Marah Abdin and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Abdelrahman Abouelenin and 1 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. *arXiv preprint arXiv:2503.01743*.

Steen Andreassen, Finn V. Jensen, Stig Kjær Andersen, B. Falck, U. Kjærulff, M. Woldbye, A. R. Sørensen, A. Rosenfalck, and F. Jensen. 1989. MUNIN: An expert EMG assistant. In J. E. Desmedt, editor, *Computer-Aided Electromyography and Expert Systems*, pages 255–277. Pergamon Press.

Nikolay Babakov, Ehud Reiter, and Alberto Bugarín-Diz. 2025. CausalGraphBench: A benchmark for evaluating language models capabilities of causal graph discovery. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 240–258, Vienna, Austria. Association for Computational Linguistics.

Taiyu Ban, Lyuzhou Chen, Derui Lyu, Xiangyu Wang, and Huanhuan Chen. 2023a. Causal structure learning supervised by large language model. *arXiv preprint arXiv:2311.11689*.

Taiyu Ban, Lyuzhou Chen, Xiangyu Wang, and Huanhuan Chen. 2023b. From query tools to causal architects: Harnessing large language models for advanced causal discovery from data. *arXiv preprint arXiv:2306.16902*.

Glenn W. Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2023. Unleashing the potential of prompt engineering in large language models: A comprehensive review. *arXiv preprint arXiv:2310.14735*.

David Maxwell Chickering. 2002. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554.

Jesse Davis and Mark Goadrich. 2006. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 233–240. ACM.

Tao Feng, Lizhen Qu, Niket Tandon, Zhuang Li, Xiaoxi Kang, and Gholamreza Haffari. 2025. On the reliability of large language models for causal discovery. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9565–9590, Vienna, Austria. Association for Computational Linguistics.

Clark Glymour, Kun Zhang, and Peter Spirtes. 2019. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10:524.

Google DeepMind. 2026. Gemma 4 model card. `https://ai.google.dev/gemma/docs/core/model_card_4`. Accessed: 2026.

Aaron Grattafiori and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

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.

Shiv Kampani, David Hidary, Constantijn van der Poel, Martin Ganahl, and Brenda Miao. 2024. Llm-initialized differentiable causal discovery. *arXiv preprint arXiv:2406.06406*.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. 2024. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. 2020. Gradient-Based Neural DAG Learning. In *Proceedings of the Eighth International Conference on Learning Representations*, ICLR 2020, Addis Ababa, Ethiopia. OpenReview.net.

Steffen L. Lauritzen and David J. Spiegelhalter. 1988. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):157–224.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

Yunpeng Liu, Shen Wang, Qian Liu, Dongpeng Liu, Yang Yang, Yong Dan, and Wei Wu. 2022. Failure risk assessment of coal gasifier based on the integration of bayesian network and trapezoidal intuitionistic fuzzy number-based similarity aggregation method (tpifn-sam). *Processes*, 10(9):1863.

Stephanie Long, Alexandre Piché, Valentina Zantedeschi, Tibor Schuster, and Alexandre Drouin. 2023. Causal discovery with language models as imperfect experts. In *ICML 2023 Workshop on Structured Probabilistic Inference and Generative Modeling*.

Helen J. Mayfield, Colleen L. Lau, Jane E. Sinclair, Samuel J. Brown, Andrew Baird, John Litt, Aapeli Vuorinen, Kirsty R. Short, Michael Waller, and Kerrie Mengersen. 2022. Designing an evidence-based bayesian network for estimating the risk versus benefits of astrazeneca covid-19 vaccine. *Vaccine*, 40(22):3072–3084.

Meta AI. 2024. Llama-3.3-70B-Instruct model card. `https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct`. Accessed: 2024.

Mistral AI. 2024. Ministral-8B-Instruct-2410. `https://huggingface.co/mistralai/Ministral-8B-Instruct-2410`. Accessed: 2024.

Eugenio Molina-Navarro, Pedro Segurado, Paulo Branco, Carina Almeida, and Hans E. Andersen. 2020. Predicting the ecological status of rivers and streams under different climatic and socioeconomic scenarios using bayesian belief networks. *Limnologica*, 80:125742.

Agnieszka Onisko. 2003. *Probabilistic Causal Models in Medicine: Application to Diagnosis of Liver Disorders*. Ph.D. thesis, Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Sciences, Warsaw, Poland.

Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference*, 2 edition. Cambridge University Press.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, Toronto, Canada. Association for Computational Linguistics.

Peter Spirtes, Clark Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search*, 2nd edition. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Ioannis Tsamardinos, Laura E. Brown, and Constantin F. Aliferis. 2006. The Max-Min Hill-Climbing Bayesian Network structure learning algorithm. *Machine Learning*, 65(1):31–78.

Dennis Ulmer, Martin Gubri, Hwaran Lee, Sangdoo Yun, and Seong Oh. 2024. Calibrating Large Language Models using their generations only. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15440–15459, Bangkok, Thailand. Association for Computational Linguistics.

Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth N. Balasubramanian, and Amit Sharma. 2023. Causal inference using llm-guided discovery. *arXiv preprint arXiv:2310.15117*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le,

and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Yuxi Xia, Pedro Henrique Luz De Araujo, Klim Zaporojets, and Benjamin Roth. 2025. Influences on LLM calibration: A study of response agreement, loss functions, and prompt styles. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3740–3761, Vienna, Austria. Association for Computational Linguistics.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations*.

An Yang and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

An Yang and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Min Zhang, Jianfeng He, Taoran Ji, and Chang-Tien Lu. 2024. Don't go to extremes: Revealing the excessive sensitivity and calibration limitations of LLMs in implicit hate speech detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12073–12086, Bangkok, Thailand. Association for Computational Linguistics.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. 2018. DAGs with NO TEARS: Continuous Optimization for Structure Learning. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

A Appendix

A.1 Prompt Templates

A.1.1 Name-only Prompt

Does A cause B? Answer Yes or No. Provide your confidence (0-100%) that A directly causes B.
Context: `{context}` Variable A: `{var_a}` Variable B: `{var_b}`
Answer (Yes/No, Confidence%):

A.1.2 Metadata Prompt

Given variable definitions, does A cause B? Answer Yes or No. Provide your confidence (0-100%) that A directly causes B.
Context: `{context}` Variable A: `{var_a}` Definition: `{def_a}` Variable B: `{var_b}` Definition: `{def_b}`
Answer (Yes/No, Confidence%):

A.1.3 Chain-of-Thought Prompt

Given variable definitions, does A cause B? Think step by step before answering.
Context: `{context}` Variable A: `{var_a}` Definition: `{def_a}` Variable B: `{var_b}` Definition: `{def_b}`
Answer (Yes/No): Confidence (0-100%): Reasoning:

A.1.4 Few-shot Prompt

Does A cause B? Answer Yes or No. Provide your confidence (0-100%) that A directly causes B.
Context: `{context}`
Example 1: Variable A: Smoking (tobacco intake) Variable B: Lung Cancer (malignant tumor) Answer: Yes, 95%
Example 2: Variable A: Rain (precipitation) Variable B: Wet Roads (surface moisture) Answer: Yes, 98%
Example 3: Variable A: Exercise (physical activity) Variable B: Income (earnings) Answer: No, 90%
Example 4: Variable A: Height (body length) Variable B: Intelligence (cognitive ability) Answer: No, 88%
Example 5: Variable A: UV Radiation (sun exposure) Variable B: Skin Cancer (malignant melanoma) Answer: Yes, 93%
Now answer: Variable A: `{var_a}` Definition: `{def_a}` Variable B: `{var_b}` Definition: `{def_b}`
Answer (Yes/No, Confidence%):

A.1.5 Few-shot + Chain-of-Thought Prompt

Given variable definitions, does A cause B? Think step by step. Provide answer, confidence, reasoning.
Context: `{context}`
Example 1: Variable A: Smoking (tobacco intake) Variable B: Lung Cancer (malignant tumor) Answer: Yes, 95% Reasoning: Tobacco contains carcinogens that directly damage lung tissue leading to malignant tumor formation.
Example 2: Variable A: Rain (precipitation) Variable B: Wet Roads (surface moisture) Answer: Yes, 98% Reasoning: Rain directly causes water to accumulate on road surfaces.
Example 3: Variable A: Exercise (physical activity) Variable B: Income (earnings) Answer: No, 90% Reasoning: Physical activity level has no direct causal effect on earnings.
Example 4: Variable A: Height (body length) Variable B: Intelligence (cognitive ability) Answer: No, 88% Reasoning: Body height does not causally determine cognitive ability.
Example 5: Variable A: UV Radiation (sun exposure) Variable B: Skin Cancer (malignant melanoma) Answer: Yes, 93% Reasoning: UV radiation directly damages DNA in skin cells, triggering mutations that lead to malignant tumor formation.
Now answer: Variable A: `{var_a}` Definition: `{def_a}` Variable B: `{var_b}` Definition: `{def_b}`
Answer (Yes/No): Confidence (0-100%): Reasoning:

A.1.6 Three-Way Ablation Prompt

```
Given the context and variable definitions, determine the direct causal relationship between Variable A and Variable B.

Choose exactly one label:
A_TO_B: Variable A directly causes Variable B.
B_TO_A: Variable B directly causes Variable A.
NO_EDGE: There is no direct causal edge in either direction.
```

```
Provide your confidence (0-100%) in the
selected label.
Return only the selected label and confidence.
Do not provide reasoning or an explanation.

Context: {context}
Variable A: {var_a}
Definition: {def_a}
Variable B: {var_b}
Definition: {def_b}

Answer (A_TO_B/B_TO_A/NO_EDGE, Confidence%):
```

A.2 Confidence Estimation Details

Formal definitions are provided in Section 2.3. Verbalized confidence is extracted from the standardized answer–confidence response after deterministic parsing of labeled fields and answer-first forms. Logit-based confidence aggregates accepted surface forms of YES and NO before renormalization over the binary label set. Cross-prompt agreement uses predictions from the five prompt styles, whereas cross-model agreement is computed separately within the small- and large-model groups under a fixed prompt. Agreement-based predictions use the majority label.

A.3 Benchmark Dataset Details

A.3.1 Dataset Statistics

Table A1: Reference-graph statistics. No-edge:edge denotes class imbalance over ordered variable pairs.

Dataset	Domain	Source	Nodes	Edges	Pairs	No-edge:Edge
Small networks ($n < 20$)						
AsiaM	Respiratory diagnosis	bnlearn; (Lauritzen and Spiegelhalter, 1988)	7	8	42	4.25:1
River Status	Ecological quality	(Molina-Navarro et al., 2020)	15	25	210	7.40:1
Medium networks ($20 \leq n \leq 50$)						
COVID	Vaccine risk-benefit	BayesFusion; (Mayfield et al., 2022)	20	26	380	13.62:1
Coal Gasifier	Industrial risk	(Liu et al., 2022)	39	39	1,482	37.00:1
Large networks ($51 \leq n \leq 100$)						
Hepar2	Hepatic diagnosis	bnlearn; (Onisko, 2003)	70	123	4,830	38.27:1
Very large networks ($n > 100$)						
Munin1	Electromyography	bnlearn; (Andreassen et al., 1989)	186	273	34,410	125.04:1

Dataset	Number of variables	Mean words per description	Code-like variable names (%)	Descriptions with acronyms (%)
AsiaM	7	5.3	0.0	0.0
River Status	15	3.7	6.7	0.0
COVID	20	6.8	100.0	45.0
Coal Gasifier	39	3.9	100.0	5.1
Hepar2	70	3.4	15.7	0.0
Munin1	186	7.6	100.0	27.4

Table A2: Metadata-level indicators of variable-description complexity across datasets. The table reports simple descriptive measures of the node metadata used in the classification prompts. Lower values for the number of variables, code-like variable names, and acronym-containing descriptions generally indicate easier metadata conditions. Mean words per description is included as a descriptive measure of label length, but is not assumed to have a strictly monotonic relationship with performance because longer descriptions may either clarify or complicate variable meaning.

A.3.2 Dataset Preprocessing

To avoid leaking structural cues to the LLM, we remove explicit references to "Bayesian Network" from the dataset-level context before prompting. Since this term may implicitly suggest that directed edges exist among the variables, we replace it with the neutral term "system," while preserving the original domain meaning of each dataset.

A.4 Model Classification Performance Ranking

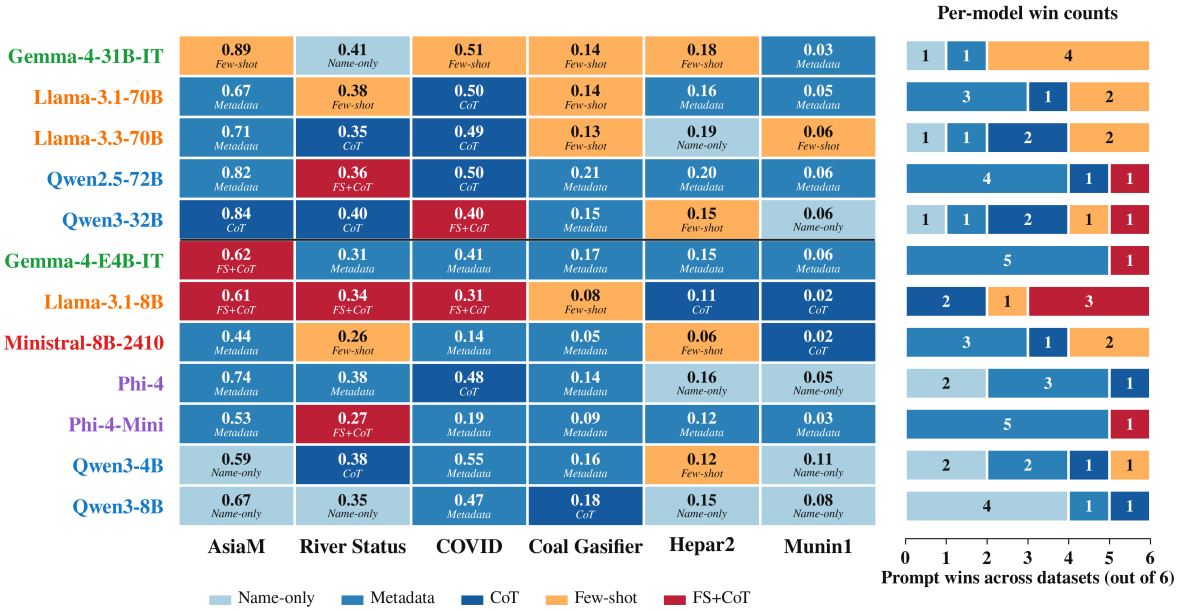

Figure A1: Prompt-level F1 performance across models and datasets. Each cell shows the best F1 score and prompt, with right-side bars summarizing prompt wins per model. The horizontal line separates large and small models.

A.5 Complete Results Tables

A.5.1 Primary Classification Tables

Model	Dataset	Name-only				Metadata				CoT				Few-shot				Few-shot+CoT			
		P↑	R↑	F1↑	nSHD↓	P↑	R↑	F1↑	nSHD↓	P↑	R↑	F1↑	nSHD↓	P↑	R↑	F1↑	nSHD↓	P↑	R↑	F1↑	nSHD↓
Gemma-4-E4B-IT	asiam	0.500	0.375	0.429	0.190	0.417	0.625	0.500	0.214	1.000	0.125	0.222	0.167	0.400	0.500	0.444	0.190	0.625	0.625	0.625	0.143
	river	0.200	0.240	0.218	0.205	0.239	0.440	0.310	0.233	0.212	0.560	0.308	0.300	0.167	0.920	0.282	0.557	0.169	0.880	0.284	0.529
	covid	0.071	0.038	0.050	0.097	0.435	0.385	0.408	0.074	0.421	0.308	0.356	0.076	0.202	0.654	0.309	0.197	0.162	0.654	0.260	0.255
	coal	0.042	0.026	0.032	0.041	0.125	0.256	0.168	0.067	0.096	0.308	0.146	0.093	0.038	0.846	0.073	0.562	0.043	0.718	0.082	0.423
	hepar2	0.096	0.325	0.148	0.094	0.093	0.415	0.152	0.115	0.096	0.179	0.125	0.062	0.046	0.626	0.086	0.335	0.049	0.561	0.091	0.281
	munin1	0.014	0.051	0.022	0.037	0.033	0.194	0.056	0.051	0.018	0.242	0.034	0.109	0.010	0.850	0.019	0.687	0.009	0.802	0.018	0.675
Llama-3.1-8B-Instruct	asiam	0.333	0.875	0.483	0.357	0.375	0.750	0.500	0.262	0.556	0.625	0.588	0.143	0.304	0.875	0.452	0.405	0.467	0.875	0.609	0.214
	river	0.172	0.680	0.274	0.414	0.192	0.560	0.286	0.310	0.245	0.480	0.324	0.224	0.153	0.840	0.259	0.567	0.222	0.720	0.340	0.329
	covid	0.136	0.462	0.211	0.226	0.200	0.462	0.279	0.158	0.095	0.077	0.085	0.111	0.134	0.692	0.225	0.324	0.198	0.692	0.308	0.203
	coal	0.029	0.513	0.055	0.457	0.029	0.359	0.053	0.329	0.018	0.154	0.033	0.233	0.042	0.667	0.078	0.411	0.037	0.410	0.069	0.287
	hepar2	0.044	0.561	0.081	0.320	0.049	0.504	0.090	0.257	0.068	0.366	0.115	0.140	0.046	0.740	0.087	0.391	0.060	0.683	0.110	0.278
	munin1	0.009	0.571	0.018	0.488	0.012	0.509	0.023	0.337	0.013	0.462	0.025	0.288	0.010	0.835	0.020	0.649	0.010	0.615	0.019	0.506
Ministral-8B-Instruct-2410	asiam	0.182	0.750	0.293	0.667	0.292	0.875	0.438	0.405	0.171	0.750	0.279	0.690	0.214	0.750	0.333	0.524	0.241	0.875	0.378	0.524
	river	0.128	0.640	0.213	0.529	0.152	0.680	0.248	0.467	0.113	0.640	0.193	0.614	0.154	0.800	0.258	0.533	0.124	0.760	0.213	0.643
	covid	0.059	0.500	0.106	0.555	0.079	0.577	0.139	0.476	0.053	0.400	0.093	0.495	0.057	0.500	0.102	0.582	0.058	0.500	0.104	0.568
	coal	0.024	0.564	0.047	0.594	0.027	0.564	0.052	0.539	0.026	0.615	0.051	0.601	0.024	0.564	0.046	0.611	0.024	0.641	0.047	0.681
	hepar2	0.026	0.602	0.049	0.587	0.026	0.553	0.050	0.529	0.027	0.577	0.052	0.532	0.034	0.642	0.064	0.470	0.030	0.610	0.058	0.499
	munin1	0.008	0.575	0.015	0.590	0.008	0.601	0.017	0.559	0.009	0.637	0.018	0.539	0.008	0.560	0.016	0.559	0.009	0.641	0.017	0.571
Phi-4	asiam	0.417	0.625	0.500	0.214	0.636	0.875	0.737	0.119	0.500	0.875	0.636	0.190	0.421	1.000	0.593	0.262	0.375	0.750	0.500	0.262
	river	0.238	0.400	0.299	0.224	0.333	0.440	0.379	0.171	0.210	0.680	0.321	0.343	0.220	0.520	0.310	0.276	0.179	0.680	0.283	0.405
	covid	0.450	0.346	0.391	0.071	0.444	0.308	0.364	0.074	0.464	0.500	0.481	0.071	0.296	0.615	0.400	0.126	0.200	0.560	0.295	0.176
	coal	0.000	0.000	0.000	0.074	0.080	0.667	0.143	0.210	0.051	0.744	0.096	0.364	0.050	0.641	0.093	0.327	0.041	0.744	0.079	0.455
	hepar2	0.095	0.537	0.161	0.141	0.088	0.593	0.153	0.165	0.077	0.772	0.141	0.239	0.057	0.732	0.106	0.313	0.055	0.777	0.103	0.337
	munin1	0.029	0.110	0.047	0.035	0.022	0.322	0.041	0.119	0.018	0.480	0.034	0.214	0.015	0.527	0.028	0.287	0.011	0.681	0.021	0.495
Phi-4-Mini-Instruct	asiam	0.375	0.375	0.375	0.190	0.571	0.500	0.533	0.167	0.500	0.250	0.333	0.167	0.000	0.000	0.000	0.262	0.333	0.250	0.286	0.167
	river	0.200	0.208	0.204	0.176	0.227	0.208	0.217	0.171	0.175	0.280	0.215	0.229	0.152	0.400	0.220	0.329	0.208	0.400	0.274	0.252
	covid	0.158	0.115	0.133	0.097	0.192	0.192	0.192	0.108	0.154	0.154	0.154	0.108	0.105	0.160	0.127	0.139	0.080	0.077	0.078	0.121
	coal	0.055	0.105	0.072	0.067	0.064	0.179	0.094	0.089	0.060	0.205	0.092	0.103	0.045	0.282	0.078	0.171	0.044	0.231	0.074	0.148
	hepar2	0.071	0.228	0.109	0.092	0.084	0.203	0.119	0.074	0.075	0.260	0.117	0.098	0.063	0.369	0.107	0.152	0.065	0.382	0.111	0.152
	munin1	0.012	0.185	0.022	0.129	0.016	0.275	0.031	0.137	0.011	0.198	0.021	0.142	0.013	0.292	0.024	0.182	0.013	0.260	0.024	0.164
Qwen3-4B-Instruct	asiam	0.421	1.000	0.593	0.262	0.438	0.875	0.583	0.238	0.333	0.875	0.483	0.357	0.381	1.000	0.552	0.310	0.320	1.000	0.485	0.405
	river	0.222	0.640	0.330	0.310	0.213	0.760	0.333	0.362	0.244	0.833	0.377	0.319	0.207	0.760	0.325	0.376	0.211	0.920	0.343	0.414
	covid	0.615	0.333	0.432	0.061	0.619	0.500	0.553	0.053	0.436	0.654	0.523	0.082	0.321	0.692	0.439	0.121	0.198	0.769	0.315	0.229
	coal	0.000	0.000	0.000	0.028	0.097	0.474	0.161	0.126	0.067	0.513	0.119	0.199	0.060	0.410	0.105	0.184	0.043	0.667	0.080	0.402
	hepar2	0.064	0.516	0.114	0.202	0.058	0.642	0.107	0.271	0.054	0.733	0.101	0.324	0.066	0.650	0.120	0.241	0.052	0.724	0.097	0.340
	munin1	0.089	0.136	0.107	0.017	0.027	0.377	0.050	0.113	0.020	0.509	0.038	0.206	0.017	0.692	0.032	0.329	0.011	0.832	0.021	0.609
Qwen3-8B-Instruct	asiam	0.714	0.625	0.667	0.119	0.538	0.875	0.667	0.167	0.467	0.875	0.609	0.214	0.333	1.000	0.500	0.381	0.389	0.875	0.538	0.286
	river	0.333	0.360	0.346	0.162	0.200	0.320	0.246	0.233	0.183	0.440	0.259	0.300	0.162	0.840	0.271	0.519	0.168	0.800	0.278	0.476
	covid	0.500	0.231	0.316	0.068	0.524	0.423	0.468	0.066	0.600	0.346	0.439	0.061	0.194	0.769	0.310	0.234	0.212	0.846	0.338	0.226
	coal	–	0.000	0.000	0.026	0.082	0.231	0.121	0.088	0.119	0.385	0.182	0.091	0.031	0.667	0.060	0.549	0.035	0.615	0.067	0.454
	hepar2	0.087	0.472	0.147	0.138	0.084	0.553	0.145	0.164	0.071	0.480	0.124	0.170	0.048	0.780	0.090	0.400	0.051	0.772	0.095	0.372
	munin1	0.064	0.117	0.083	0.020	0.026	0.190	0.046	0.061	0.017	0.205	0.031	0.101	0.011	0.941	0.021	0.701	0.010	0.916	0.020	0.714

Table A3: Primary causal edge classification performance for small language models across six benchmark datasets and five prompt styles. Cell colors encode performance percentile within the full table (yellow = good, orange = bad), with intensity scaled per metric direction (P/R/F1 higher is better; nSHD lower is better). A dash denotes undefined precision because the model made no positive predictions.

Small LLMs						Large LLMs					
Model	**Prompt**	**P ↑**	**R ↑**	**F1 ↑**	**Avg. nSHD ↓**	**Model**	**Prompt**	**P ↑**	**R ↑**	**F1 ↑**	**Avg. nSHD ↓**
Gemma-4-E4B-IT	Name-only	0.154	0.176	0.150	**0.111**	Gemma-4-31B-IT	Name-only	0.236	0.524	0.278	**0.150**
	Metadata	0.224	0.386	**0.266**	0.126		Metadata	0.241	**0.841**	0.335	0.222
	CoT	**0.307**	0.287	0.199	0.135		CoT	0.243	0.816	0.330	0.238
	Few-shot	0.144	**0.733**	0.202	0.421		Few-shot	**0.276**	0.738	**0.355**	0.184
	Few-shot + CoT	0.176	0.707	0.227	0.384		Few-shot + CoT	0.234	0.840	0.326	0.255
Llama-3.1-8B-Instruct	Name-only	0.121	0.610	0.187	0.377	Llama-3.1-70B-Instruct	Name-only	0.183	0.273	0.201	0.128
	Metadata	0.143	0.524	0.205	0.275		Metadata	**0.247**	0.318	0.251	**0.096**
	CoT	**0.166**	0.361	0.195	**0.190**		CoT	0.184	**0.869**	**0.272**	0.372
	Few-shot	0.115	**0.775**	0.187	0.458		Few-shot	0.176	0.695	0.261	0.243
	Few-shot + CoT	0.166	0.666	**0.243**	0.303		Few-shot + CoT	0.178	0.830	0.255	0.401
Ministral-8B-Instruct-2410	Name-only	0.071	0.605	0.121	0.587	Llama-3.3-70B-Instruct	Name-only	0.291	0.181	0.211	0.074
	Metadata	**0.097**	0.642	**0.157**	**0.496**		Metadata	**0.295**	0.163	0.199	**0.066**
	CoT	0.067	0.603	0.114	0.579		CoT	0.204	**0.844**	0.293	0.320
	Few-shot	0.082	0.636	0.136	0.547		Few-shot	0.206	0.650	**0.298**	0.178
	Few-shot + CoT	0.081	**0.671**	0.136	0.581		Few-shot + CoT	0.168	0.842	0.251	0.403
Phi-4	Name-only	0.205	0.336	0.233	**0.127**	Qwen2.5-72B-Instruct	Name-only	**0.387**	0.178	0.204	**0.076**
	Metadata	**0.267**	0.534	**0.303**	0.143		Metadata	0.328	0.380	0.304	0.084
	CoT	0.220	0.675	0.285	0.237		CoT	0.267	0.684	**0.320**	0.198
	Few-shot	0.176	0.672	0.255	0.265		Few-shot	0.213	0.658	0.285	0.213
	Few-shot + CoT	0.144	**0.699**	0.213	0.355		Few-shot + CoT	0.214	**0.805**	0.297	0.293
Phi-4-Mini-Instruct	Name-only	0.145	0.203	0.152	0.125	Qwen3-32B-Instruct	Name-only	0.207	0.344	0.208	0.144
	Metadata	**0.192**	0.260	**0.198**	**0.124**		Metadata	0.265	0.535	0.279	0.172
	CoT	0.163	0.225	0.155	0.141		CoT	**0.278**	0.545	**0.314**	**0.143**
	Few-shot	0.063	0.251	0.093	0.206		Few-shot	0.212	0.578	0.276	0.196
	Few-shot + CoT	0.124	**0.267**	0.141	0.167		Few-shot + CoT	0.196	**0.655**	0.275	0.241
Qwen3-4B-Instruct	Name-only	0.235	0.438	0.263	**0.147**						
	Metadata	**0.242**	0.605	**0.298**	0.194						
	CoT	0.192	0.686	0.274	0.248						
	Few-shot	0.175	0.701	0.262	0.260						
	Few-shot + CoT	0.139	**0.819**	0.224	0.400						
Qwen3-8B-Instruct	Name-only	**0.283**	0.301	0.260	**0.089**						
	Metadata	0.242	0.432	**0.282**	0.130						
	CoT	0.243	0.455	0.274	0.156						
	Few-shot	0.130	**0.833**	0.209	0.464						
	Few-shot + CoT	0.144	0.804	0.223	0.421						

Table A4: Macro-averaged primary causal edge classification performance for small (left) and large (right) LLMs across all datasets. Precision (P), recall (R), and F1 are macro-averaged over the six benchmark datasets for each model–prompt pair. Avg. nSHD denotes the macro-average of normalized structural Hamming distance, computed for each dataset as $\mathrm{SHD}/(n_d(n_d-1))$, where n_d is the number of variables in dataset d. Higher P, R, and F1 are better, while lower Avg. nSHD is better. Bold values indicate the best prompt for each model and metric.

	Small LLMs					Large LLMs				
Dataset	**Model**	**Avg. P ↑**	**Avg. R ↑**	**Avg. F1 ↑**	**Avg. nSHD ↓**	**Model**	**Avg. P ↑**	**Avg. R ↑**	**Avg. F1 ↑**	**Avg. nSHD ↓**
asiam	Gemma-4-E4B-IT	**0.588**	0.450	0.444	**0.181**	Qwen3-32B-Instruct	0.603	0.825	0.694	0.124
	Llama-3.1-8B-Instruct	0.407	0.800	0.526	0.276	Gemma-4-31B-IT	**0.705**	**0.900**	**0.780**	**0.090**
	Ministral-8B-Instruct-2410	0.220	0.800	0.344	0.562	Qwen2.5-72B-Instruct	0.680	0.850	0.733	0.114
	Phi-4	0.470	0.825	0.593	0.210	Llama-3.3-70B-Instruct	0.612	0.800	0.638	0.167
	Phi-4-Mini-Instruct	0.356	0.275	0.305	0.190	Llama-3.1-70B-Instruct	0.438	0.775	0.552	0.210
	Qwen3-4B-Instruct	0.379	**0.950**	0.539	0.314					
	Qwen3-8B-Instruct	0.488	0.850	**0.596**	0.233					
river	Gemma-4-E4B-IT	0.197	0.608	0.280	0.365	Qwen3-32B-Instruct	0.223	0.616	0.326	0.303
	Llama-3.1-8B-Instruct	0.197	0.656	0.297	0.369	Gemma-4-31B-IT	**0.258**	**0.810**	**0.391**	0.296
	Ministral-8B-Instruct-2410	0.134	0.704	0.225	0.557	Qwen2.5-72B-Instruct	0.160	0.456	0.220	**0.271**
	Phi-4	**0.236**	0.544	0.318	0.284	Llama-3.3-70B-Instruct	0.228	0.496	0.261	0.272
	Phi-4-Mini-Instruct	0.192	0.299	0.226	**0.231**	Llama-3.1-70B-Instruct	0.213	0.608	0.295	0.315
	Qwen3-4B-Instruct	0.219	**0.783**	**0.342**	0.356					
	Qwen3-8B-Instruct	0.209	0.552	0.280	0.338					
covid	Gemma-4-E4B-IT	0.258	0.408	0.276	0.140	Qwen3-32B-Instruct	0.418	0.362	0.341	0.089
	Llama-3.1-8B-Instruct	0.153	0.477	0.221	0.204	Gemma-4-31B-IT	0.352	**0.800**	**0.486**	0.116
	Ministral-8B-Instruct-2410	0.061	0.495	0.109	0.535	Qwen2.5-72B-Instruct	**0.581**	0.438	0.424	**0.086**
	Phi-4	0.371	0.466	0.386	**0.104**	Llama-3.3-70B-Instruct	0.359	0.531	0.370	0.108
	Phi-4-Mini-Instruct	0.138	0.140	0.137	0.115	Llama-3.1-70B-Instruct	0.352	0.592	0.394	0.122
	Qwen3-4B-Instruct	**0.438**	**0.590**	**0.453**	0.109					
	Qwen3-8B-Instruct	0.406	0.523	0.374	0.131					
coal	Gemma-4-E4B-IT	**0.069**	0.431	**0.100**	0.237	Qwen3-32B-Instruct	0.052	0.501	0.093	0.215
	Llama-3.1-8B-Instruct	0.031	0.421	0.058	0.344	Gemma-4-31B-IT	0.052	**0.608**	0.096	0.241
	Ministral-8B-Instruct-2410	0.025	**0.590**	0.048	0.605	Qwen2.5-72B-Instruct	**0.091**	0.513	**0.121**	**0.178**
	Phi-4	0.045	0.559	0.082	0.286	Llama-3.3-70B-Instruct	0.088	0.436	0.073	0.252
	Phi-4-Mini-Instruct	0.053	0.201	0.082	**0.116**	Llama-3.1-70B-Instruct	0.069	0.467	0.093	0.241
	Qwen3-4B-Instruct	0.053	0.413	0.093	0.188					
	Qwen3-8B-Instruct	0.067	0.379	0.086	0.242					
hepar2	Gemma-4-E4B-IT	**0.076**	0.421	0.120	0.177	Qwen3-32B-Instruct	0.074	0.481	0.128	0.164
	Llama-3.1-8B-Instruct	0.053	0.571	0.097	0.277	Gemma-4-31B-IT	0.093	**0.769**	**0.165**	0.196
	Ministral-8B-Instruct-2410	0.029	0.597	0.055	0.523	Qwen2.5-72B-Instruct	**0.102**	0.541	0.162	**0.151**
	Phi-4	0.074	**0.682**	**0.133**	0.239	Llama-3.3-70B-Instruct	0.091	0.553	0.136	0.194
	Phi-4-Mini-Instruct	0.072	0.288	0.113	**0.113**	Llama-3.1-70B-Instruct	0.072	0.620	0.123	0.249
	Qwen3-4B-Instruct	0.059	0.653	0.108	0.275					
	Qwen3-8B-Instruct	0.068	0.611	0.120	0.249					
munin1	Gemma-4-E4B-IT	0.017	0.428	0.030	0.312	Qwen3-32B-Instruct	**0.021**	0.405	**0.039**	**0.180**
	Llama-3.1-8B-Instruct	0.011	0.599	0.021	0.453	Gemma-4-31B-IT	0.015	**0.625**	0.030	0.319
	Ministral-8B-Instruct-2410	0.008	**0.603**	0.017	0.564	Qwen2.5-72B-Instruct	0.017	0.447	0.032	0.238
	Phi-4	0.019	0.424	0.034	0.230	Llama-3.3-70B-Instruct	0.019	0.401	0.024	0.257
	Phi-4-Mini-Instruct	0.013	0.242	0.024	**0.151**	Llama-3.1-70B-Instruct	0.017	0.519	0.030	0.352
	Qwen3-4B-Instruct	**0.032**	0.509	**0.050**	0.255					
	Qwen3-8B-Instruct	0.026	0.474	0.040	0.320					

Table A5: Dataset-wise prompt-averaged primary causal edge classification performance for small and large LLMs on the six benchmark datasets. For each model–dataset pair, precision (P), recall (R), F1, and normalized structural Hamming distance (nSHD) are averaged across the five prompt styles. nSHD is computed as $\mathrm{SHD}/(n_d(n_d-1))$, where n_d is the number of variables in dataset d. Higher P, R, and F1 are better, while lower Avg. nSHD is better. Bold values indicate the best model within each dataset and metric, with small and large LLMs evaluated independently.

Model	Dataset	Name-only				Metadata				CoT				Few-shot				Few-shot+CoT			
		P↑	R↑	F1↑	nSHD↓	P↑	R↑	F1↑	nSHD↓	P↑	R↑	F1↑	nSHD↓	P↑	R↑	F1↑	nSHD↓	P↑	R↑	F1↑	nSHD↓
Qwen3-32B-Instruct	asiam	0.500	0.625	0.556	0.190	0.667	0.750	0.706	0.095	0.727	1.000	0.842	0.071	0.583	0.875	0.700	0.119	0.538	0.875	0.667	0.143
	river	0.203	0.560	0.298	0.314	0.217	0.520	0.306	0.281	0.273	0.720	0.396	0.262	0.209	0.560	0.304	0.305	0.212	0.720	0.327	0.352
	covid	0.444	0.154	0.229	0.071	0.538	0.269	0.359	0.066	0.500	0.269	0.350	0.068	0.324	0.423	0.367	0.100	0.281	0.692	0.400	0.139
	coal	0.000	0.000	0.000	0.028	0.082	0.684	0.146	0.206	0.076	0.513	0.132	0.176	0.055	0.667	0.101	0.310	0.046	0.641	0.087	0.355
	hepar2	0.060	0.472	0.107	0.196	0.067	0.463	0.117	0.173	0.072	0.463	0.124	0.161	0.086	0.528	0.148	0.150	0.086	0.480	0.146	0.139
	munin1	0.032	0.256	0.057	0.067	0.020	0.524	0.038	0.209	0.021	0.304	0.039	0.117	0.017	0.418	0.034	0.189	0.013	0.524	0.025	0.319
Gemma-4-31B-IT	asiam	0.667	0.500	0.571	0.143	0.667	1.000	0.800	0.095	0.727	1.000	0.842	0.071	0.800	1.000	0.889	0.048	0.667	1.000	0.800	0.095
	river	0.279	0.773	0.410	0.248	0.266	0.875	0.408	0.295	0.247	0.840	0.382	0.324	0.257	0.720	0.379	0.281	0.241	0.840	0.375	0.333
	covid	0.364	0.769	0.494	0.108	0.333	0.808	0.472	0.124	0.318	0.808	0.457	0.132	0.400	0.692	0.507	0.092	0.343	0.923	0.500	0.126
	coal	0.000	0.000	0.000	0.033	0.069	0.784	0.127	0.267	0.059	0.769	0.110	0.326	0.077	0.692	0.138	0.225	0.056	0.795	0.105	0.355
	hepar2	0.092	0.746	0.165	0.190	0.093	0.798	0.167	0.194	0.091	0.813	0.164	0.210	0.105	0.699	0.183	0.157	0.082	0.789	0.148	0.229
	munin1	0.016	0.359	0.030	0.181	0.017	0.783	0.033	0.356	0.014	0.667	0.028	0.367	0.016	0.626	0.032	0.299	0.014	0.692	0.027	0.394
Qwen2.5-72B-Instruct	asiam	0.800	0.500	0.615	0.119	0.778	0.875	0.824	0.071	0.571	1.000	0.727	0.143	0.583	0.875	0.700	0.143	0.667	1.000	0.800	0.095
	river	0.000	0.000	0.000	0.152	0.154	0.080	0.105	0.162	0.213	0.760	0.333	0.362	0.213	0.520	0.302	0.286	0.221	0.920	0.357	0.395
	covid	1.000	0.269	0.424	0.050	0.727	0.308	0.432	0.055	0.611	0.423	0.500	0.058	0.311	0.538	0.394	0.113	0.258	0.654	0.370	0.153
	coal	–	0.000	0.000	0.026	0.144	0.385	0.210	0.074	0.092	0.718	0.163	0.194	0.075	0.718	0.136	0.239	0.053	0.744	0.098	0.358
	hepar2	0.125	0.260	0.168	0.064	0.133	0.374	0.196	0.077	0.094	0.585	0.162	0.152	0.081	0.724	0.146	0.214	0.075	0.764	0.137	0.246
	munin1	0.008	0.040	0.014	0.045	0.034	0.256	0.060	0.063	0.017	0.615	0.033	0.281	0.016	0.571	0.031	0.286	0.012	0.751	0.023	0.512
Llama-3.3-70B-Instruct	asiam	0.750	0.375	0.500	0.143	0.833	0.625	0.714	0.095	0.533	1.000	0.696	0.167	0.471	1.000	0.640	0.214	0.471	1.000	0.640	0.214
	river	0.357	0.200	0.256	0.138	0.143	0.040	0.062	0.143	0.216	0.880	0.346	0.386	0.234	0.600	0.337	0.281	0.190	0.760	0.304	0.414
	covid	0.462	0.231	0.308	0.068	0.400	0.154	0.222	0.071	0.357	0.769	0.488	0.111	0.333	0.769	0.465	0.121	0.244	0.731	0.365	0.171
	coal	0.000	0.000	0.000	0.031	0.286	0.051	0.087	0.028	0.042	0.769	0.079	0.470	0.073	0.590	0.130	0.207	0.037	0.769	0.072	0.523
	hepar2	0.148	0.276	0.193	0.057	0.097	0.106	0.101	0.047	0.065	0.813	0.120	0.303	0.090	0.691	0.159	0.184	0.056	0.878	0.106	0.378
	munin1	0.029	0.004	0.007	0.009	0.010	0.004	0.005	0.011	0.014	0.835	0.027	0.485	0.034	0.253	0.060	0.062	0.010	0.912	0.020	0.717
Llama-3.1-70B-Instruct	asiam	0.333	0.375	0.353	0.238	0.538	0.875	0.667	0.143	0.381	1.000	0.552	0.310	0.400	0.750	0.522	0.214	0.538	0.875	0.667	0.143
	river	0.214	0.360	0.269	0.233	0.200	0.160	0.178	0.176	0.230	0.920	0.368	0.371	0.250	0.800	0.381	0.310	0.171	0.800	0.282	0.486
	covid	0.333	0.346	0.340	0.092	0.538	0.269	0.359	0.063	0.380	0.731	0.500	0.100	0.250	0.808	0.382	0.179	0.256	0.808	0.389	0.174
	coal	0.111	0.051	0.070	0.036	0.070	0.128	0.091	0.065	0.046	0.821	0.086	0.455	0.077	0.564	0.136	0.186	0.042	0.769	0.080	0.462
	hepar2	0.086	0.382	0.140	0.117	0.107	0.341	0.163	0.087	0.055	0.813	0.103	0.359	0.063	0.740	0.116	0.286	0.051	0.821	0.095	0.396
	munin1	0.020	0.121	0.034	0.054	0.029	0.136	0.048	0.042	0.011	0.927	0.023	0.638	0.014	0.505	0.028	0.282	0.010	0.908	0.019	0.745

Table A6: Primary causal edge classification performance for large language models across six benchmark datasets and five prompt styles. Cell colors encode performance percentile within the full table (yellow = good, orange = bad), with intensity scaled per metric direction (P/R/F1 higher is better; nSHD lower is better). A dash denotes undefined precision because the model made no positive predictions.

A.5.2 Calibration Tables

Model	Dataset	Method	Name-only			Metadata			CoT			Few-shot			Few-shot + CoT		
			AUROC ↑	ECE ↓	Brier ↓	AUROC ↑	ECE ↓	Brier ↓	AUROC ↑	ECE ↓	Brier ↓	AUROC ↑	ECE ↓	Brier ↓	AUROC ↑	ECE ↓	Brier ↓
Gemma-4-E4B-IT	asiam	Verb.	0.397	0.513	0.641	0.244	0.417	0.539	0.878	0.010	0.103	0.375	0.051	0.219	0.384	0.024	0.142
		Logit	0.602	0.023	0.143	0.872	0.021	0.063	0.500	0.655	0.810	0.562	0.044	0.205	0.665	0.054	0.141
	river	Verb.	0.320	0.356	0.485	0.192	0.389	0.503	0.395	0.088	0.280	0.195	0.316	0.460	0.266	0.255	0.432
		Logit	0.691	0.019	0.138	0.635	0.016	0.155	0.503	0.552	0.742	0.485	0.262	0.509	0.547	0.294	0.525
	covid	Verb.	0.398	0.550	0.634	0.342	0.590	0.657	0.661	0.009	0.074	0.233	0.081	0.217	0.349	0.057	0.234
		Logit	0.624	0.011	0.063	0.613	0.005	0.058	0.500	0.590	0.768	0.660	0.036	0.178	0.606	0.045	0.225
	coal	Verb.	0.541	0.521	0.561	0.245	0.500	0.547	0.405	0.059	0.143	0.176	0.298	0.461	0.163	0.260	0.379
		Logit	0.499	0.001	0.026	0.774	0.002	0.044	0.500	0.819	0.905	0.492	0.263	0.510	0.585	0.142	0.381
	hepar2	Verb.	0.144	0.533	0.583	0.132	0.548	0.598	0.423	0.037	0.097	0.160	0.130	0.297	0.240	0.108	0.254
		Logit	0.776	0.003	0.065	0.725	0.006	0.086	0.500	0.932	0.965	0.571	0.087	0.306	0.597	0.105	0.328
	munin1	Verb.	0.136	0.523	0.544	0.104	0.499	0.527	0.382	0.040	0.134	0.036	0.502	0.575	0.065	0.465	0.551
		Logit	0.806	0.001	0.015	0.824	0.000	0.030	0.499	0.854	0.924	0.454	0.416	0.619	0.460	0.409	0.615
Llama-3.1-8B-Instruct	asiam	Verb.	0.220	0.311	0.472	0.435	0.231	0.377	0.473	0.023	0.155	0.209	0.313	0.482	0.438	0.037	0.198
		Logit	0.803	0.059	0.229	0.895	0.020	0.138	0.704	0.015	0.115	0.580	0.068	0.303	0.690	0.054	0.203
	river	Verb.	0.226	0.285	0.460	0.343	0.156	0.354	0.512	0.017	0.192	0.337	0.251	0.458	0.436	0.097	0.293
		Logit	0.580	0.018	0.256	0.723	0.004	0.158	0.785	0.014	0.162	0.499	0.164	0.394	0.651	0.032	0.240
	covid	Verb.	0.289	0.368	0.522	0.332	0.331	0.459	0.564	0.051	0.149	0.295	0.174	0.366	0.468	0.058	0.220
		Logit	0.777	0.031	0.095	0.777	0.011	0.069	0.678	0.005	0.065	0.677	0.002	0.193	0.816	0.005	0.101
	coal	Verb.	0.127	0.362	0.510	0.195	0.263	0.432	0.449	0.028	0.205	0.235	0.236	0.427	0.443	0.073	0.271
		Logit	0.936	0.050	0.215	0.748	0.082	0.114	0.835	0.012	0.097	0.671	0.025	0.240	0.809	0.006	0.156
	hepar2	Verb.	0.208	0.340	0.503	0.232	0.322	0.478	0.518	0.033	0.153	0.210	0.243	0.422	0.502	0.053	0.246
		Logit	0.809	0.026	0.151	0.866	0.032	0.094	0.864	0.003	0.066	0.613	0.046	0.278	0.740	0.017	0.186
	munin1	Verb.	0.123	0.367	0.512	0.232	0.206	0.393	0.464	0.028	0.227	0.207	0.390	0.547	0.406	0.192	0.423
		Logit	0.643	0.030	0.257	0.874	0.064	0.095	0.829	0.008	0.121	0.467	0.303	0.456	0.638	0.159	0.333
Ministral-8B-Instruct-2410	asiam	Verb.	0.503	0.387	0.574	0.497	0.131	0.350	0.580	0.392	0.580	0.506	0.175	0.405	0.540	0.147	0.393
		Logit	0.636	0.242	0.390	0.567	0.152	0.336	0.717	0.273	0.417	0.548	0.200	0.375	0.663	0.292	0.429
	river	Verb.	0.391	0.245	0.472	0.420	0.183	0.418	0.458	0.320	0.529	0.306	0.181	0.410	0.410	0.323	0.535
		Logit	0.543	0.272	0.376	0.516	0.135	0.285	0.431	0.290	0.403	0.569	0.240	0.365	0.716	0.384	0.482
	covid	Verb.	0.423	0.209	0.442	0.470	0.162	0.388	0.519	0.168	0.408	0.364	0.225	0.453	0.438	0.238	0.475
		Logit	0.565	0.261	0.342	0.674	0.022	0.233	0.424	0.023	0.265	0.584	0.303	0.380	0.503	0.296	0.388
	coal	Verb.	0.383	0.293	0.514	0.368	0.247	0.474	0.504	0.280	0.508	0.320	0.267	0.481	0.421	0.337	0.550
		Logit	0.292	0.325	0.350	0.284	0.261	0.313	0.306	0.314	0.368	0.308	0.267	0.361	0.323	0.465	0.505
	hepar2	Verb.	0.378	0.280	0.505	0.412	0.232	0.465	0.555	0.222	0.463	0.371	0.120	0.359	0.462	0.152	0.401
		Logit	0.432	0.341	0.381	0.410	0.167	0.315	0.311	0.230	0.341	0.540	0.021	0.265	0.473	0.046	0.290
	munin1	Verb.	0.368	0.289	0.511	0.375	0.254	0.483	0.516	0.226	0.467	0.327	0.195	0.423	0.438	0.205	0.450
		Logit	0.498	0.348	0.357	0.251	0.290	0.323	0.218	0.266	0.332	0.252	0.228	0.335	0.231	0.278	0.371

(continued on next page)

(continued from previous page)

Model	Dataset	Method	Name-only			Metadata			CoT			Few-shot			Few-shot + CoT		
			AUROC ↑	ECE ↓	Brier ↓	AUROC ↑	ECE ↓	Brier ↓	AUROC ↑	ECE ↓	Brier ↓	AUROC ↑	ECE ↓	Brier ↓	AUROC ↑	ECE ↓	Brier ↓
Phi-4	asiam	Verb.	0.680	0.179	0.274	0.624	0.079	0.165	0.756	0.009	0.143	0.677	0.026	0.196	0.602	0.031	0.233
		Logit	0.508	0.019	0.120	0.784	0.016	0.103	0.893	0.655	0.810	0.869	0.092	0.171	0.664	0.079	0.240
	river	Verb.	0.772	0.018	0.156	0.660	0.039	0.162	0.836	0.062	0.210	0.539	0.015	0.212	0.556	0.072	0.309
		Logit	0.771	0.024	0.152	0.783	0.012	0.142	0.627	0.742	0.857	0.829	0.059	0.237	0.373	0.237	0.488
	covid	Verb.	0.671	0.012	0.077	0.699	0.020	0.083	0.715	0.002	0.062	0.627	0.004	0.110	0.760	0.021	0.134
		Logit	0.813	0.007	0.053	0.775	0.002	0.047	0.671	0.755	0.869	0.829	0.005	0.086	0.544	0.020	0.185
	coal	Verb.	0.468	0.558	0.627	0.522	0.183	0.308	0.849	0.072	0.215	0.483	0.035	0.251	0.621	0.105	0.337
		Logit	0.636	0.001	0.026	0.778	0.019	0.164	0.572	0.913	0.954	0.728	0.060	0.265	0.410	0.236	0.484
	hepar2	Verb.	0.825	0.026	0.119	0.810	0.041	0.143	0.888	0.050	0.153	0.626	0.035	0.237	0.662	0.054	0.260
		Logit	0.813	0.007	0.104	0.800	0.010	0.124	0.528	0.913	0.954	0.749	0.061	0.262	0.490	0.133	0.377
	munin1	Verb.	0.436	0.370	0.404	0.683	0.074	0.166	0.872	0.029	0.136	0.480	0.020	0.226	0.523	0.125	0.370
		Logit	0.905	0.000	0.015	0.875	0.003	0.075	0.272	0.895	0.944	0.721	0.023	0.201	0.332	0.278	0.515
Phi-4-Mini-Instruct	asiam	Verb.	0.367	0.410	0.560	0.431	0.343	0.464	0.605	0.136	0.242	0.332	0.183	0.359	0.522	0.125	0.271
		Logit	0.746	0.003	0.138	0.882	0.028	0.109	0.890	0.026	0.122	0.784	0.012	0.114	0.840	0.021	0.124
	river	Verb.	0.315	0.386	0.516	0.287	0.424	0.548	0.381	0.121	0.297	0.212	0.184	0.348	0.356	0.058	0.237
		Logit	0.594	0.009	0.103	0.684	0.010	0.099	0.742	0.003	0.096	0.792	0.003	0.141	0.784	0.007	0.122
	covid	Verb.	0.365	0.399	0.489	0.340	0.432	0.523	0.467	0.077	0.180	0.348	0.163	0.283	0.411	0.101	0.205
		Logit	0.703	0.002	0.059	0.816	0.003	0.057	0.838	0.003	0.058	0.718	0.005	0.063	0.682	0.003	0.062
	coal	Verb.	0.289	0.565	0.624	0.243	0.520	0.590	0.435	0.104	0.200	0.211	0.200	0.314	0.269	0.093	0.216
		Logit	0.550	0.001	0.026	0.790	0.004	0.030	0.705	0.001	0.027	0.765	0.017	0.048	0.696	0.009	0.035
	hepar2	Verb.	0.177	0.626	0.692	0.211	0.597	0.658	0.453	0.083	0.174	0.176	0.237	0.339	0.342	0.101	0.227
		Logit	0.852	0.002	0.036	0.845	0.001	0.031	0.805	0.002	0.028	0.863	0.009	0.060	0.856	0.008	0.050
	munin1	Verb.	0.127	0.573	0.645	0.185	0.482	0.573	0.422	0.067	0.190	0.181	0.126	0.242	0.279	0.090	0.218
		Logit	0.777	0.016	0.025	0.904	0.016	0.034	0.708	0.012	0.020	0.881	0.030	0.048	0.745	0.018	0.028
Qwen3-4B-Instruct	asiam	Verb.	0.166	0.300	0.415	0.320	0.226	0.331	0.354	0.185	0.309	0.231	0.126	0.272	0.358	0.119	0.329
		Logit	0.625	0.514	0.682	0.548	0.529	0.717	0.625	0.212	0.431	0.501	0.097	0.308	0.529	0.163	0.402
	river	Verb.	0.154	0.269	0.370	0.145	0.286	0.398	0.208	0.133	0.299	0.165	0.102	0.289	0.185	0.202	0.362
		Logit	0.524	0.272	0.518	0.500	0.281	0.522	0.547	0.100	0.318	0.502	0.143	0.375	0.499	0.174	0.416
	covid	Verb.	0.386	0.400	0.448	0.384	0.342	0.390	0.398	0.094	0.168	0.194	0.061	0.151	0.235	0.061	0.218
		Logit	0.494	0.435	0.651	0.531	0.249	0.495	0.592	0.077	0.304	0.573	0.014	0.111	0.566	0.051	0.222
	coal	Verb.	0.425	0.273	0.299	0.101	0.253	0.326	0.417	0.032	0.190	0.175	0.057	0.185	0.419	0.115	0.320
		Logit	0.500	0.948	0.974	0.514	0.187	0.433	0.609	0.038	0.204	0.569	0.029	0.171	0.528	0.147	0.386
	hepar2	Verb.	0.080	0.383	0.454	0.099	0.286	0.382	0.265	0.126	0.297	0.128	0.134	0.228	0.230	0.118	0.288
		Logit	0.454	0.418	0.639	0.512	0.169	0.411	0.524	0.159	0.403	0.531	0.054	0.236	0.526	0.107	0.331
	munin1	Verb.	0.118	0.444	0.452	0.029	0.379	0.408	0.269	0.083	0.200	0.085	0.225	0.306	0.403	0.269	0.481
		Logit	0.505	0.971	0.985	0.455	0.629	0.783	0.508	0.224	0.471	0.517	0.101	0.322	0.479	0.346	0.583

(continued on next page)

(continued from previous page)

Model	Dataset	Method	Name-only			Metadata			CoT			Few-shot			Few-shot + CoT		
			AUROC ↑	ECE ↓	Brier ↓	AUROC ↑	ECE ↓	Brier ↓	AUROC ↑	ECE ↓	Brier ↓	AUROC ↑	ECE ↓	Brier ↓	AUROC ↑	ECE ↓	Brier ↓
Qwen3-8B-Instruct	asiam	Verb.	0.346	0.417	0.499	0.247	0.467	0.523	0.658	0.086	0.171	0.255	0.160	0.324	0.243	0.114	0.266
		Logit	0.500	0.655	0.810	0.500	0.655	0.810	0.500	0.655	0.810	0.515	0.206	0.439	0.481	0.068	0.240
	river	Verb.	0.249	0.481	0.577	0.179	0.436	0.532	0.325	0.076	0.240	0.186	0.232	0.416	0.143	0.316	0.436
		Logit	0.500	0.776	0.881	0.500	0.768	0.876	0.500	0.776	0.881	0.438	0.234	0.483	0.513	0.222	0.466
	covid	Verb.	0.266	0.611	0.663	0.198	0.669	0.715	0.554	0.013	0.069	0.165	0.115	0.229	0.128	0.126	0.220
		Logit	0.497	0.615	0.782	0.501	0.843	0.918	0.500	0.868	0.932	0.604	0.062	0.242	0.574	0.049	0.220
	coal	Verb.	0.546	0.478	0.503	0.112	0.552	0.576	0.180	0.051	0.118	0.069	0.374	0.452	0.041	0.375	0.411
		Logit	0.500	0.948	0.974	0.500	0.947	0.973	0.500	0.948	0.974	0.496	0.294	0.538	0.532	0.171	0.415
	hepar2	Verb.	0.047	0.585	0.614	0.048	0.588	0.624	0.417	0.034	0.156	0.098	0.250	0.345	0.054	0.294	0.346
		Logit	0.498	0.909	0.953	0.497	0.934	0.966	0.500	0.950	0.975	0.522	0.152	0.392	0.524	0.125	0.357
	munin1	Verb.	0.132	0.563	0.571	0.040	0.561	0.576	0.240	0.037	0.116	0.141	0.381	0.510	0.025	0.595	0.621
		Logit	0.500	0.984	0.992	0.498	0.983	0.992	0.500	0.984	0.992	0.442	0.468	0.671	0.421	0.485	0.680

Table A7: Prompt-specific calibration results for small LLMs on `asiam`, `river`, `covid`, `coal`, `hepar2`, and `munin1`. Cell colors encode performance percentile within the full table (yellow = good, orange = bad), with intensity scaled separately per metric direction.

Dataset	Small LLMs				Large LLMs			
	Model	AUROC ↑	ECE ↓	Brier ↓	Model	AUROC ↑	ECE ↓	Brier ↓
asiam	Gemma-4-E4B-IT	**0.845**	0.005	**0.110**	Qwen3-32B-Instruct	0.747	0.002	0.077
	Llama-3.1-8B-Instruct	0.689	0.004	0.158	Gemma-4-31B-IT	0.774	**0.001**	**0.056**
	Ministral-8B-Instruct-2410	0.540	0.228	0.428	Qwen2.5-72B-Instruct	0.822	0.006	0.065
	Phi-4	0.773	**0.002**	0.118	Llama-3.3-70B-Instruct	**0.863**	0.014	0.094
	Phi-4-Mini-Instruct	0.779	0.010	0.138	Llama-3.1-70B-Instruct	0.753	0.010	0.148
	Qwen3-4B-Instruct	0.588	0.046	0.240				
	Qwen3-8B-Instruct	0.774	0.018	0.151				
river	Gemma-4-E4B-IT	0.616	0.027	0.234	Qwen3-32B-Instruct	0.642	0.057	0.252
	Llama-3.1-8B-Instruct	0.628	**0.006**	0.212	Gemma-4-31B-IT	0.527	0.071	0.267
	Ministral-8B-Instruct-2410	0.403	0.169	0.392	Qwen2.5-72B-Instruct	**0.806**	0.017	**0.161**
	Phi-4	**0.698**	0.023	0.197	Llama-3.3-70B-Instruct	0.777	**0.009**	0.162
	Phi-4-Mini-Instruct	0.647	0.012	**0.125**	Llama-3.1-70B-Instruct	0.740	0.023	0.202
	Qwen3-4B-Instruct	0.555	0.084	0.306				
	Qwen3-8B-Instruct	0.574	0.038	0.232				
covid	Gemma-4-E4B-IT	0.706	0.011	0.073	Qwen3-32B-Instruct	0.767	0.002	0.056
	Llama-3.1-8B-Instruct	**0.799**	0.005	0.098	Gemma-4-31B-IT	0.734	0.008	0.089
	Ministral-8B-Instruct-2410	0.440	0.126	0.359	Qwen2.5-72B-Instruct	0.729	0.004	**0.054**
	Phi-4	0.782	**0.002**	0.057	Llama-3.3-70B-Instruct	0.882	0.004	0.060
	Phi-4-Mini-Instruct	0.681	0.004	0.071	Llama-3.1-70B-Instruct	**0.888**	**0.001**	0.066
	Qwen3-4B-Instruct	0.784	0.004	**0.056**				
	Qwen3-8B-Instruct	0.788	0.007	0.066				
coal	Gemma-4-E4B-IT	0.704	0.023	0.106	Qwen3-32B-Instruct	0.818	0.022	0.133
	Llama-3.1-8B-Instruct	0.659	**0.001**	0.176	Gemma-4-31B-IT	0.855	0.062	0.166
	Ministral-8B-Instruct-2410	0.360	0.229	0.428	Qwen2.5-72B-Instruct	**0.857**	0.003	**0.100**
	Phi-4	0.766	0.018	0.174	Llama-3.3-70B-Instruct	0.828	0.006	0.123
	Phi-4-Mini-Instruct	0.709	0.014	**0.049**	Llama-3.1-70B-Instruct	0.806	**0.002**	0.118
	Qwen3-4B-Instruct	**0.817**	0.002	0.106				
	Qwen3-8B-Instruct	0.591	0.042	0.120				
hepar2	Gemma-4-E4B-IT	0.691	0.008	0.103	Qwen3-32B-Instruct	0.721	0.007	0.119
	Llama-3.1-8B-Instruct	0.709	**0.002**	0.151	Gemma-4-31B-IT	0.638	0.021	0.163
	Ministral-8B-Instruct-2410	0.466	0.091	0.337	Qwen2.5-72B-Instruct	0.775	0.004	**0.099**
	Phi-4	0.683	0.010	0.159	Llama-3.3-70B-Instruct	**0.821**	**0.001**	0.108
	Phi-4-Mini-Instruct	**0.776**	0.008	**0.052**	Llama-3.1-70B-Instruct	0.762	0.004	0.148
	Qwen3-4B-Instruct	0.677	0.030	0.204				
	Qwen3-8B-Instruct	0.629	0.018	0.168				
munin1	Gemma-4-E4B-IT	0.622	0.029	0.142	Qwen3-32B-Instruct	0.813	0.003	0.107
	Llama-3.1-8B-Instruct	0.551	0.021	0.262	Gemma-4-31B-IT	0.718	0.057	0.233
	Ministral-8B-Instruct-2410	0.423	0.141	0.371	Qwen2.5-72B-Instruct	**0.827**	0.002	0.125
	Phi-4	0.772	**0.001**	0.120	Llama-3.3-70B-Instruct	0.764	0.045	**0.101**
	Phi-4-Mini-Instruct	**0.846**	0.017	**0.055**	Llama-3.1-70B-Instruct	0.699	**0.002**	0.170
	Qwen3-4B-Instruct	0.725	0.004	0.136				
	Qwen3-8B-Instruct	0.511	0.055	0.148				

Table A8: Cross-prompt agreement calibration results for small (left) and large (right) LLMs across all datasets. This method combines predictions across the five prompt styles for each fixed model and dataset. Bold values indicate the best model within each dataset for each metric (small and large LLMs evaluated independently).

Model	Dataset	Method	Name-only			Metadata			CoT			Few-shot			Few-shot + CoT		
			AUROC ↑	ECE ↓	Brier ↓	AUROC ↑	ECE ↓	Brier ↓	AUROC ↑	ECE ↓	Brier ↓	AUROC ↑	ECE ↓	Brier ↓	AUROC ↑	ECE ↓	Brier ↓
Qwen3-32B-Instruct	asiam	Verb.	0.265	0.525	0.637	0.211	0.516	0.603	0.440	0.036	0.097	0.331	0.038	0.154	0.306	0.029	0.162
		Logit	0.615	0.319	0.531	0.676	0.482	0.654	0.749	0.023	0.125	0.667	0.177	0.384	0.711	0.065	0.162
	river	Verb.	0.142	0.512	0.616	0.161	0.515	0.614	0.354	0.055	0.220	0.213	0.049	0.246	0.235	0.085	0.270
		Logit	0.675	0.530	0.688	0.790	0.378	0.544	0.525	0.218	0.458	0.565	0.101	0.319	0.542	0.080	0.300
	covid	Verb.	0.376	0.747	0.806	0.324	0.731	0.783	0.571	0.062	0.124	0.500	0.045	0.136	0.368	0.025	0.146
		Logit	0.607	0.730	0.803	0.744	0.633	0.723	0.654	0.090	0.276	0.614	0.012	0.123	0.710	0.008	0.104
	coal	Verb.	0.503	0.839	0.865	0.055	0.699	0.730	0.154	0.185	0.284	0.059	0.167	0.291	0.113	0.164	0.295
		Logit	0.607	0.722	0.776	0.689	0.630	0.713	0.678	0.039	0.195	0.602	0.121	0.352	0.602	0.080	0.305
	hepar2	Verb.	0.105	0.601	0.664	0.087	0.610	0.665	0.380	0.040	0.165	0.164	0.080	0.180	0.104	0.107	0.169
		Logit	0.630	0.675	0.775	0.667	0.675	0.778	0.574	0.090	0.303	0.538	0.026	0.166	0.686	0.012	0.120
	munin1	Verb.	0.046	0.848	0.859	0.033	0.767	0.796	0.216	0.123	0.208	0.018	0.189	0.248	0.072	0.149	0.280
		Logit	0.818	0.761	0.807	0.704	0.756	0.815	0.513	0.185	0.432	0.648	0.032	0.199	0.681	0.047	0.244
Gemma-4-31B-IT	asiam	Verb.	0.593	0.241	0.361	0.919	0.058	0.100	0.833	0.003	0.065	0.900	0.001	0.040	0.928	0.026	0.076
		Logit	0.544	0.631	0.782	0.458	0.245	0.471	0.500	0.655	0.810	0.500	0.002	0.048	0.500	0.354	0.595
	river	Verb.	0.403	0.447	0.537	0.703	0.165	0.319	0.833	0.091	0.261	0.787	0.052	0.212	0.833	0.089	0.261
		Logit	0.455	0.287	0.498	0.531	0.138	0.388	0.500	0.776	0.881	0.497	0.079	0.281	0.487	0.674	0.819
	covid	Verb.	0.709	0.156	0.235	0.811	0.076	0.146	0.901	0.036	0.110	0.843	0.011	0.076	0.913	0.024	0.102
		Logit	0.410	0.380	0.571	0.588	0.032	0.240	0.500	0.868	0.932	0.504	0.008	0.085	0.495	0.175	0.414
	coal	Verb.	0.453	0.481	0.507	0.392	0.384	0.507	0.877	0.137	0.255	0.715	0.016	0.173	0.838	0.117	0.269
		Logit	0.432	0.921	0.956	0.356	0.408	0.592	0.500	0.948	0.974	0.526	0.048	0.219	0.480	0.807	0.897
	hepar2	Verb.	0.372	0.486	0.560	0.558	0.292	0.392	0.877	0.073	0.171	0.811	0.011	0.121	0.852	0.064	0.177
		Logit	0.308	0.849	0.902	0.348	0.396	0.577	0.500	0.950	0.975	0.521	0.024	0.156	0.492	0.645	0.803
	munin1	Verb.	0.013	0.845	0.855	0.257	0.555	0.652	0.871	0.151	0.287	0.552	0.031	0.240	0.696	0.094	0.304
		Logit	0.436	0.969	0.979	0.463	0.216	0.465	0.500	0.984	0.992	0.554	0.081	0.288	0.487	0.280	0.528
Qwen2.5-72B-Instruct	asiam	Verb.	0.516	0.002	0.107	0.632	0.101	0.141	0.722	0.003	0.113	0.688	0.005	0.116	0.368	0.034	0.107
		Logit	0.558	0.016	0.135	0.500	0.002	0.048	0.500	0.655	0.810	0.571	0.031	0.153	0.408	0.029	0.113
	river	Verb.	0.662	0.001	0.124	0.670	0.012	0.136	0.614	0.036	0.254	0.368	0.038	0.224	0.360	0.068	0.299
		Logit	0.538	0.021	0.133	0.560	0.023	0.141	0.500	0.776	0.881	0.503	0.069	0.257	0.506	0.141	0.379
	covid	Verb.	0.714	0.022	0.066	0.765	0.012	0.059	0.664	0.004	0.057	0.776	0.002	0.090	0.699	0.003	0.120
		Logit	0.526	0.004	0.049	0.550	0.004	0.049	0.500	0.868	0.932	0.520	0.011	0.097	0.609	0.016	0.133
	coal	Verb.	0.433	0.070	0.096	0.434	0.026	0.095	0.600	0.003	0.154	0.349	0.033	0.198	0.318	0.062	0.280
		Logit	0.500	0.001	0.026	0.536	0.003	0.054	0.500	0.948	0.974	0.557	0.048	0.222	0.557	0.100	0.325
	hepar2	Verb.	0.507	0.015	0.075	0.474	0.025	0.095	0.649	0.001	0.125	0.620	0.004	0.164	0.493	0.007	0.193
		Logit	0.551	0.003	0.053	0.533	0.004	0.064	0.500	0.950	0.975	0.576	0.038	0.199	0.579	0.045	0.221
	munin1	Verb.	0.649	0.158	0.201	0.282	0.070	0.127	0.550	0.013	0.212	0.266	0.068	0.229	0.197	0.162	0.367
		Logit	0.500	0.000	0.008	0.617	0.001	0.034	0.500	0.984	0.992	0.559	0.054	0.233	0.511	0.215	0.464

(continued on next page)

(continued from previous page)

Model	Dataset	Method	Name-only			Metadata			CoT			Few-shot			Few-shot + CoT		
			AUROC ↑	ECE ↓	Brier ↓	AUROC ↑	ECE ↓	Brier ↓	AUROC ↑	ECE ↓	Brier ↓	AUROC ↑	ECE ↓	Brier ↓	AUROC ↑	ECE ↓	Brier ↓
Llama-3.3-70B-Instruct	asiam	Verb.	0.331	0.395	0.507	0.145	0.543	0.613	0.696	0.006	0.131	0.497	0.009	0.175	0.367	0.029	0.191
		Logit	0.525	0.029	0.131	0.598	0.021	0.076	0.500	0.028	0.167	0.608	0.046	0.208	0.556	0.044	0.198
	river	Verb.	0.311	0.369	0.476	0.248	0.467	0.570	0.553	0.048	0.281	0.318	0.048	0.231	0.347	0.117	0.339
		Logit	0.562	0.016	0.126	0.616	0.024	0.124	0.520	0.159	0.395	0.561	0.060	0.259	0.524	0.169	0.404
	covid	Verb.	0.304	0.607	0.668	0.407	0.753	0.818	0.799	0.009	0.088	0.509	0.019	0.116	0.506	0.006	0.150
		Logit	0.613	0.008	0.051	0.519	0.004	0.050	0.540	0.012	0.102	0.674	0.012	0.095	0.542	0.031	0.172
	coal	Verb.	0.470	0.939	0.967	0.340	0.487	0.513	0.580	0.090	0.322	0.368	0.012	0.173	0.301	0.226	0.423
		Logit	0.500	0.001	0.026	0.525	0.001	0.026	0.506	0.207	0.456	0.657	0.019	0.163	0.501	0.266	0.514
	hepar2	Verb.	0.164	0.514	0.562	0.140	0.607	0.645	0.748	0.047	0.215	0.433	0.012	0.157	0.523	0.065	0.296
		Logit	0.629	0.003	0.051	0.657	0.002	0.039	0.521	0.087	0.296	0.599	0.024	0.163	0.514	0.135	0.369
	munin1	Verb.	0.339	0.688	0.697	0.417	0.464	0.475	0.507	0.088	0.337	0.076	0.075	0.089	0.150	0.490	0.594
		Logit	0.502	0.000	0.008	0.522	0.000	0.008	0.508	0.218	0.467	0.795	0.001	0.042	0.467	0.502	0.700
Llama-3.1-70B-Instruct	asiam	Verb.	0.317	0.289	0.448	0.273	0.441	0.519	0.637	0.040	0.229	0.686	0.019	0.190	0.406	0.016	0.151
		Logit	0.776	0.026	0.169	0.734	0.026	0.109	0.587	0.045	0.223	0.702	0.042	0.201	0.634	0.021	0.130
	river	Verb.	0.278	0.361	0.502	0.356	0.322	0.457	0.464	0.060	0.285	0.364	0.057	0.253	0.452	0.144	0.385
		Logit	0.760	0.007	0.124	0.714	0.006	0.107	0.477	0.125	0.357	0.705	0.050	0.231	0.520	0.142	0.383
	covid	Verb.	0.191	0.635	0.698	0.287	0.602	0.655	0.576	0.002	0.092	0.516	0.012	0.155	0.524	0.006	0.150
		Logit	0.865	0.001	0.041	0.812	0.001	0.041	0.883	0.009	0.067	0.817	0.011	0.106	0.694	0.010	0.139
	coal	Verb.	0.446	0.384	0.418	0.236	0.396	0.453	0.486	0.087	0.325	0.368	0.033	0.178	0.325	0.165	0.378
		Logit	0.501	0.000	0.027	0.713	0.004	0.028	0.555	0.128	0.373	0.829	0.001	0.105	0.562	0.112	0.357
	hepar2	Verb.	0.167	0.452	0.538	0.235	0.386	0.461	0.610	0.036	0.256	0.495	0.023	0.219	0.511	0.078	0.314
		Logit	0.803	0.001	0.064	0.868	0.001	0.045	0.604	0.063	0.283	0.700	0.024	0.197	0.609	0.089	0.320
	munin1	Verb.	0.112	0.428	0.469	0.201	0.366	0.403	0.386	0.225	0.438	0.201	0.108	0.260	0.221	0.467	0.612
		Logit	0.869	0.002	0.011	0.786	0.003	0.013	0.446	0.316	0.532	0.842	0.012	0.092	0.356	0.497	0.644

Table A9: Prompt-specific calibration results for large LLMs on `asiam`, `river`, `covid`, `coal`, `hepar2`, and `munin1`. ECE is the 10-bin squared-gap measure defined in Section 3.3. Cell colors encode performance percentile within the full table (yellow = good, orange = bad), with intensity scaled separately per metric direction.

Dataset	Prompt	Small LLMs			Large LLMs		
		AUROC ↑	ECE ↓	Brier ↓	AUROC ↑	ECE ↓	Brier ↓
asiam	Name-only	0.772	**0.008**	0.135	0.632	0.022	0.114
	Metadata	0.711	0.014	**0.105**	0.656	0.012	**0.054**
	CoT	0.655	0.024	0.131	0.829	**0.008**	0.099
	Few-shot	0.671	0.012	0.207	**0.871**	0.018	0.095
	Few-shot + CoT	**0.833**	0.019	0.158	0.523	0.016	0.100
river	Name-only	**0.722**	0.003	**0.147**	0.665	**0.009**	0.136
	Metadata	0.711	**0.001**	0.150	**0.726**	0.010	**0.124**
	CoT	0.654	0.009	0.193	0.631	0.086	0.292
	Few-shot	0.478	0.051	0.291	0.708	0.036	0.216
	Few-shot + CoT	0.566	0.046	0.286	0.526	0.102	0.338
covid	Name-only	0.620	0.015	0.068	0.774	0.005	0.045
	Metadata	0.729	0.010	**0.056**	0.794	**0.001**	**0.043**
	CoT	0.748	0.011	0.057	**0.903**	0.005	0.052
	Few-shot	**0.759**	**0.005**	0.111	0.828	0.005	0.080
	Few-shot + CoT	0.733	0.005	0.123	0.691	0.007	0.113
coal	Name-only	0.532	0.032	**0.058**	0.513	**0.001**	**0.027**
	Metadata	**0.802**	0.014	0.080	**0.868**	0.002	0.062
	CoT	0.768	**0.011**	0.102	0.594	0.029	0.228
	Few-shot	0.629	0.014	0.235	0.699	0.010	0.159
	Few-shot + CoT	0.650	0.027	0.249	0.554	0.077	0.315
hepar2	Name-only	0.738	0.008	**0.099**	0.743	0.003	0.071
	Metadata	0.757	**0.005**	0.102	**0.790**	**0.002**	**0.064**
	CoT	**0.800**	0.006	0.099	0.689	0.009	0.160
	Few-shot	0.687	0.006	0.189	0.695	0.007	0.137
	Few-shot + CoT	0.701	0.006	0.186	0.691	0.019	0.193
munin1	Name-only	0.725	0.037	**0.054**	0.790	0.011	**0.025**
	Metadata	**0.808**	0.020	0.069	**0.878**	0.009	0.056
	CoT	0.798	**0.011**	0.093	0.627	0.030	0.248
	Few-shot	0.523	0.044	0.293	0.760	**0.001**	0.122
	Few-shot + CoT	0.492	0.108	0.349	0.467	0.151	0.396

Table A10: Cross-model agreement calibration results for small (left) and large (right) LLMs across all datasets. This method combines predictions across all models in each group for each fixed prompt style and dataset. Bold values indicate the best prompt within each dataset for each metric (small and large LLMs evaluated independently).

A.6 Statistical Comparison of Confidence Methods

Table A11 reports pairwise comparisons among the four confidence methods. Tests are performed separately within each model group and metric using the six paired dataset-level scores.

		Small LLMs				Large LLMs		
Metric	**Comparison**	$\widetilde{\Delta}$	**Raw** p	**Holm** p	**Comparison**	$\widetilde{\Delta}$	**Raw** p	**Holm** p
ECE	CM vs. CP	-0.022	0.0625	0.1875	CM vs. CP	0.007	0.0312	0.1875
	CM vs. Logit	-0.196	0.0312	0.1875	CM vs. Logit	-0.189	0.0312	0.1875
	CM vs. Verb.	-0.207	0.0312	0.1875	CM vs. Verb.	-0.179	0.0312	0.1875
	CP vs. Logit	-0.173	0.0312	0.1875	CP vs. Logit	-0.196	0.0312	0.1875
	CP vs. Verb.	-0.191	0.0312	0.1875	CP vs. Verb.	-0.179	0.0312	0.1875
	Logit vs. Verb.	-0.001	0.6875	0.6875	Logit vs. Verb.	0.019	0.4375	0.4375
Brier	CM vs. CP	-0.029	0.0312	0.1875	CM vs. CP	0.008	0.0938	0.1875
	CM vs. Logit	-0.200	0.0312	0.1875	CM vs. Logit	-0.216	0.0312	0.1875
	CM vs. Verb.	-0.229	0.0312	0.1875	CM vs. Verb.	-0.189	0.0312	0.1875
	CP vs. Logit	-0.170	0.0312	0.1875	CP vs. Logit	-0.219	0.0312	0.1875
	CP vs. Verb.	-0.199	0.0312	0.1875	CP vs. Verb.	-0.188	0.0312	0.1875
	Logit vs. Verb.	-0.011	0.4375	0.4375	Logit vs. Verb.	0.032	0.1562	0.1875
AUROC	CM vs. CP	0.026	0.0312	0.1875	CM vs. CP	-0.054	0.0312	0.1875
	CM vs. Logit	0.089	0.0312	0.1875	CM vs. Logit	0.118	0.0312	0.1875
	CM vs. Verb.	0.321	0.0312	0.1875	CM vs. Verb.	0.226	0.0312	0.1875
	CP vs. Logit	0.059	0.0625	0.1875	CP vs. Logit	0.181	0.0312	0.1875
	CP vs. Verb.	0.309	0.0312	0.1875	CP vs. Verb.	0.286	0.0312	0.1875
	Logit vs. Verb.	0.246	0.0312	0.1875	Logit vs. Verb.	0.131	0.0312	0.1875

Table A11: Pairwise two-sided Wilcoxon signed-rank tests over six paired dataset-level scores, small (left) and large (right) LLMs evaluated separately. $\widetilde{\Delta}$ is the median difference (first method minus second). Holm correction is applied across the six pairwise comparisons within each model group and metric. CM = cross-model agreement, CP = cross-prompt agreement, Verb. = verbalized confidence. With $N = 6$ paired observations, the smallest attainable two-sided raw p-value is $2/2^6 \approx 0.031$; no adjusted (Holm) p-value falls below 0.05.

A.7 Overconfident False Positives on Non-Edges

(a) Confidence buckets among false positives.

Non-edge type	False-positive predictions	Conf. < 50	Conf. 50–79	Conf. ≥ 80
Reversed direct	10,635	0.8%	14.7%	84.6%
Indirect	64,503	0.6%	18.6%	80.8%
Other	636,333	0.9%	20.7%	78.4%

(b) High-confidence false positives by prompt.

Prompt	Reversed direct non-edge	Indirect non-edge	Other non-edge
Few-shot + CoT	46.4%	55.2%	42.5%
Few-shot	36.6%	35.8%	25.0%
CoT	30.3%	35.2%	21.5%
Metadata	21.8%	21.7%	12.1%
Name-only	16.9%	13.6%	9.5%

Table A12: Overconfident false-positive summaries on reference-graph non-edges. Panel (a) reports confidence distributions using false positives with a parseable verbal confidence score (99.9% of all false positives); its counts therefore differ slightly from the all-inclusive counts in Table 4. Panel (b) reports high-confidence false-positive rates over valid non-edge queries by prompt and non-edge type.

Non-edge type	Model	False-positive rate	High-conf. false-positive rate	False positives with conf. ≥ 80
Reversed direct	Ministral-8B-Instruct-2410	58.6%	45.0%	76.8%
Indirect	Gemma-4-31B-IT	56.3%	52.2%	92.7%
Other	Ministral-8B-Instruct-2410	56.4%	41.0%	72.7%

Table A13: Worst model-level high-confidence false-positive behavior by non-edge type. For each non-edge category, we report the model with the largest high-confidence false-positive rate, along with its overall false-positive rate and the share of its false positives that have verbal confidence of at least 80.

A.8 Additional Classification Figures

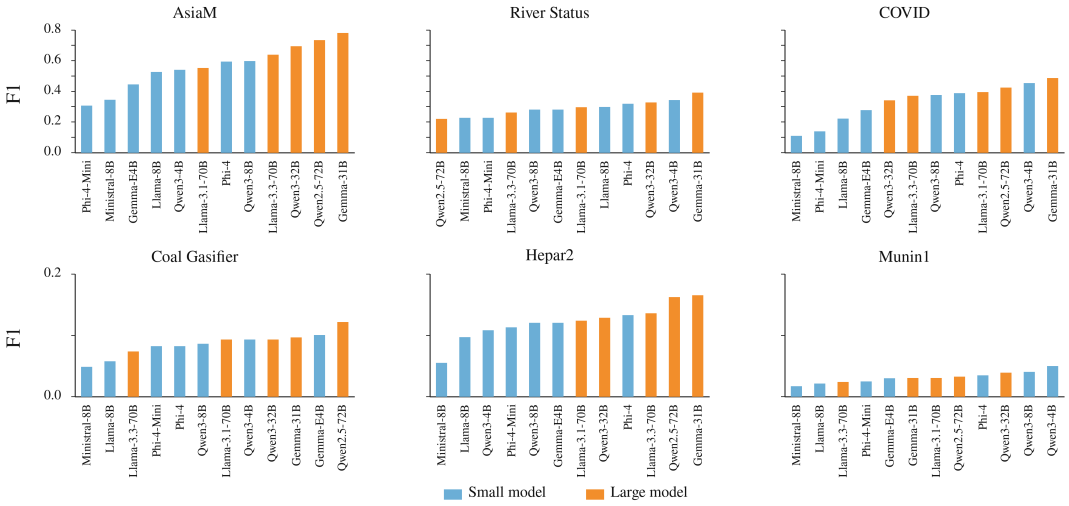

Figure A2: Model classification performance ranking across datasets. Performance is averaged over five prompt styles.

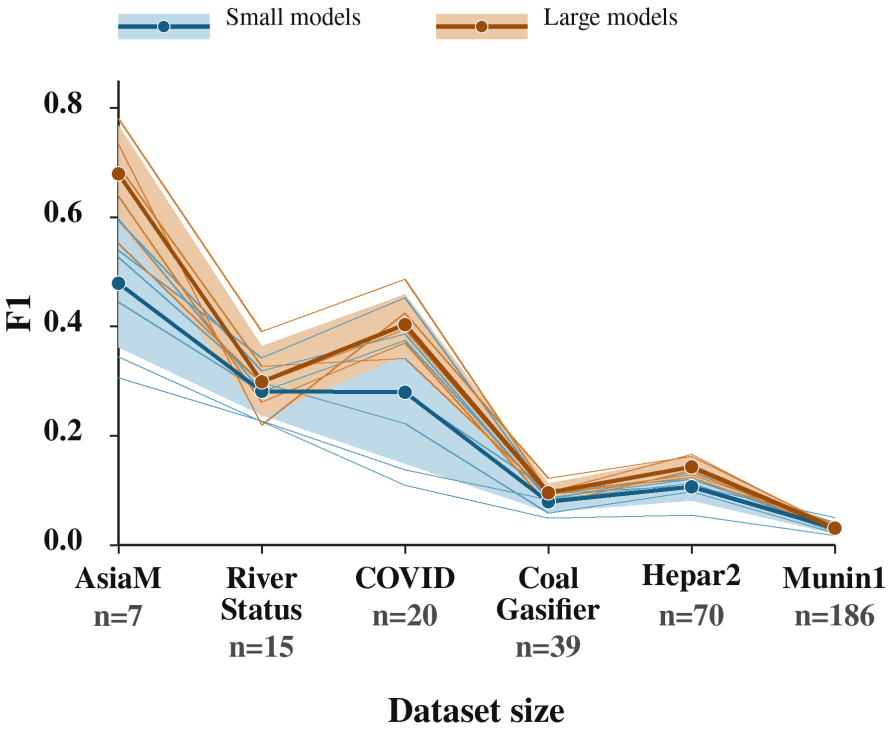

Figure A3: Model classification performance as a function of dataset size. Datasets are ordered by increasing number of variables. Points show mean F1 averaged over prompt styles, and shaded bands show variation across models within each group.

A.9 Reliability Diagram Examples

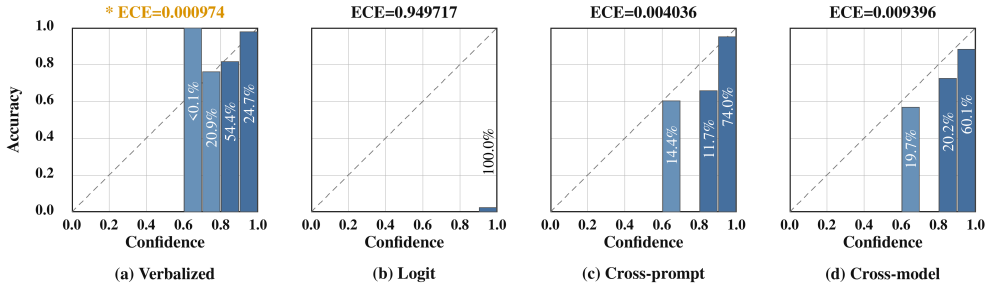

Figure A4: Reliability diagrams for Qwen2.5-72B-Instruct on Hepar2 with CoT prompting across four confidence sources: (a) verbalized, (b) logit-based, (c) cross-prompt, and (d) cross-model. Each panel reports ECE; ⋆ marks the lowest-ECE source, here verbalized confidence.

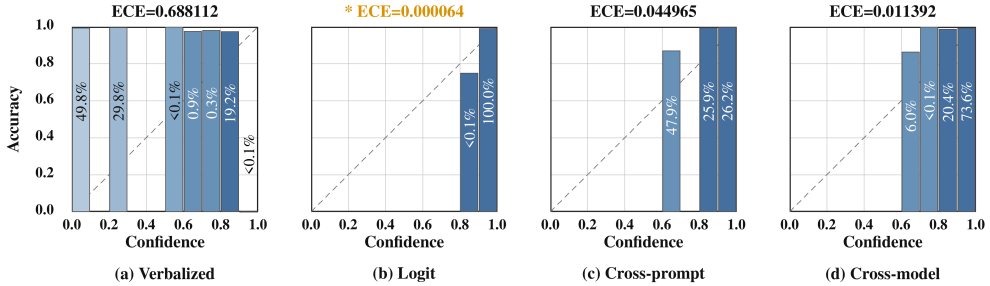

Figure A5: Reliability diagrams for Llama-3.3-70B-Instruct on Munin1 with name-only prompting across four confidence sources: (a) verbalized, (b) logit-based, (c) cross-prompt, and (d) cross-model. Each panel reports ECE; ⋆ marks the lowest-ECE source, here logit-based confidence.

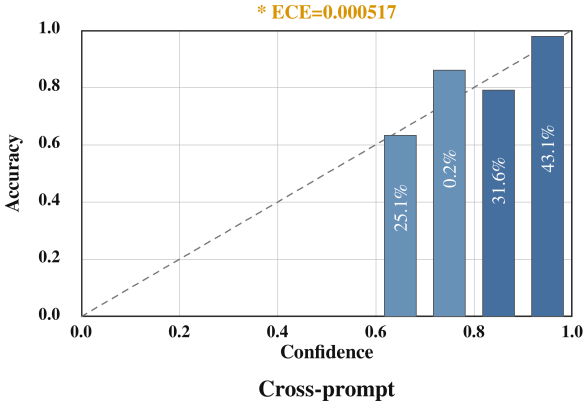

Figure A6: Reliability diagram for cross-prompt agreement on Munin1 with Phi-4.

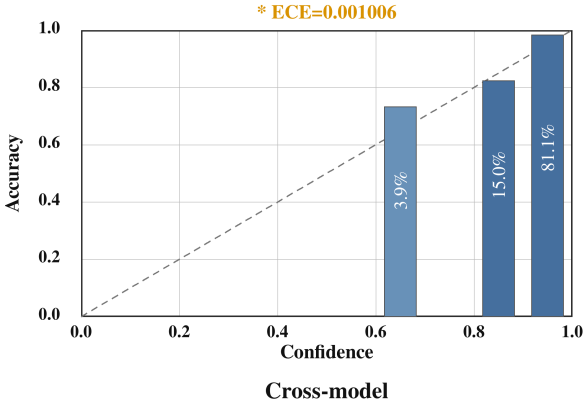

Figure A7: Reliability diagram for cross-model agreement on COVID under metadata prompting.

A.10 Additional Confidence Distributions

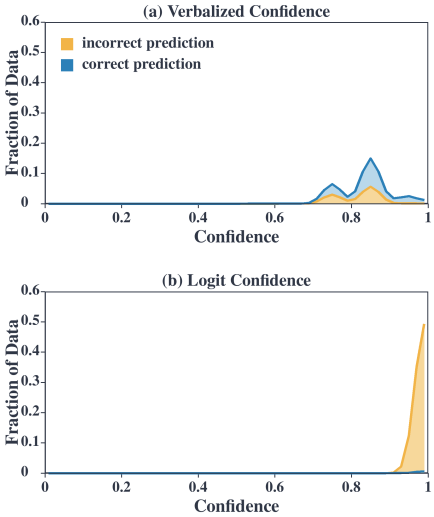

Figure A8: Confidence distributions for Qwen2.5-72B-Instruct under the Chain-of-Thought prompt. Predictions are pooled across all six benchmark datasets. Blue curves denote correctly classified variable pairs, while orange curves denote misclassified pairs. Panel (a) shows verbalized confidence, and panel (b) shows logit-based confidence. Both confidence sources concentrate in the high-confidence region, with logit-based confidence collapsing almost entirely near 1.0 for both correct and incorrect predictions, indicating strong overconfidence and limited separation between reliable and unreliable predictions.

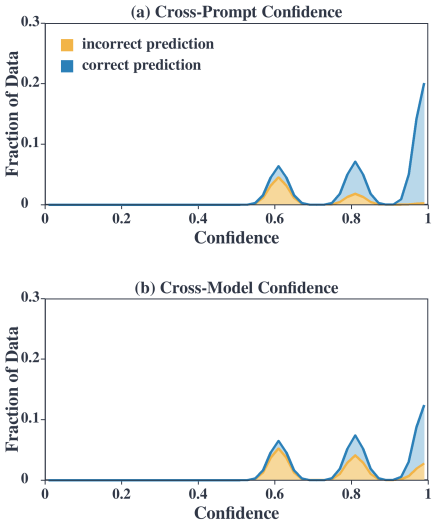

Figure A9: Agreement-based confidence distributions for Qwen2.5-72B-Instruct and the large-model ensemble. Panel (a) shows cross-prompt agreement for Qwen2.5-72B-Instruct, while panel (b) shows cross-model agreement under the Chain-of-Thought prompt. Correct predictions concentrate more strongly near full agreement, indicating that agreement-based confidence is more informative than raw confidence for identifying reliable predictions.

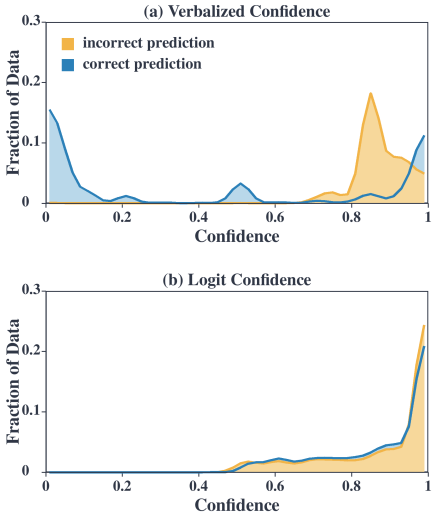

Figure A10: Confidence distributions for Gemma-4-31B-IT using the Metadata prompt.

A.11 Three-Way Label-Space Ablation Results

Model	Dataset	P↑	R↑	F1↑	nSHD↓	Non-edge FP Rate↓	Orientation Error↓	$A \to B$ Recall↑	$B \to A$ Recall↑
Qwen3-4B	River Status	0.247	0.760	0.373	0.305	72.5%	0.0%	0.760	–
	COVID	0.167	0.500	0.250	0.197	39.6%	18.8%	0.812	0.000
	Hepar2	0.067	0.813	0.124	0.291	60.6%	5.7%	0.909	0.000
	Munin1	0.016	0.637	0.031	0.311	62.9%	22.7%	0.879	0.000
Phi-4	River Status	0.287	0.920	0.438	0.281	71.3%	0.0%	0.920	–
	COVID	0.471	0.615	0.533	0.071	11.0%	5.9%	0.812	0.300
	Hepar2	0.102	0.675	0.177	0.159	32.0%	8.8%	0.718	0.308
	Munin1	0.035	0.297	0.062	0.070	13.2%	17.3%	0.404	0.013
Gemma-4-31B	River Status	0.293	0.880	0.440	0.267	66.2%	0.0%	0.880	–
	COVID	0.358	0.731	0.481	0.108	20.7%	0.0%	0.812	0.600
	Hepar2	0.079	0.724	0.142	0.220	45.5%	15.2%	0.764	0.385
	Munin1	0.018	0.520	0.035	0.224	45.2%	30.4%	0.556	0.427
Llama-3.3-70B	River Status	0.278	1.000	0.435	0.310	81.2%	0.0%	1.000	–
	COVID	0.242	0.615	0.348	0.153	30.5%	11.1%	0.875	0.200
	Hepar2	0.075	0.846	0.138	0.267	55.6%	5.5%	0.900	0.385
	Munin1	0.013	0.718	0.025	0.448	91.0%	23.1%	0.990	0.000

Table A14: Dataset-level three-way classification results under Metadata prompting. Non-edge FP rate is the fraction of true no-edge pairs assigned either directed label. Orientation error is the fraction of detected true-edge pairs assigned the reversed direction. A dash indicates that the dataset contains no $B \to A$ instances under the fixed unordered-pair ordering.

A.12 Benchmark Familiarity Audit

	Small LLMs						Large LLMs					
Dataset	**Model**	**Nodes**	**Matches**	**Recall**	**Dev.**	**Risk**	**Model**	**Nodes**	**Matches**	**Recall**	**Dev.**	**Risk**
asiam	Gemma-4-E4B-IT	10/7	2	0.286	0.429	No	Qwen3-32B-Instruct	7/7	6	0.857	0.000	**Yes**
	Llama-3.1-8B-Instruct	8/7	1	0.143	0.143	No	Gemma-4-31B-IT	7/7	7	1.000	0.000	**Yes**
	Ministral-8B-Instruct-2410	1/7	1	0.143	0.857	No	Qwen2.5-72B-Instruct	8/7	7	1.000	0.143	**Yes**
	Phi-4	8/7	6	0.857	0.143	**Yes**	Llama-3.3-70B-Instruct	8/7	7	1.000	0.143	**Yes**
	Phi-4-Mini-Instruct	5/7	3	0.429	0.286	No	Llama-3.1-70B-Instruct	10/7	7	1.000	0.429	No
	Qwen3-4B-Instruct	2/7	1	0.143	0.714	No						
	Qwen3-8B-Instruct	9/7	6	0.857	0.286	No						
river	Gemma-4-E4B-IT	16/15	6	0.400	0.067	No	Qwen3-32B-Instruct	9/15	8	0.533	0.400	No
	Llama-3.1-8B-Instruct	11/15	3	0.200	0.267	No	Gemma-4-31B-IT	10/15	9	0.600	0.333	No
	Ministral-8B-Instruct-2410	9/15	7	0.467	0.400	No	Qwen2.5-72B-Instruct	8/15	6	0.400	0.467	No
	Phi-4	8/15	2	0.133	0.467	No	Llama-3.3-70B-Instruct	9/15	7	0.467	0.400	No
	Phi-4-Mini-Instruct	9/15	7	0.467	0.400	No	Llama-3.1-70B-Instruct	13/15	3	0.200	0.133	No
	Qwen3-4B-Instruct	10/15	1	0.067	0.333	No						
	Qwen3-8B-Instruct	3/15	1	0.067	0.800	No						
covid	Gemma-4-E4B-IT	12/20	6	0.300	0.400	No	Qwen3-32B-Instruct	15/20	8	0.400	0.250	No
	Llama-3.1-8B-Instruct	14/20	1	0.050	0.300	No	Gemma-4-31B-IT	6/20	3	0.150	0.700	No
	Ministral-8B-Instruct-2410	13/20	7	0.350	0.350	No	Qwen2.5-72B-Instruct	12/20	2	0.100	0.400	No
	Phi-4	13/20	2	0.100	0.350	No	Llama-3.3-70B-Instruct	15/20	2	0.100	0.250	No
	Phi-4-Mini-Instruct	6/20	5	0.250	0.700	No	Llama-3.1-70B-Instruct	13/20	11	0.550	0.350	No
	Qwen3-4B-Instruct	16/20	1	0.050	0.200	No						
	Qwen3-8B-Instruct	3/20	0	0.000	0.850	No						
coal	Gemma-4-E4B-IT	8/39	7	0.179	0.795	No	Qwen3-32B-Instruct	1/39	0	0.000	0.974	No
	Llama-3.1-8B-Instruct	10/39	7	0.179	0.744	No	Gemma-4-31B-IT	9/39	0	0.000	0.769	No
	Ministral-8B-Instruct-2410	11/39	10	0.256	0.718	No	Qwen2.5-72B-Instruct	10/39	7	0.179	0.744	No
	Phi-4	7/39	6	0.154	0.821	No	Llama-3.3-70B-Instruct	19/39	10	0.256	0.513	No
	Phi-4-Mini-Instruct	6/39	6	0.154	0.846	No	Llama-3.1-70B-Instruct	8/39	7	0.179	0.795	No
	Qwen3-4B-Instruct	15/39	8	0.205	0.615	No						
	Qwen3-8B-Instruct	3/39	0	0.000	0.923	No						
hepar2	Gemma-4-E4B-IT	8/70	8	0.114	0.886	No	Qwen3-32B-Instruct	14/70	11	0.157	0.800	No
	Llama-3.1-8B-Instruct	15/70	5	0.071	0.786	No	Gemma-4-31B-IT	13/70	12	0.171	0.814	No
	Ministral-8B-Instruct-2410	1/70	0	0.000	0.986	No	Qwen2.5-72B-Instruct	13/70	12	0.171	0.814	No
	Phi-4	13/70	13	0.186	0.814	No	Llama-3.3-70B-Instruct	14/70	11	0.157	0.800	No
	Phi-4-Mini-Instruct	12/70	9	0.129	0.829	No	Llama-3.1-70B-Instruct	12/70	11	0.157	0.829	No
	Qwen3-4B-Instruct	13/70	12	0.171	0.814	No						
	Qwen3-8B-Instruct	3/70	2	0.029	0.957	No						
munin1	Gemma-4-E4B-IT	13/186	8	0.043	0.930	No	Qwen3-32B-Instruct	7/186	7	0.038	0.962	No
	Llama-3.1-8B-Instruct	14/186	8	0.043	0.925	No	Gemma-4-31B-IT	12/186	12	0.065	0.935	No
	Ministral-8B-Instruct-2410	9/186	9	0.048	0.952	No	Qwen2.5-72B-Instruct	15/186	14	0.075	0.919	No
	Phi-4	7/186	7	0.038	0.962	No	Llama-3.3-70B-Instruct	14/186	13	0.070	0.925	No
	Phi-4-Mini-Instruct	3/186	3	0.016	0.984	No	Llama-3.1-70B-Instruct	21/186	17	0.091	0.887	No
	Qwen3-4B-Instruct	15/186	12	0.065	0.919	No						
	Qwen3-8B-Instruct	2/186	1	0.005	0.989	No						

Table A15: Full node-level contamination results for all model–dataset pairs, small (left) and large (right) LLMs evaluated separately. Nodes reports generated/true counts, and Matches reports semantically aligned nodes. High-risk pairs (node-count deviation below 15% and node recall above 0.85) are highlighted.

A.13 Post-hoc Temperature Calibration

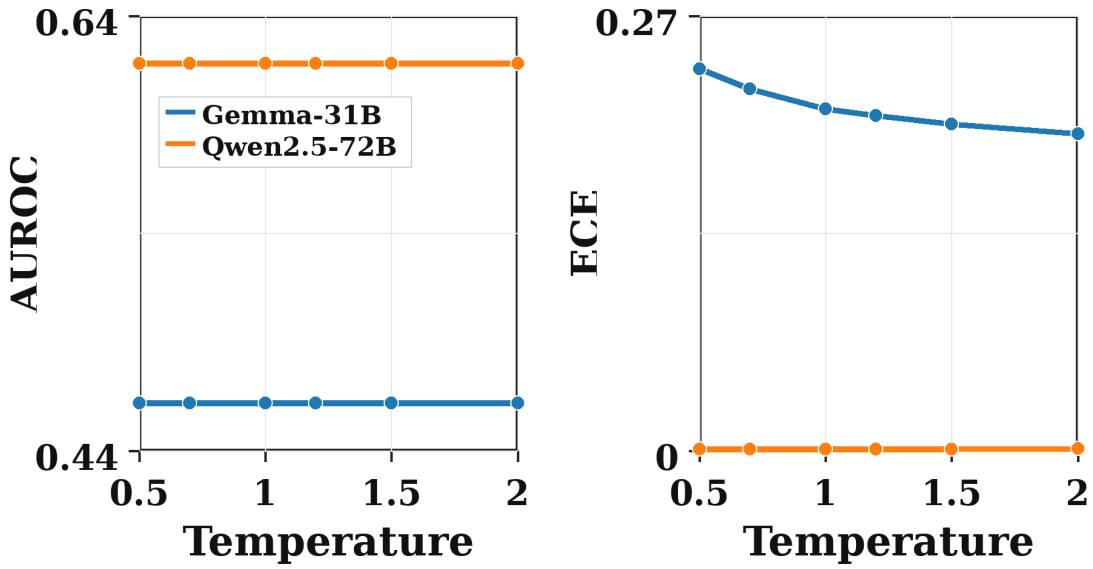

Figure A11: Post-hoc temperature sensitivity of logit-based calibration on Munin1.

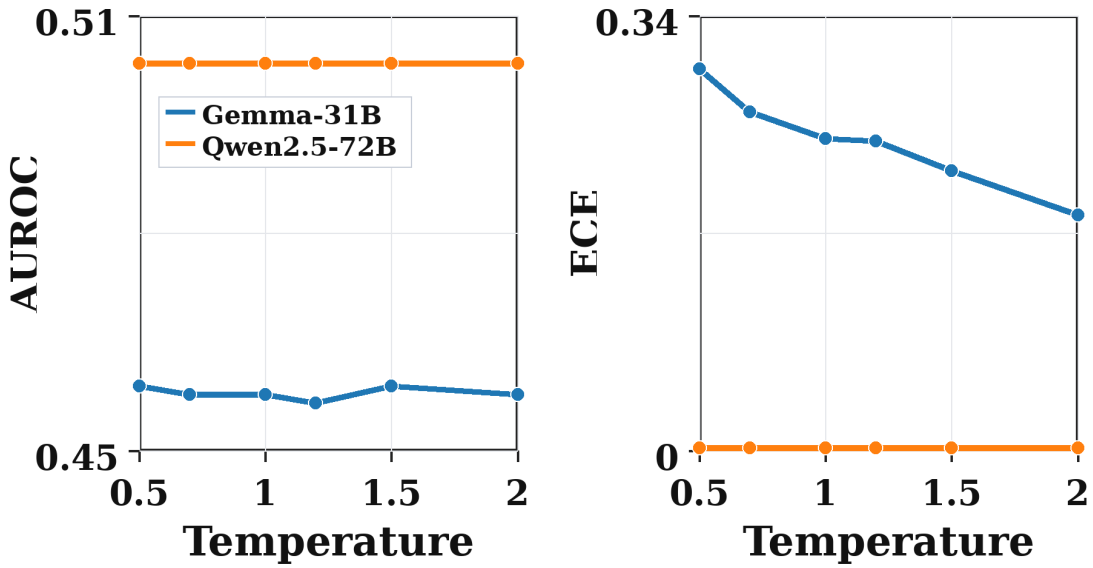

Figure A12: Post-hoc temperature sensitivity of logit-based calibration on AsiaM. AUROC is nearly unchanged because temperature scaling preserves the logit ranking, while Gemma-31B's ECE decreases at higher temperatures, indicating overconfident Yes/No logits. Qwen2.5-72B remains stable, suggesting better calibrated logit confidence.

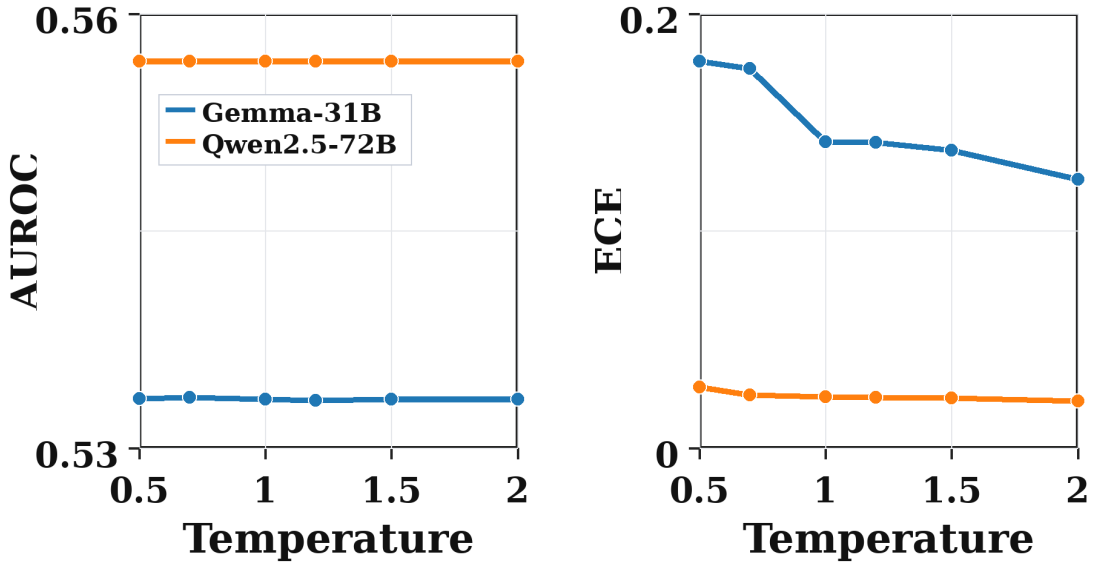

Figure A13: Post-hoc temperature sensitivity of logit-based calibration on River Status. The main effect of temperature appears in ECE rather than AUROC, showing that temperature changes confidence sharpness rather than discrimination. Gemma-31B benefits from softer logits, while Qwen2.5-72B is comparatively insensitive.

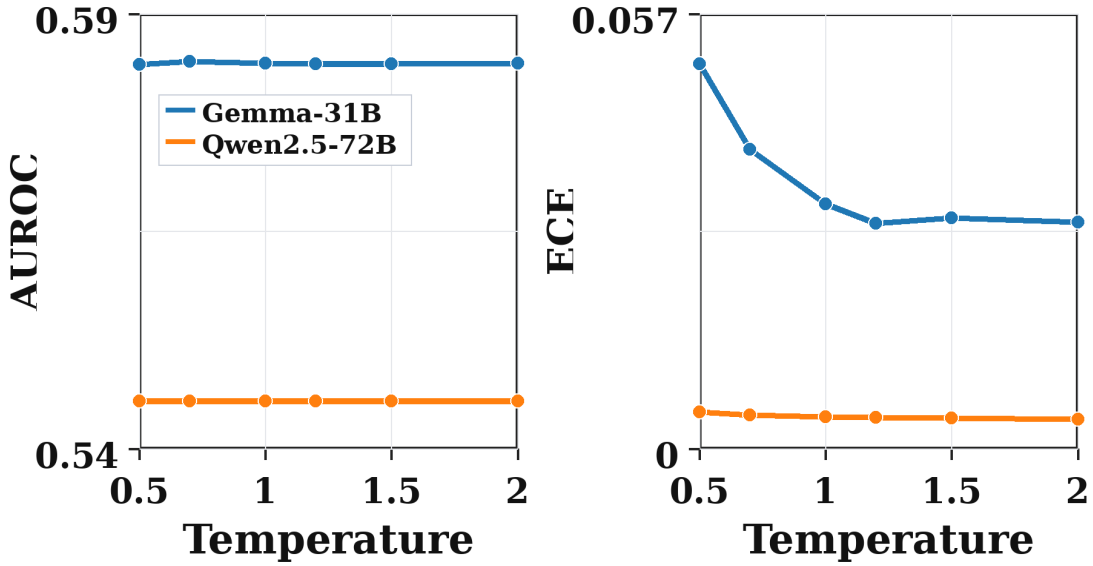

Figure A14: Post-hoc temperature sensitivity of logit-based calibration on COVID. AUROC remains stable across temperatures, confirming that prediction ranking is preserved. ECE varies modestly, with Gemma-31B showing some calibration benefit from temperature adjustment and Qwen2.5-72B remaining consistently well calibrated.

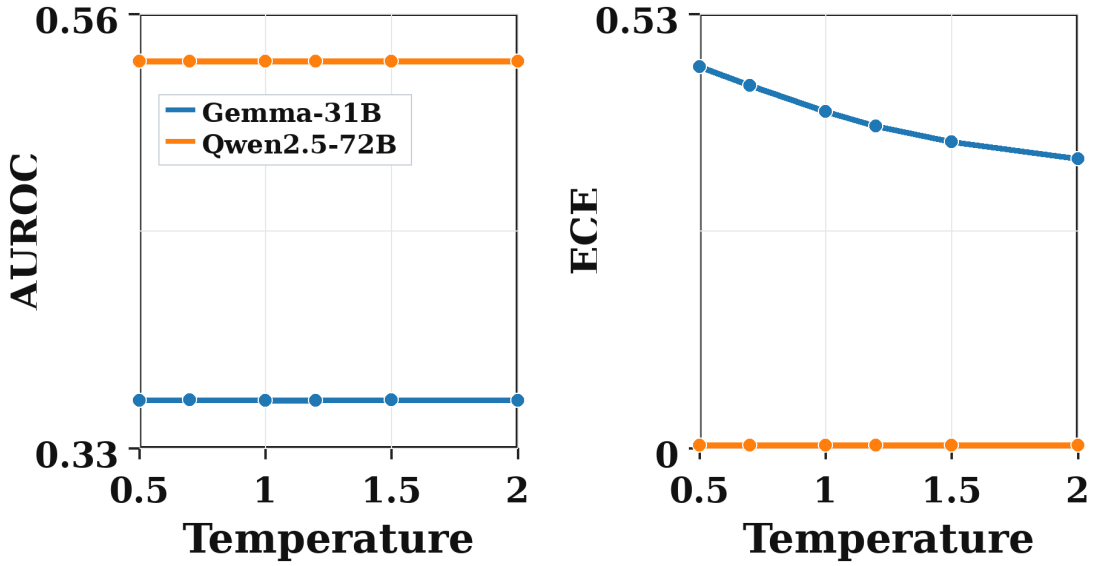

Figure A15: Post-hoc temperature sensitivity of logit-based calibration on Coal Gasifier. Gemma-31B shows a clear ECE reduction as temperature increases, suggesting its saved Yes/No logits are too sharp for reliable confidence estimates. Qwen2.5-72B has low and stable ECE, indicating stronger post-hoc calibration.

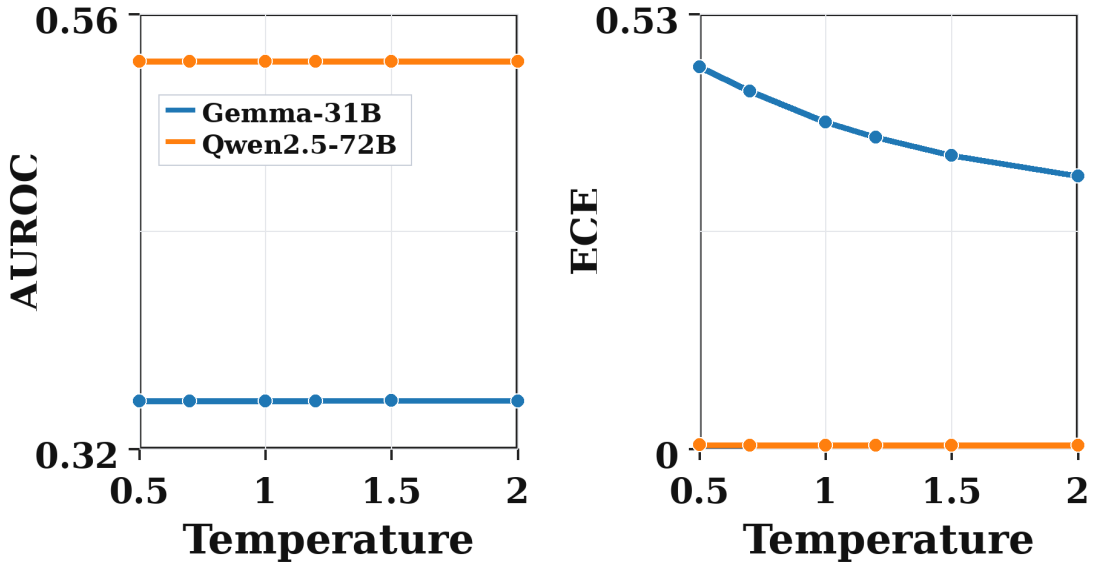

Figure A16: Post-hoc temperature sensitivity of logit-based calibration on Hepar2. Gemma-31B calibration improves substantially under higher-temperature softening, while AUROC remains nearly constant. This supports the project's finding that logit confidence can be recalibrated without changing causal-edge predictions.