Unveiling Causal Calibration: How LLM Scale Influences Statistical Information Interpolation

Markus Englberger M.ENGLBERGER@TUE.NL
Devendra Singh Dhami D.S.DHAMI@TUE.NL

Department of Mathematics and Computer Science, Eindhoven University of Technology

Editors: Bijan Mazaheri and Niels Richard Hansen

Abstract

There are two ways an LLM can learn statistical and causal knowledge during training: either by explicit statements about the causal scenario in the training corpus (e.g., "the probability of developing lung cancer is 20% if one has smoked for at least 10 years", "X and Y are independent conditioned on Z", "X is a cause of Y" etc.) or by implicit presentations of the same information via tabular data from surveys or scenario descriptions. In this paper, we probe whether the LLM internally unifies these two sources of information. To this end, we test whether fine-tuning LLMs on implicit presentations of statistical knowledge influences the LLM's answers when explicitly prompted for such knowledge. By creating in-context learning tasks where both explicit and implicit statistical information are present, we also test whether language models can unify these two modes during inference. Experiments suggest that larger language models have some shared method of storing these two sources of information, whereas smaller language models seem to be less capable of unifying these sources.

Keywords: large language models, knowledge alignment, explicit and implicit knowledge

1. Introduction

Large language models (LLMs) have become a part of our everyday life (Choi et al., 2024; Haque and Li, 2025) and their impact on several domains has been well documented (for example, nuclear medicine (Alberts et al., 2023), politics (Ryan et al., 2024) and industry (Maatouk et al., 2024)). In machine learning, papers on using and adapting LLMs have also seen a remarkable rise (Lu et al., 2022; Maini et al., 2024; Snell et al., 2025). They thus increasingly serve as general-purpose reasoning systems, capable of synthesizing information (Tao et al., 2025), generating explanations (Gat et al., 2024; Zhao et al., 2024), and offering guidance across scientific, technical, and social domains (Ge et al., 2023; Zhang et al., 2025).

Despite their growing capabilities, these models often struggle to reliably communicate what they implicitly "know". This issue is especially exacerbated in cases where causal reasoning is required, as LLMs have been shown to memorize (causal) information and then provide answers based on these memorized facts rather than by performing explicit reasoning (Zečević et al., 2024; Chi et al., 2024). On the contrary, some recent research has shown promise in this direction (Jin et al., 2023; Kiciman et al., 2023; Yu et al., 2025) and indicates that LLMs are capable of causal reasoning under some strict assumptions.

LLMs can acquire their knowledge during training in two primary ways. First, such knowledge may be encoded explicitly in the training corpus, for example, through direct statements of statistical and causal relationships and formal explanations (Yildirim and Paul, 2024). Second, the same

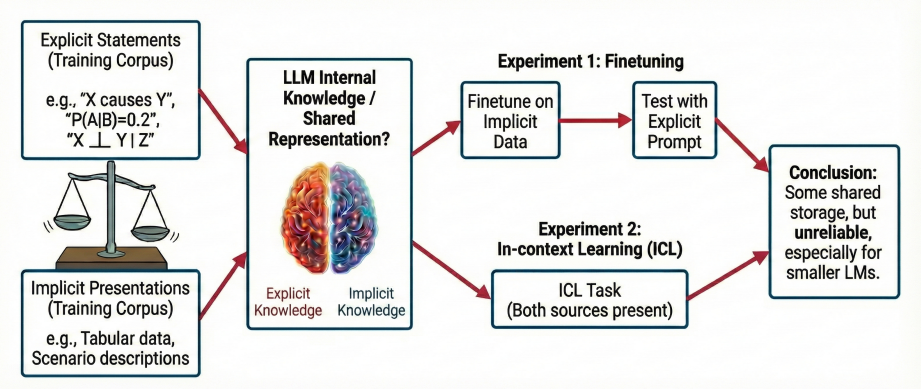

Figure 1: **Causal Calibration.** Pipeline of the experiments to determine the effect of scale on the explicit and implicit knowledge alignment of LLMs.

information can be provided in the form of tabular data that is sampled from the underlying domain distribution (Dong and Wang, 2024). This leads to the following critical questions:

1. *How do the LLMs actually learn the relationships between domain variables?*
2. *Does the scale of the LLM matter for learning these relationships?*

We propose the *Causal Calibration* property to answer the above set of questions. Causal calibration investigates how LLM scale influences the interpolation and internal representation of statistical information, with a focus on causally relevant patterns. This amounts to empirically ascertaining whether the LLMs learn the relationships between domain variables from explicit knowledge (statements regarding the data-generating process) or implicit knowledge (data generated from the process) by conducting two sets of experiments, namely fine-tuning and in-context learning. See Figure 1 for a detailed pipeline. In the first set of experiments, we fine-tune the LLM on implicit knowledge and test it with explicit knowledge, and in the second set of experiments, we create in-context learning tasks with both modes of information present in few-shot samples at inference.

By systematically probing models across a range of sizes within controlled statistical environments, we aim to clarify when and how LLMs begin to unify these two sources of knowledge. Causal calibration demonstrates the effect of the scale of the utilized LLMs on unifying these two sources of knowledge, as larger LLMs seem to be more capable of using a *joint* representation of the available sources of information, i.e., implicit and explicit.

We proceed as follows: first, we introduce the concept of causal calibration before highlighting our extensive empirical evaluation to demonstrate the effect of LLM scale on the extent of causal calibration. We then touch upon the related work before concluding.

2. Causal Calibration: Explicit vs. Implicit Knowledge

We differentiate between *explicit* and *implicit* statistical information about a data-generating process that may both be present in the training corpus. By *explicit* information, we mean declarative statements that directly describe properties of the data-generating process, such as "A causes B," "If A holds, then B holds with probability 60%," or "A is independent of B given C." By contrast,

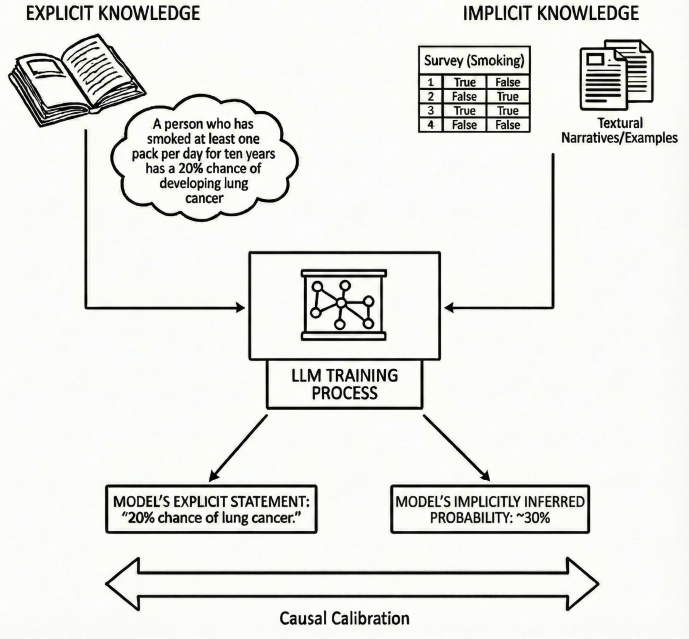

Figure 2: LLM training with Explicit and Implicit information

implicit statistical information consists of data or narratives generated by the process itself, for example, in the form of tabular datasets or scenario descriptions.

To illustrate, during training, an LLM may encounter an explicit statement such as "A person who has smoked at least one pack per day for ten years has a 20% chance of developing lung cancer." At the same time, it may also be exposed to implicit evidence, such as survey tables relating smoking habits to health outcomes or numerous textual accounts describing individuals who did or did not smoke and who did or did not develop lung cancer.

Crucially, these two sources of information differ not only in their perceived forms but also in the inferential demands they place on the model. Explicit statements provide direct access to aspects of the underlying data-generating process, often compressing the underlying causal structure into a single declarative sentence. Implicit information, by contrast, requires the model to infer such structure indirectly by aggregating patterns across many observations and contexts. Learning from implicit information, therefore, entails a form of statistical abstraction where the model must internally estimate potential (causal) relations without being explicitly told that such relations exist. Whether language models successfully unify these two sources of information into a coherent internal representation is not obvious; we refer to this property as *causal calibration*. In the lung cancer example, this would correspond to whether a model merely reproduces the learned risk when prompted, or whether it can infer comparable risk estimates from the tabular evidence alone (see Figure 2). Furthermore, the question is whether it can resolve inconsistencies between the two when they arise. Understanding how the LLM scale modulates this balance informs whether increasing model capacity leads to more coherent internal representations of data-generating processes or merely to improved memorization of statistical facts.

Causal calibration requires more than accurate marginal or conditional predictions; it necessitates consistency between a model's implicit statistical beliefs and its explicit verbalized judgments.

A model may, for instance, correctly predict outcomes sampled from a data-generating process while simultaneously endorsing incorrect causal statements about that same process. Such discrepancies suggest that high prediction results are insufficient evidence of structured causal understanding. Instead, causal calibration demands alignment between what the model does and what it says when queried about causal, probabilistic, or independence relations. This alignment is particularly challenging for LLMs, whose training objective optimizes next-token prediction rather than explicit causal inference, and whose internal representations may therefore encode statistical regularities in a form that is not trivially accessible for performing reasoning.

Model scale plays a central role in determining whether such alignment emerges. Increasing scale may enable richer internal abstractions that support both interpolation from implicit data and reconciliation with explicit statements, but it may also amplify spurious correlations or overconfident generalizations if calibration does not occur proportionately. To clarify, we conduct two sets of experiments:

1. We fine-tune LLMs on implicit data generated by a structural causal model (SCM) with two to four variables, and subsequently prompt them to explicitly reason about conditional probabilities, independence relations, and causal directions. We thus tackle the following questions: If we only fine-tune on implicit information regarding synthetic variables that the LLM has no prior knowledge about, can the LLM explicitly reason about that information? Can we modify an LLM's stance on whether random variables are independent by feeding it statistical data? Can implicit statistical data update an LLM's stance on whether a causal effect between two variables exists?

2. We construct in-context learning tasks in which the model is shown samples generated by an SCM with two variables A and B, and is asked to predict B given a value of A. Optionally, we prefix the prompt with an explicit statement describing the relationship between A and B. We then examine how this explicit information influences the model's predictions.

After postulating our questions, we are now ready to present our set of experiments, which can demonstrate the effect of LLM scale on causal calibration.

3. Experiments

To investigate how language model scale influences causal calibration, we design a set of controlled experiments that disentangle the roles of implicit statistical evidence and explicit causal statements.

3.1. Fine-tuning Experiments

We first present fine-tuning experiments in which an LLM is trained on implicit data generated by a structural causal model (SCM) with two to four binary variables, and is then prompted to explicitly reason about conditional probabilities, independence relations, and causal directions. Specifically, implicit data in this context refers to samples generated from a probability distribution present in the training data. The training data consist of prompts formatted as conditional prediction queries, such as: "Given that A is True, predict whether B is True. Answer only with True or False!". [1] For

1. To avoid semantic priors and help the model treat variables as abstract symbols, the variable names A, B, and C are replaced by randomly generated 4-5 letter words (e.g., `Ajkl`, `Bsdlf`, `Yont`) in all experiments.

each experiment, the training set contains samples drawn i.i.d. from the underlying data-generating process.

We evaluate the following models: Qwen2.5-1.5B-Instruct (Qwen et al., 2025), Gemma-2-9b & 27b-it (Team et al., 2024), Qwen3-32B (Yang et al., 2025), Meta-Llama-3-8B & 70B-Instruct (Grattafiori et al., 2024), and GPT-4.1 (OpenAI, 2025). [2] GPT-4.1 is accessed via the OpenAI API, while all other models are fine-tuned using LoRA adapters of rank 16 with 8-bit quantization. Models are trained with a batch size of 16 and a learning rate of 1×10^{-5} with AdamW (Loshchilov and Hutter, 2017) for 35 epochs; each dataset contains 400 samples. While including a broader range of models in the experiments would be desirable, many models refuse to answer the prompts included in the experiments, i.e., they decline to provide estimates of conditional probabilities or to assess independence statements, often citing insufficient information to do so.

At inference time, we probe both implicit and explicit knowledge. To assess *implicit* knowledge, we use the same binary prediction format as during training, for example: "Given that A is False, predict whether B is True. Answer only with True or False!". To assess *explicit* knowledge, we prompt the model for probability estimates, e.g.: "Given that A is True, state the probability that B is True. Answer only with the probability!".

To extract estimates of $P(B \mid A)$ from the model's implicit responses, we inspect the logits assigned to the tokens `True` and `False` and compute the normalized probability

$$P(B = \text{True} \mid A) = \frac{p_{\text{logit}}(\text{True})}{p_{\text{logit}}(\text{True}) + p_{\text{logit}}(\text{False})}. \tag{1}$$

For explicit probability prompts, we report the most frequently generated probability values along with their empirical frequencies, derived from the logits. For example, reporting $0.5(43\%), 1(21\%)$ indicates that the model outputs 0.5 in 43% of cases and 1 in 21% of cases.

In addition, we probe explicit judgments about independence and causal direction using prompts such as: "Are random variables A and B independent? Answer only with Yes or No!" and "Assuming causal sufficiency of variables A,B,C and faithfulness, does A have a causal effect on C?". We use the shorthand notation $P(B \mid \overline{A})$ to denote $P(B = \text{True} \mid A = \text{False})$ and consider three experimental configurations:

- **Experiment 1 (Two variables).** Two binary variables A and B with distribution

$$P(A) = 0.5, P(B|A) = 0.2, P(B|\overline{A}) = 0.6$$

- **Experiment 2 (Three variables).** Three binary variables A, B, and C with mechanisms

$$P(A) = P(B) = 0.5,$$

$$P(C \mid A, B) = 0.8, \quad P(C \mid A, \overline{B}) = 0.6, \quad P(C \mid \overline{A}, B) = 0.4, \quad P(C \mid \overline{A}, \overline{B}) = 0.2.$$

  Assuming causal sufficiency and faithfulness, this construction enforces the causal structure $A \to C$ and $B \to C$, allowing us to test whether the model can infer causal directions solely from implicit data. The prompts in the training set are divided equally between predicting B conditioned on A and predicting C conditioned on A and B.

2. specifically the `gpt-4.1-2025-04-14` snapshot.

	$P(B \mid A) = 0.2$		$P(B \mid \overline{A}) = 0.6$		
Model	Implicit	Explicit	Implicit	Explicit	Independent?
Qwen2.5-1.5B-Instr.	0.14	0.5 (87%) 1 (11%)	0.53	0.5 (96%)	No (100%)
Llama-3-8B-Instr.	0.15	0.5 (100%)	0.62	0.5 (99%)	No (100%)
Gemma-2-9b-it	0.24	0.5 (85%) 1 (7.0%)	0.65	0.5 (85%) 1 (11%)	Yes (86%)
Gemma-2-27b-it	0.22	0.5 (99%)	0.50	0.5 (60%) 0.2(32%)	Yes (94%)
Qwen3-32B	0.086	0.5 (94%) 0 (5.9%)	0.70	0 (94%)	No (100%)
Llama-3-70B-Instr.	0.31	0.5 (57%) 0.4 (14%)	0.68	0.5 (64%) 0.8 (11%)	Yes (100%)
GPT-4.1	0.002	0 (83%)	0.97	0.75 (48%) 1 (20%)	No (76%)

Table 1: **Fine-tuning experiment including two binary variables A, B with probabilities $P(A) = 0.5, P(B|A) = 0.2, P(B|\overline{A}) = 0.6$** . Models are queried on conditional probabilities $P(B|A)$ both implicitly (by letting the model predict B) and by explicitly querying the probability $P(B|A)$ and on whether A and B are independent. Regarding *explicit* columns, 0.5 (43%) means the model returns 0.5, 43% of the time.

- **Experiment 3 (Two variables).** Two independent binary variables A and B with

$$P(A) = 0.5, \quad P(B \mid A) = P(B \mid \overline{A}) = P(B) = 0.8.$$

Experiments on three additional data-generating processes can be found in the appendix Section A, one of which has four variables.

Results. The results for the three sets of experiments are shown in Table 1, Table 2, and Table 3, respectively. From the *implicit* columns, we observe that while models often fail to exactly recover the true underlying probabilities, they nonetheless capture the dominant statistical relationships in the data. For instance, in Experiment 1, all models correctly learn that B is more likely to be true when A is false than when A is true.

The *explicit* columns reveal that some models are capable of relating their explicitly stated conditional probabilities to their implicit predictions and that this capability appears to depend on model scale. For the two smallest models (Qwen2.5-1.5B-Instruct and Llama-3-8B-Instruct), the explicitly stated probabilities seem largely unrelated to the implicitly learned probabilities. For the slightly larger models (Gemma-2-9B-it, Gemma-2-27B-it, and Qwen3-32B), an emerging ability to relate the two modes can be observed, although this alignment remains inconsistent. *The larger models (Llama-3-70B-Instruct, and even more so GPT-4.1) exhibit a more consistent and closer alignment between implicit and explicit probability estimates*. GPT-4.1 shows this alignment consistently across all conditional probabilities. Llama-3-70B-Instruct also demonstrates consistent alignment,

(a) Models are queried on conditional probabilities $P(C|A,B)$.

	$P(C \mid A, B) = 0.8$		$P(C \mid A, \overline{B}) = 0.6$		$P(C \mid \overline{A}, B) = 0.4$		$P(C \mid \overline{A}, \overline{B}) = 0.2$	
Model	Impl.	Expl.	Impl.	Expl.	Impl.	Expl.	Impl.	Expl.
Qwen2.5-1.5B-Instr.	0.75	0.5 (100%)	0.73	0.5 (100%)	0.50	0.5 (99%)	0.27	0.5 (100%)
Llama-3-8B Instr.	0.59	0.5 (93%) 0.25 (5.8%)	0.59	0.5(100%)	0.41	0.5 (98%) 0.25 (1.2%)	0.29	0.5(100%)
Gemma-2-9b-it	0.64	0.6 (30%) 0.3 (18%) 0.8 (17%)	0.70	0.3 (61%) 0.6 (18%)	0.52	0.3 (54%) 0.6 (18%)	0.18	0.3 (33%) 0.2 (11%)
Gemma-2-27b-it	0.80	0.8 (94%) 0.6 (3.0%)	0.62	0.2 (96%)	0.41	0.2 (84%) 0.3 (7.8%)	0.13	0.2 (97%)
Qwen3-32B	0.83	0.5 (92%) 1 (8%)	0.82	0.6 (48%) 0.7 (45%)	0.7	0.8 (66%) 0.7 (34%)	0.09	0.5 (100%)
Llama-3-70B-Instr.	0.81	0.9 (97%) 0.7 (3%)	0.86	0.6 (28%) 0.9 (25%) 0.7 (23%)	0.54	0.7 (44%) 0.6 (37%)	0.18	0.5 (85%) 0.4 (8%)
GPT-4.1	0.73	1 (44%) 0.9 (24%) 0.8 (16%)	0.82	1 (32%) 0.8 (25%) 0.7 (17%)	0.56	0.7 (21%) 0.8 (21%) 0.6 (16%)	0.17	0 (42%) 0.2 (19%) 0.1 (13%)

(b) Models are queried on independence between variables.

Model	$A \perp B$	$A \perp C$	$B \perp C$
Qwen2.5-1.5B-Instr.	No (100%)	No (100%)	No (100%)
Llama-3-8B-Instr.	No (100%)	No (100%)	No (100%)
Gemma-2-9b-it	Yes (91%)	Yes (91%)	Yes (92%)
Gemma-2-27b-it	Yes (75%)	Yes (93%)	No (65%)
Qwen3-32B	No (100%)	No (100%)	No (100%)
Llama-3-70B-Instr.	No (63%)	No (63%)	No (64%)
GPT-4.1	No (75%)	No (71%)	No (79%)

(c) Models are queried on causal directions.

Model	$A \to B$	$A \to C$	$B \to A$	$B \to C$	$C \to A$	$C \to B$
Qwen2.5-1.5B-Instr.	No (100%)	No (100%)	No (100%)	No (100%)	No (100%)	No (100%)
Llama-3-8B-Instr.	No (100%)	No (100%)	No (100%)	No (100%)	No (100%)	No (100%)
Gemma-2-9b-it	Yes (74%)	Yes (71%)	Yes (75%)	Yes (72%)	Yes (62%)	Yes (55%)
Gemma-2-27b-it	No (98%)	No (99%)	No (97%)	No (100%)	No (98%)	No (100%)
Qwen3-32B	No (100%)	No (100%)	No (100%)	No (100%)	No (100%)	No (100%)
Llama-3-70B-Instr.	No (54%)	No (65%)	Yes (100%)	No (88%)	Yes (82%)	No (74%)
GPT-4.1	No (89%)	No (94%)	No (96%)	No (99%)	No (97%)	No (96%)

Table 2: **Fine-tuning experiment including three binary variables A, B, C with probabilities $P(A) = P(B) = 0.5, P(C|A,B) = 0.8, P(C|\overline{A},B) = 0.6, P(C|A,\overline{B}) = 0.4, P(C|\overline{A},\overline{B}) = 0.2$** . Subtable (a) shows the fine-tuned model's implicit and explicit beliefs about conditional probabilities, subtable (b) shows the model's beliefs about independence relations, and subtable (c) shows the model's beliefs on causal directions. In the *explicit* columns, 1 (30%) means the model returns 1, 30% of the time.

	$P(B \mid A) = 0.8$		$P(B \mid \overline{A}) = 0.8$		
Model	Implicit	Explicit	Implicit	Explicit	Independent?
Qwen2.5-1.5B-Instr.	0.72	0.5 (100%)	0.80	0.5 (100%)	No (100%)
Llama-3-8B-Instr.	0.79	0.5 (73%) 0 (18%)	0.85	0.5 (100%)	No (83%)
Gemma-2-9b-it	0.80	0.6 (30%) 0.8 (26%) 0.9 (12%)	0.79	0.6 (37%) 0.8 (28%)	Yes (96%)
Gemma-2-27b-it	0.88	0.8 (83%) 0.7(10%)	0.71	0.2 (38%) 0.7 (18%)	Yes (56%)
Qwen3-32B	1.0	0.7 (73%) 0.8 (27%)	1.0	0.5 (100%)	No (100%)
Llama-3-70B-Instr.	0.89	0.95 (96%) 0.8 (4.0%)	1.0	0.75 (28%) 0.7 (20%)	No (86%)
GPT-4.1	0.81	0.8 (42%) 0.6 (13%) 0.7 (12%)	0.73	0.8 (47%) 0.6 (21%)	No (62%)

Table 3: **Fine-tuning experiment including two binary variables $\boldsymbol{A, B}$ with probabilities $\boldsymbol{P(A) = 0.5, P(B \mid A) = P(B \mid \overline{A}) = 0.8}$.** Models are queried on conditional probabilities $P(B|A)$ both implicitly (by letting the model predict B) and by explicitly querying the probability $P(B|A)$ as well as on independence of A and B. In the *explicit* columns, 0.5 (73%) means model returns 0.5, 73% of the time.

although the gap between implicit and explicitly assigned probabilities is sometimes larger than for GPT-4.1. For example, the largest discrepancy between implicit and explicit probabilities for Llama-3-70B-Instruct is 32% (the estimate for $P(C|\overline{A}, \overline{B})$ in Table 2) and 27% for GPT-4.1 (the estimate for $P(C|A, B)$ in Table 2).

Finally, across all settings, none of the models reliably infer independence relations or correctly identify causal directions based solely on implicit data.

3.2. In-Context Learning Experiments

We further evaluate the ability of language models to integrate explicit and implicit statistical information during inference by constructing in-context learning tasks in which both forms of information are simultaneously present. The data for these experiments are generated from two to three binary random variables. Implicit data in this context refers to samples generated from a probability distribution present in the prompt.

Our prompting setup follows techniques similar to those proposed in Requeima et al. (2024) and Shysheya et al. (2025). Specifically, the model is provided with 10 samples of (A, B), one per line, where the final line contains only the value of A. The model is then expected to continue the sequence by predicting the corresponding value of B, effectively estimating $P(B \mid A)$ from the in-

Model	No Explicit Info	Info: $P(B\|\overline{A})$ low
Qwen2.5-1.5B	0.814	0.853
Mistral-7B v0.2	0.721	0.679
Llama-3-8B	0.673	0.578
Gemma-2-9b	0.786	0.728
Gemma-2-27b	0.696	0.521
Qwen3-32B	0.765	0.628
Llama-3-70B	0.886	0.754
Qwen-2.5-72B	0.839	0.733

Table 4: **ICL Experiment including binary random variables A and B with probabilities $(A) = 0, P(B|\overline{A}) = 0.8$.** Given are the predicted probabilities for B conditioned on A =False either with no explicit information given (first column) or with contradictory explicit information "If A is False, B has a very low chance of being True. Specifically, if A is False, B has a 10% chance of being True." (second column).

context samples. Each prompt is prefixed with the sentence: "The data contains samples of binary random variables A and B." Optionally, we prepend an additional *explicit* statement describing the properties of the data-generating process, such as: "If A is True, B holds with a probability of 70%." "If A is False, B has a very low probability." and "The random variables A and B are independent." Importantly, these explicit statements are chosen to conflict with the true data-generating process underlying the provided in-context samples. The goal is to assess how the model's predictions are influenced by explicit information when it contradicts the implicit statistical evidence. An example prompt is shown in Figure 3 in the appendix.

As in the fine-tuning experiments, we extract estimates of $P(B \mid A)$ from the logits of the tokens `True` and `False` (see Equation (1)). Each experiment is repeated 200 times, with variance arising solely from resampling the in-context examples (the prediction itself is deterministic, as probabilities are read directly from logits). We evaluate Qwen2.5-1.5B (Qwen et al., 2025), Mistral-7B-v0.2 (Jiang et al., 2023), Gemma-2-9B and Gemma-2-27B (Team et al., 2024), Qwen3-32B (Yang et al., 2025) and Llama-3-8B & 70B (Grattafiori et al., 2024). Again, including a broader range of models in the experiments would be desirable, but many models fail to continue the prompts as required; in particular, they do not respond with a simple `True` or `False` answer.

Experimental settings. We consider five data-generating processes, of which the first four include two variables (A and B) and the fifth contains three variables (A, B, and C):

- **Experiment 1 (Degenerate A).** $P(A) = 0$, $P(B \mid \overline{A}) = 0.9$. All samples satisfy $A =$ False, making this task simple.
- **Experiment 2 (Asymmetric dependence).** $P(A) = 0.5$, $P(B \mid A) = 0.1$, $P(B \mid \overline{A}) = 0.9$.
- **Experiment 3 (Asymmetric dependence).** $P(A) = 0.5$, $P(B \mid A) = 0.6$, $P(B \mid \overline{A}) = 0.3$.
- **Experiment 4 (Independence).** $P(A) = 0.5$, $P(B \mid A) = P(B \mid \overline{A}) = 0.8$.

	No expl. info		Info: $P(B\|\overline{A})$ low		Info: $P(B\|A)$ high		Info: $A \perp B$	
Model	$P(B\|A)$	$P(B\|\overline{A})$	$P(B\|A)$	$P(B\|\overline{A})$	$P(B\|A)$	$P(B\|\overline{A})$	$P(B\|A)$	$P(B\|\overline{A})$
Qwen2.5-1.5B	0.311	0.638	0.256	0.645	0.330	0.692	0.340	0.660
Mistral-7B-v0.2	0.253	0.701	0.259	0.695	0.394	0.649	0.276	0.648
Llama-3-8B	0.427	0.682	0.370	0.637	0.425	0.670	0.465	0.664
Gemma-2-9b	0.334	0.634	0.390	0.636	0.442	0.604	0.347	0.637
Gemma-2-27b	0.344	0.680	0.396	0.631	0.448	0.643	0.347	0.665
Qwen3-32B	0.271	0.719	0.245	0.640	0.554	0.721	0.361	0.638
Llama-3-70B	0.287	0.758	0.250	0.699	0.405	0.727	0.320	0.735
Qwen2.5-72B	0.314	0.736	0.235	0.672	0.454	0.723	0.297	0.742

Table 5: **ICL Experiment including binary random variables A and B with probabilities $P(A) = 0.5, P(B|A) = 0.1, P(B|\overline{A}) = 0.9$.** Given are the predicted probabilities for B conditioned on A either with no explicit information given (first column) or with contradictory explicit information "If A is False, B has a very low chance of being True. Specifically, if A is False, B has a 10% chance of being True." (second column), "If A is True, B has a very high chance of being True. Specifically, if A is True, B has a 90% chance of being True." (third column) or "A and B are independent random variables" (forth column).

	No explicit information		Info: $P(B\|\overline{A})$ low		Info: $A \perp B$	
Model	$P(B\|A)$	$P(B\|\overline{A})$	$P(B\|A)$	$P(B\|\overline{A})$	$P(B\|A)$	$P(B\|\overline{A})$
Qwen2.5-1.5B	0.493	0.468	0.470	0.438	0.493	0.473
Mistral-7B-v0.2	0.493	0.379	0.496	0.387	0.489	0.386
Llama-3-8B	0.536	0.494	0.497	0.351	0.534	0.514
Gemma-2-9b	0.515	0.426	0.515	0.437	0.522	0.431
Gemma-2-27b	0.516	0.461	0.547	0.452	0.507	0.464
Qwen3-32B	0.553	0.376	0.659	0.256	0.535	0.419
Llama-3-70B	0.496	0.434	0.494	0.363	0.487	0.449
Qwen2.5-3-72B	0.551	0.401	0.560	0.274	0.532	0.426

Table 6: **ICL Experiment including binary random variables A and B with probabilities $P(A) = 0.5, P(B|A) = 0.6, P(B|\overline{A}) = 0.3$.** Given are the predicted probabilities for B conditioned on A either when no explicit information is given (first column) or when contradictory explicit information "If A is False, B has a very low chance of being True. Specifically, if A is False, B has a 10% chance of being True." (second column) or "A and B are independent random variables" (third column) is given.

- **Experiment 5 (three variables).** $P(A) = P(B) = 0.5, P(C|A, B) = 0.9, P(C|A, \overline{B}) = 0.3, P(C|\overline{A}, B) = 0.6, P(C|\overline{A}, \overline{B}) = 0.0$:

The corresponding results are reported in Table 4, Table 5, Table 6, Table 7, and Table 8 respectively. For each experiment, we evaluate two variants: one in which no additional explicit information is provided, and one in which an explicit statement contradicting the true data-generating process is included. The specific statements used are listed in the table captions.

	No explicit information		Info: $P(B|\overline{A})$ high	
Model	$P(B|A)$	$P(B|\overline{A})$	$P(B|A)$	$P(B|\overline{A})$
Qwen2.5-1.5B	0.291	0.296	0.311	0.255
Mistral-7B-v0.2	0.289	0.256	0.389	0.252
Llama-3-8B	0.405	0.361	0.456	0.271
Gemma-2-9b	0.335	0.305	0.415	0.325
Gemma-2-27b	0.334	0.317	0.457	0.360
Qwen3-32B	0.287	0.202	0.574	0.194
Llama-3-70B	0.256	0.466	0.466	0.214
Qwen2.5-72B	0.313	0.245	0.456	0.142

Table 7: **ICL experiment including binary random variables A and B with probabilities $P(B|A) = P(B|\overline{A}) = 0.1$.** Given are the predicted probabilities for B conditioned on A either when no explicit information is given (first column) or when contradictory explicit information "If A is True, B has a very high chance of being True. Specifically, if A is True, B has a 90% chance of being True." (second column) is given.

Results. We observe that larger models (Gemma-2-27b, Qwen3-32B, Llama-3-70B, and Qwen2.5-72B) can consistently interpolate between explicit and implicit information at inference time. For example, in Experiment 1, when the present few-shot examples are sampled according to $P(B|\overline{A}) = 0.8$ (Table 4), these models reduce their estimated $P(B \mid \overline{A})$ by more than 10% when the prompt is prefixed with: "If A is False, B has a very low chance of being True. Specifically, if A is False, B has a 10% chance of being True.". In Experiment 4 (Table 7), when the few-shot examples follow the distribution $P(B|A) = 0.1$, these models raise their estimation for $P(B|A)$ by more than 10% when the prompt is prefixed with "If A is True, B has a very high chance of being True. Specifically, if A is True, B has a 90% chance of being True." Similarly, in Experiment 2 (Table 5), where the few-shot examples are sampled from the conditional distribution $P(B \mid A) = 0.1, P(B \mid \overline{A}) = 0.9$, the prediction probabilities for $P(B|A)$ and $P(B|\overline{A})$ are closer to one another when the in-context samples are prefixed with the statement "A and B are independent variables." (see the fourth column of Table 5 and the third column of Table 6). The same pattern appears in the three-variable setting (Table 8). Larger models again shift their estimates of $P(C \mid A, B)$ in the direction of the explicit statements, interpolating between explicitly and implicitly given information. Smaller models show weaker and less consistent shifts.

The smallest included model (Qwen2.5-1.5B) appears incapable of incorporating explicit information in its predictions, and the remaining mid-size models seem to be able to incorporate explicit information at times, however, this behavior is less consistent than in the larger models.

Overall, these results suggest that *larger models exhibit a greater degree of causal calibration at inference time*, blending explicit causal or statistical cues with implicit evidence from observed samples, whereas smaller models rely more rigidly on one source of information.

4. Related Work

There is relatively little prior work explicitly examining the misalignment between knowledge acquired from *implicit* statistical evidence and knowledge conveyed through *explicit* statements. Gu

Queried prob.	$P(C \mid A, B)$		$P(C \mid A, \overline{B})$		$P(C \mid \overline{A}, B)$		$P(C \mid \overline{A}, \overline{B})$	
Implicit prob.	0.9		0.3		0.6		0.0	
Expl. info	None	low pr.	None	high pr.	None	low pr.	None	high pr.
Qwen2.5-1.5B	0.65	0.61	0.43	0.43	0.55	0.51	0.24	0.23
Mistral-7B v0.3	0.64	0.65	0.40	0.50	0.49	0.48	0.28	0.26
Llama-3-8B	0.59	0.61	0.46	0.50	0.53	0.27	0.28	0.27
Gemma-2-9B	0.60	0.60	0.44	0.46	0.51	0.51	0.33	0.34
Gemma-2-27B	0.62	0.60	0.44	0.53	0.53	0.52	0.35	0.43
Qwen3-32B	0.75	0.49	0.37	0.82	0.57	0.36	0.15	0.64
Llama-3-70B	0.67	0.45	0.35	0.59	0.51	0.49	0.31	0.43
Qwen2.5-72B	0.73	0.55	0.45	0.60	0.59	0.43	0.23	0.40

Table 8: **ICL experiment with binary random variables A, B, and C.** The data-generating process is given by $P(A) = P(B) = 0.5$, $P(C \mid A, B) = 0.9$, $P(C \mid A, \overline{B}) = 0.3$, $P(C \mid \overline{A}, B) = 0.6$, and $P(C \mid \overline{A}, \overline{B}) = 0.0$. We report model-predicted probabilities for C conditioned on (A, B) under two settings: without explicit information (first column in each block) and with contradictory explicit statements (second column). The explicit statements respectively assert that $P(C \mid A, B)$ is low, $P(C \mid A, \overline{B})$ is high, $P(C \mid \overline{A}, B)$ is low, and $P(C \mid \overline{A}, \overline{B})$ is high.

et al. (2024) shows that probabilities explicitly stated by LLMs in response to direct probability queries often conflict with the probabilities implicitly encoded in the model's logits when asked to predict events. Notably, the implicit probabilities derived from logits are frequently better calibrated to real-world data than the explicitly stated probabilities, with this discrepancy being more pronounced in smaller models.

However, even in cases where a model's implicit and explicit probability estimates coincide, such observational results do not establish whether the model internally unifies these two forms of information. Apparent alignment may simply reflect consistency between the explicit statements and implicit statistical regularities present in the training corpus, rather than a genuinely coherent internal representation. Wang et al. (2024) demonstrates that LLMs struggle to identify and resolve conflicts when presented with information at inference time that contradicts their training data.

Misalignment between different sources of information has also been investigated by Zhang et al. (2024), who show that language models often fail to reconcile information provided in natural language with structured information produced through tool use. More broadly, there is ongoing research into how LLMs acquire their (potentially incorrect) knowledge (Betley et al., 2025) and what causes them to reproduce incorrect statements or to hallucinate (Kalai et al., 2025).

Large language models have also been proposed as tools for causal inference (Cai et al., 2024; Kiciman et al., 2023; Gao et al., 2023; Jin et al., 2024), for example, by directly querying models about causal directions. Understanding how LLMs arrive at such predictions is therefore crucial, particularly, if LLMs can learn such information only when it is explicitly stated in the training data or whether they can infer it from tabular data in the training corpus. Studies such as ours contribute to this goal. Our results suggest that while language models can infer certain relationships between random variables from implicit data alone, they often fail to recover causal directions when implicit data impose a specific underlying causal structure.

Requeima et al. (2024) demonstrated that LLMs can effectively condition on natural language descriptions to improve predictive performance in few-shot numerical regression tasks.

5. Conclusion

We investigated whether large language models internally unify implicit and explicit statistical knowledge, a property we refer to as *causal calibration*. Our experiments show that larger models are, to a limited extent, capable of explicitly reasoning about aspects of a data-generating process when trained solely on implicit data, that is, data generated from the process itself. In particular, fine-tuning on implicit samples can update an LLM's explicit beliefs about conditional probabilities. Furthermore, our in-context learning experiments demonstrate that larger LLMs are able to interpolate between contradictory explicit statements and implicit statistical evidence provided through few-shot samples. This suggests that, at inference time, these models can partially reconcile competing sources of information rather than relying exclusively on one. Despite these gains, substantial limitations remain. Even the largest models we evaluate fail to reliably update their internal beliefs about independence relations or to infer causal directions when such information is only implicitly available.

Probing the models on more difficult scenarios, such as more variables, more complex variable domains, and concrete values of interventional probability distributions, is interesting and is the immediate next step in our future work. However, we observe that the simple scenarios presented in our paper already prove quite challenging, even for the better performing larger models. In particular, all models struggle with predicting causal directions from implicit information, and probing for interventional probability distributions is a strictly more challenging task. Overall, our findings suggest that current LLMs exhibit only partial causal calibration, pointing to fundamental challenges in understanding and integrating statistical and causal information from data alone.

Future work includes developing training methods that encourage LLMs to internally unify implicit and explicit information about random variables, for example, by augmenting implicit data (such as tabular observations) with explicit statements describing the underlying data-generating process. Evaluating a broader range of models is also an immediate future step. However, as mentioned, selecting suitable candidate models is nontrivial, as some models refuse to answer prompts, such as those requesting concrete probabilities or requiring the continuation of in-context learning setups, as described in Sections 3.1 and 3.2.

Acknowledgments

The TU Eindhoven authors received support from their Department of Mathematics and Computer Science and the Eindhoven Artificial Intelligence Systems Institute. The authors also thank the support from the Dutch Research Council (NWO) via project NGF.1609.242.024.

References

Ian L Alberts, Lorenzo Mercolli, Thomas Pyka, George Prenosil, Kuangyu Shi, Axel Rominger, and Ali Afshar-Oromieh. Large language models (llm) and chatgpt: what will the impact on nuclear medicine be? *European journal of nuclear medicine and molecular imaging*, 50(6):1549–1552, 2023.

Jan Betley, Daniel Chee Hian Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 4043–4068, 2025.

Hengrui Cai, Shengjie Liu, and Rui Song. Is knowledge all large language models needed for causal reasoning?, 2024. URL https://arxiv.org/abs/2401.00139.

Haoang Chi, He Li, Wenjing Yang, Feng Liu, Long Lan, Xiaoguang Ren, Tongliang Liu, and Bo Han. Unveiling causal reasoning in large language models: Reality or mirage? *Advances in Neural Information Processing Systems*, 37:96640–96670, 2024.

Alexander Choi, Syeda Sabrina Akter, JP Singh, and Antonios Anastasopoulos. The llm effect: Are humans truly using llms, or are they being influenced by them instead? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22032–22054, 2024.

Haoyu Dong and Zhiruo Wang. Large language models for tabular data: Progresses and future directions. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2997–3000, 2024.

Jinglong Gao, Xiao Ding, Bing Qin, and Ting Liu. Is chatgpt a good causal reasoner? a comprehensive evaluation, 2023. URL https://arxiv.org/abs/2305.07375.

Yair Gat, Nitay Calderon, Amir Feder, Alexander Chapanin, Amit Sharma, and Roi Reichart. Faithful explanations of black-box nlp models using llm-generated counterfactuals. *The Twelfth International Conference on Learning Representations*, 2024.

Yingqiang Ge, Wenyue Hua, Kai Mei, Juntao Tan, Shuyuan Xu, Zelong Li, Yongfeng Zhang, et al. Openagi: When llm meets domain experts. *Advances in Neural Information Processing Systems*, 36:5539–5568, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, and Anirudh Goyal. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Bowen Gu, Rishi J. Desai, Kueiyu Joshua Lin, and Jie Yang. Probabilistic medical predictions of large language models, 2024. URL https://arxiv.org/abs/2408.11316.

Md Asraful Haque and Shuai Li. Exploring chatgpt and its impact on society. *AI and Ethics*, 5(2): 791–803, 2025.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, et al. Cladder: Assessing causal reasoning in language models. *Advances in Neural Information Processing Systems*, 36: 31038–31065, 2023.

Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. Can large language models infer causation from correlation?, 2024. URL https://arxiv.org/abs/2306.05836.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models hallucinate, 2025. URL https://arxiv.org/abs/2509.04664.

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1711.05101.

Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5206–5215, 2022.

Ali Maatouk, Nicola Piovesan, Fadhel Ayed, Antonio De Domenico, and Merouane Debbah. Large language models for telecom: Forthcoming impact on the industry. *IEEE communications magazine*, 2024.

Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. Llm dataset inference: Did you train on my dataset? *Advances in Neural Information Processing Systems*, 37:124069–124092, 2024.

OpenAI. Gpt-4.1 model documentation, 2025. URL https://platform.openai.com/docs/models/gpt-4.1.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

James Requeima, John Bronskill, Dami Choi, Richard E. Turner, and David Duvenaud. Llm processes: Numerical predictive distributions conditioned on natural language. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 109609–109671. Curran Associates, Inc., 2024. doi: 10.52202/079017-3479. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/c5ec22711f3a4a2f4a0a8ffd92167190-Paper-Conference.pdf.

Michael J Ryan, William Held, and Diyi Yang. Unintended impacts of llm alignment on global representation. *arXiv preprint arXiv:2402.15018*, 2024.

Aliaksandra Shysheya, John Bronskill, James Requeima, Shoaib Ahmed Siddiqui, Javier Gonzalez, David Duvenaud, and Richard E. Turner. Jolt: Joint probabilistic predictions on tabular data using llms, 2025. URL https://arxiv.org/abs/2502.11877.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Zhengwei Tao, Jialong Wu, Wenbiao Yin, Junkai Zhang, Baixuan Li, Haiyang Shen, Kuan Li, Liwen Zhang, Xinyu Wang, Yong Jiang, et al. Webshaper: Agentically data synthesizing via information-seeking formalization. *arXiv preprint arXiv:2507.15061*, 2025.

Gemma Team, Morgane Riviere, Shreya Pathak, et al. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. Resolving knowledge conflicts in large language models, 2024. URL https://arxiv.org/abs/2310.00935.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, and Dayiheng Liu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Ilker Yildirim and LA Paul. From task structures to world models: what do llms know? *Trends in Cognitive Sciences*, 28(5):404–415, 2024.

Longxuan Yu, Delin Chen, Siheng Xiong, Qingyang Wu, Dawei Li, Zhikai Chen, Xiaoze Liu, and Liangming Pan. Causaleval: Towards better causal reasoning in language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12512–12540, 2025.

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. Causal parrots: Large language models may talk causality but are not causal. *Transactions on Machine Learning Research*, 2024.

Shuo Zhang, Liangming Pan, Junzhou Zhao, and William Yang Wang. The knowledge alignment problem: Bridging human and external knowledge for large language models, 2024. URL https://arxiv.org/abs/2305.13669.

Yutong Zhang, Lixing Chen, Shenghong Li, Nan Cao, Yang Shi, Jiaxin Ding, Zhe Qu, Pan Zhou, and Yang Bai. Way to specialist: Closing loop between specialized llm and evolving domain knowledge graph. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 1996–2007, 2025.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38, 2024.

Appendix A. Additional Fine-Tuning Experiments

In this appendix, we report additional fine-tuning experiments that complement the results in Section 3.1. These experiments follow the same setup but consider alternative parameterizations of the data-generating process and an increased number of variables.

For each experiment, we report both *implicit* probabilities (obtained from model logits) and *explicit* probabilities (obtained via direct probability queries). When applicable, we also report model judgments on independence relations.

A.1. Two-variable setting (asymmetric distribution)

We first consider two binary variables A and B with

$$P(A) = 0.4, \quad P(B \mid A) = 0.1, \quad P(B \mid \overline{A}) = 0.9.$$

	$P(B \mid A) = 0.1$		$P(B \mid \overline{A}) = 0.9$		
Model	Impl.	Expl.	Impl.	Expl.	$A \perp B$
Qwen2.5-1.5B-Instr.	0.0	0.5 (100%)	0.83	0.5 (100%)	No (100%)
Llama-3-8B-Instr.	0.0	0 (81%)	1.0	0.5 (94%)	No (100%)
Gemma-2-9b-it	0.06	0.5 (62%)	0.9	0.6 (96%)	Yes (70%)
Gemma-2-27b-it	0.09	0.3 (45%)	0.65	0.5 (84%)	No (89%)
Qwen3-32B	0.0	0.5 (100%)	0.78	0.5 (100%)	No (100%)
Llama-3-70B-Instr.	0.06	0 (70%)	0.98	0.8 (91%)	No (100%)
GPT-4.1	0.05	0 (82%)	0.94	1 (99%)	No (86%)

Table 9: Fine-tuning results for a two-variable asymmetric distribution.

The results are shown in Table 9.

A.2. Three-variable setting

We next consider three binary variables A, B, C with

$$P(A) = 0.6, \quad P(B \mid A) = 0.3, \quad P(B \mid \overline{A}) = 0.6,$$

$$P(C \mid A, B) = 0.1, \quad P(C \mid A, \overline{B}) = 1.0, \quad P(C \mid \overline{A}, B) = 0.3, \quad P(C \mid \overline{A}, \overline{B}) = 0.5.$$

The results are shown in Table 10.

A.3. Four-variable setting

Finally, we consider four binary variables A, B, C, D with

$$P(A) = P(B) = P(C) = 0.5,$$

and conditional probabilities defined over $P(D \mid A, B, C)$ and $P(C \mid A, B, C)$.

Due to space constraints, we report the results in two parts (Table 11 and Table 12).

Model	$P(C \mid A, B = 0.1$ Impl.	Expl.	$P(C \mid A, \overline{B}) = 1.0$ Impl.	Expl.	$P(C \mid \overline{A}, B) = 0.3$ Impl.	Expl.	$P(C \mid \overline{A}, \overline{B}) = 0.5$ Impl.	Expl.
Qwen2.5-1.5B-Instr.	0.0	0.5	1.0	0.5	0.21	0.5	0.41	0.5
Llama-3-8B-Instr.	0.0	0.5	1.0	0.5	0.23	0.5	0.70	0.5
Gemma-2-9b-it	0.12	0.3	0.99	0.6	0.21	0.6	0.71	0.6
Gemma-2-27b-it	0.11	0.2	0.99	0.6	0.25	0.6	0.62	0.8
Qwen3-32B	0.0	0.5	1.0	0.7	0.24	0.5	0.39	0.5
Llama-3-70B-Instr.	0.19	0.2	0.89	0.8	0.23	0.1	0.59	0.7
GPT-4.1	0.01	0.25	1.0	1.0	0.27	0.25	0.41	0.5

Table 10: Fine-tuning results for three variables.

Model	$P(D \mid A, B, C) = 0.1$ Impl.	Expl.	$P(C \mid A, B, \overline{C}) = 0.8$ Impl.	Expl.	$P(D \mid A, \overline{B}, C) = 0.4$ Impl.	Expl.	$P(D \mid A, \overline{B}, \overline{C}) = 0.3$ Impl.	Expl.
Qwen2.5-1.5B-Instr.	0.0	0.5(97%)	1.0	0.5(100%)	0.5	0.5(96%)	0.26	0.5(98%)
Llama-3-8B-Instr.	0.0	0.5(74%)	0.85	0.5(71%)	0.19	0.5(69%)	0.16	0.5(65%)
Gemma-2-9b-it	0.10	0.6(57%)	0.82	0.6(51%)	0.43	0.5(62%)	0.38	0.6(55%)
Gemma-2-27b-it	0.05	0.2(68%)	0.57	0.5(92%)	0.34	0.5(93%)	0.15	0.5(96%)
Qwen3-32B	0.0	0(46%)	0.77	0.5(66%)	0.23	0.5(82%)	0.16	0.2(52%)
Llama-3-70B-Instr.	0.0	0(100%)	0.70	0.8(63%)	0.33	0.5(93%)	0.19	0.2(93%)
GPT-4.1	0.01	0(98%)	0.85	1(85%)	0.41	0.3(76%)	0.38	0.5(43%)

Table 11: Fine-tuning results for four variables (part 1).

Discussion. The additional experiments presented in this appendix exhibit trends consistent with those reported in the main text. While overall performance slightly degrades in the more complex settings including 4 variables,larger models (e.g., Llama-3-70B-Instruct and GPT-4.1) continue to show evidence of partial unification between implicit and explicit statistical information. Their explicit probability estimates are more closely aligned with their implicit predictions compared to smaller models. In contrast, smaller and mid-sized models display less stable behavior, with explicit predictions often remaining weakly related or unrelated to the implicitly learned distributions.

Model	$P(D \mid \overline{A}, B, C) = 0.8$		$P(C \mid \overline{A}, B, \overline{C}) = 0.5$		$P(D \mid \overline{A}, \overline{B}, C) = 1.0$		$P(D \mid \overline{A}, \overline{B}, \overline{C}) = 0.2$	
	Impl.	Expl.	Impl.	Expl.	Impl.	Expl.	Impl.	Expl.
Qwen2.5-1.5B-Instr.	1.0	0.5(98%)	0.74	0.5(100%)	1.0	0.5(98%)	0.0	0.5(96%)
Llama-3-8B-Instr.	0.63	0.5(77%)	0.56	0.5(70%)	1.0	0.5(78%)	0.35	0.5(81%)
Gemma-2-9b-it	0.62	0.5(62%)	0.50	0.3(56%)	0.98	0.5(48%)	0.20	0.6(60%)
Gemma-2-27b-it	0.56	0.5(91%)	0.35	0.4(59%)	0.84	0.4(84%)	0.1	0.5(98%)
Qwen3-32B	0.77	0.5(94%)	0.40	0.5(72%)	1.0	0.7(47%)	0.11	0.5(96%)
Llama-3-70B-Instr.	0.85	0.9(86%)	0.26	0.4(96%)	0.89	0.9(73%)	0.0	0.5(81%)
GPT-4.1	0.78	0.75(49%)	0.59	0.5(61%)	1.0	0.8(52%)	0.18	0.4(49%)

Table 12: Fine-tuning results for four variables (part 2).

Appendix B. Figures

Example of a prompt used in the ICL experiments

```
The data contains samples of binary random variables A and B.
If A is True, B has a very high chance of being True as well.
Specifically, if A is True, B has a 90% chance of being True.

A: True; B: False
A: False; B: True
A: True; B: False
A: True; B: False
A: False; B: True
A: True; B: False
A: False; B: True
A: True; B: False
A: True; B: False
A: True; B: False
A: False; B:
```

Figure 3: Example of a prompt used in the ICL experiments.