# Always Tell Me The Odds: Fine-grained Conditional Probability Estimation

**Liaoyaqi Wang*** **Zhengping Jiang*** **Anqi Liu** **Benjamin Van Durme**

Johns Hopkins University

{lwang240,zjiang31,aliu74,bvandur1}@jh.edu

## Abstract

We present a state-of-the-art model for fine-grained probability estimation of propositions conditioned on context. Recent advances in large language models (LLMs) have significantly enhanced their reasoning capabilities, particularly on well-defined tasks with complete information. However, LLMs continue to struggle with making accurate and well-calibrated probabilistic predictions under uncertainty or partial information. While incorporating uncertainty into model predictions often boosts performance, obtaining reliable estimates of that uncertainty remains understudied. In particular, LLM probability estimates tend to be coarse and biased towards more frequent numbers. Through a combination of human and synthetic data creation and assessment, scaling to larger models, and better supervision, we propose a set of strong and precise probability estimation models. We conduct systematic evaluations across tasks that rely on conditional probability estimation and show that our approach consistently outperforms existing fine-tuned and prompting-based methods by a large margin [1][2].

## 1 Introduction

While large language models (LLMs) have shown remarkable performance in a wide range of well-formulated, clearly solvable reasoning tasks, real-world applications often require them to operate under uncertainty and ambiguity, where purely deductive or deterministic reasoning may fail (McCarthy & Hayes, 1981). Much of real-world and commonsense knowledge is inherently probabilistic (Li et al., 2021; Moss, 2018; Glickman et al., 2005). Thus, it is crucial for LLMs to integrate evidence with prior knowledge and to evaluate the plausibility of new information (Jaynes, 2003). We want LLMs that always (and accurately) tell us the odds.[3]

Various approaches incorporate uncertainty into the LLM reasoning process. For example, a recent line of work proposes to create structured reasoning traces that allow uncertainty to propagate (Jung et al., 2022; Feng et al., 2024; Xia et al., 2024; Hou et al., 2024b; Akyürek et al., 2024; Sanders & Durme, 2025). Further, it has been argued that uncertainty-aware evaluation provides better insights into the model capabilities (Cheng et al., 2024; Jiang et al., 2024b; Yuan et al., 2024). However, these methods typically rely on estimating conditional probabilities over local structures using prompting strategies that are often underexplored and poorly optimized, effectively delegating this task to the LLM without sufficient scrutiny. Approaches typically involve examining the logits of the model (Zhao et al., 2021b), which are known to be miscalibrated (Jiang et al., 2021; Xiong et al., 2024; Jiang et al., 2022), or using verbalized confidence (Mielke et al., 2022; Tian et al., 2023) which tends to stick to a few common discrete values (Razeghi et al., 2022; Cruz et al., 2024; Feng et al., 2024).

*Equal contribution.

[1] https://github.com/zipJiang/decoding-based-regression/tree/main

[2] https://huggingface.co/collections/Zhengping/always-tell-me-the-odds-6806b1e01cb76d8c7f3a33ef

[3] https://star-wars-memes.fandom.com/wiki/Never_tell_me_the_odds

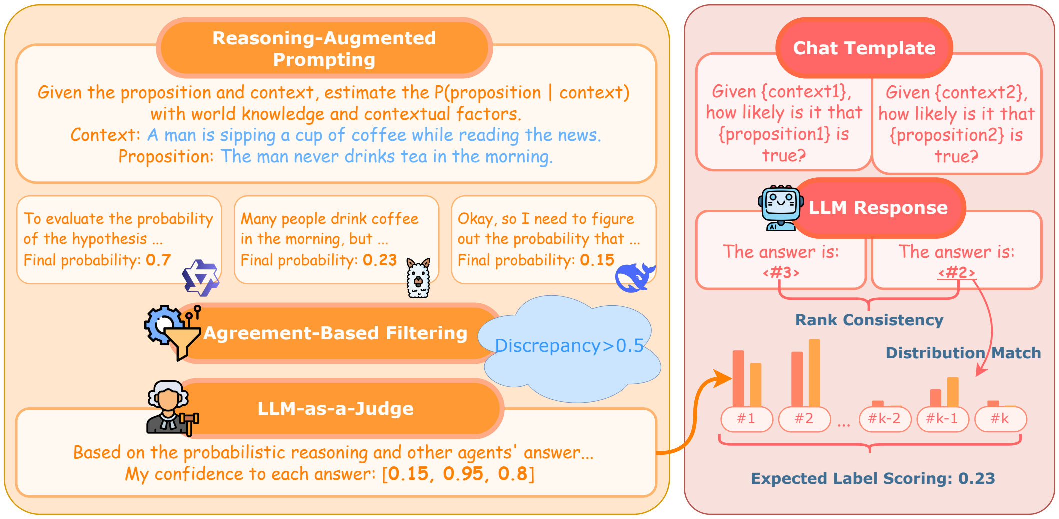

Reasoning-Augmented Prompting
Given the proposition and context, estimate the P(proposition | context) with world knowledge and contextual factors.
Context: A man is sipping a cup of coffee while reading the news.
Proposition: The man never drinks tea in the morning.
To evaluate the probability of the hypothesis ... Final probability: 0.7
Many people drink coffee in the morning, but ... Final probability: 0.23
Okay, so I need to figure out the probability that ... Final probability: 0.15
Agreement-Based Filtering
Discrepancy>0.5
LLM-as-a-Judge
Based on the probabilistic reasoning and other agents' answer... My confidence to each answer: [0.15, 0.95, 0.8]
Chat Template
Given {context1}, how likely is it that {proposition1} is true?
Given {context2}, how likely is it that {proposition2} is true?
LLM Response
The answer is: <#3>
The answer is: <#2>
Rank Consistency
Distribution Match
#1
#2
...
#k-2
#k-1
#k
Expected Label Scoring: 0.23

Figure 1: We train decoder-based models for fine-grained probability estimation, going beyond human annotations. We rely on two sources of supervision: synthetic (left), and pairwise ranking (right). For synthetic data, we first collect multiple LLMs' probability estimates with reasoning, then group instances by LLM agreement. For those with significant discrepancies, we submit them to another LLM judge, which rate the quality of each reasoning process. These ratings are then used to aggregate the probability estimates into a distribution over possible bins. For pairwise ranking, we use a margin loss to ensure consistency between the pairwise labels and the fine-grained probability estimates – which are computed via the expected label scoring rule.

Driven by the necessity of accurate, fine-grained probabilistic reasoning, in this work we focus on building strong, LLM-based models for estimating the probability of a textual proposition given a context. Direct evaluation against human annotations can be subjective and noisy; therefore, we first construct an objective and comprehensive evaluation suite by designing subproblems from existing datasets and frameworks that admit an intuitive probabilistic interpretation.

We therefore propose to train such conditional probability estimation models with (1) a modern LLM backend, (2) much more synthetic data, and (3) diverse training objectives. While most existing approaches rely on regression with smaller encoder-based models to fit human annotation (Chen et al., 2020; Nie et al., 2020b), by leveraging an expected label scoring rule that bridges regression / ranking and calibrated classification (Jiang et al., 2024a), we propose simple techniques that can provide fine-grained probability estimates with a modern, large decoder-based model. This allows us to better utilize the exceptional language understanding capabilities of pretrained LLMs. Also, our formulation allows natural integration with ensemble distillation using synthetic data as well as training with pairwise likelihood comparison. Due to the challenges in getting high quality data described above, these techniques are crucial to further improve performance.

We also propose a suite of tasks to comprehensively evaluate the performance of such models, ranging from their alignment with human and synthetic annotations, to the consistency of their plausibility rankings with human perception, and their ability to support Bayesian inference and decision-making. Our best model surpasses existing fine-tuned and prompting-based methods across multiple tasks, revealing untapped potential in improving probabilistic reasoning in LLMs through better probability estimation.

## 2 Related Work

**Uncertainty in LLMs** Like other modern neural networks (Guo et al., 2017), transformer-based language models are often overconfident in their predictions (Desai & Durrett, 2020;

Jiang et al., 2022). This makes LLM bad confidence estimator, while instruction tuning has been reported to further concentrating probability mass (Achiam et al., 2023; Hendrycks et al., 2020; Padmakumar & He, 2023). To address this, various methods have been proposed to improve uncertainty estimation, such as analyzing logits (Zhao et al., 2021b), prompting LLMs to verbalize confidence (Mielke et al., 2022; Tian et al., 2023), or probing internal representations (CH-Wang et al., 2024). Of these, verbalized uncertainty has generally been shown to produce the most accurate probability estimation (Tian et al., 2023), though it can still be overconfident (Tanneru et al., 2023; Xiong et al., 2024; Chen et al., 2023) and often relies on a small set of common discrete outputs (e.g., 10%, 75%) (Cruz et al., 2024).

Part of the problem is that existing resources for training are predominantly hard labeled, making model calibration and its evaluation particularly challenging (Kumar et al., 2019; Nixon et al., 2019). Recently, the field has seen growing interest in modeling the complete label distribution (Meissner et al., 2021; Liu et al., 2023; Zhou et al., 2021; Cheng et al., 2024), and different protocols have been proposed for curating datasets to enable such training and evaluation (Nie et al., 2020b; Chen et al., 2020). However, these datasets tend to be limited in size and are subject to ongoing debate over whether the distributions reflect true aleatoric uncertainty or merely label noise (Jiang & Marneffe, 2022; Baan et al., 2022; Wang et al., 2022; Weber-Genzel et al., 2024; Baan et al., 2024).

Thus, to mitigate the limitations of existing resources, we propose training our model with synthetic data and further supervising it with pairwise ranking supervision.

**LLM-based Probabilistic Reasoning** Chain-of-thought prompts effectively elicit reasoning in LLMs (Wei et al., 2022), but do not ensure accurate or calibrated probability estimates in real-world settings (Paruchuri et al., 2024). To address this, various methods integrate probability into reasoning. Ozturkler et al. (2023) proposed a two-stage framework where LLMs are queried in parallel and aggregated via likelihoods. Feng et al. (2024) and Nafar et al. (2024) align Bayesian Networks with LLM abductions for decision-making and QA tasks. Others decompose statements into sub-claims and aggregate them using confidence scores (Xia et al., 2024; Hou et al., 2024b; Cao et al., 2023; Jung et al., 2022; Akyürek et al., 2024). While LLMs can be tuned to mimic Bayesian models (Qiu et al., 2025), such supervision is rarely feasible in practice, limiting prior work to toy examples (Wong et al., 2023). Most existing methods solely rely on aggregating local probability estimates via prompting or probing (Jung et al., 2022), which can be suboptimal as discussed above. We show that LLMs can be directly tuned to produce fine-grained, accurate conditional probabilities for structured reasoning.

**Decoder-based Regression** Although primarily trained to predict the next token in textual sequences, large language models (LLMs) have been shown to perform effectively as regression models. Vacareanu et al. (2024) demonstrate that LLMs can conduct both linear and non-linear regression in context. In a series of papers, Song et al. (2024) and Song & Bahri (2025) develop techniques for training LLMs to perform numerical regression using only textual representations, achieving arbitrary precision. In this work, to balance efficiency and usability, we propose a hybrid approach that combines coarse-grained probability estimates via textual representations with fine-grained probability refinement through expected label scoring rule aggregation (Jiang et al., 2024a).

## 3 Methodology

We aim to develop accurate, fine-grained, and broadly applicable models for estimating the conditional probability $P(\text{proposition} \mid \text{context})$, representing the likelihood that a candidate proposition is true or occurs given a textual context. Subsection 3.1 outlines our evaluation suite, which covers intrinsic alignment with labels, consistency of probability rankings, and support for structured reasoning, beyond straightforwardly comparing predictions to noisy human annotations. Subsection 3.2 presents our strategy for scaling up training data using LLM-based pseudo-labeling, despite known inaccuracies and biases in LLM probability estimates. Finally, subsection 3.3 describes our training procedure for fine-tuning LLM-based models using a combination of human annotations, synthetic probability estimates, and pairwise ranking consistency signals.

### 3.1 Objective and Evaluation

What makes a good model for conditional probability estimation? When accurate probabilistic labels are available – e.g., from census data or controlled Bayesian setups – evaluation is straightforward. However, such labels are rare in real-world settings, where commonsense reasoning and complex patterns are involved. Human-provided estimates can help (Chen et al., 2020; Nie et al., 2020b) but are often subjective and noisy (Meissner et al., 2021).

Instead of relying solely on ground-truth alignment, we advocate evaluating models based on their effectiveness in real-world decision-making, where probabilistic reasoning aids belief modeling and evidence integration (Feng et al., 2024; Xia et al., 2024; Qiu et al., 2024). Inspired by multi-class calibration (Zhao et al., 2021a), we hypothesize that better probability estimates should support more human-aligned decisions.

To this end, we propose three task categories: (1) **Intrinsic**, comparing model estimates with human or LLM labels; (2) **Comparison**, ranking plausibility among probability estimates; and (3) **Structural**, assessing uncertainty propagation and decision-making in structured reasoning (subsection 4.1).

### 3.2 Synthetic Data Creation

This section introduces our method for generating pseudo-probabilistic labels for propositions given textual context, aiming to expand domain and distribution coverage for conditional probability estimation.

Inspired by the fact that increased inference-time compute improves both performance and confidence (Kojima et al., 2022; Muennighoff et al., 2025; Jurayj et al., 2025), we enhance annotation quality by generating multiple LLM roll-outs for each probability estimation (Wang et al., 2023). Prior work highlights a discrepancy between LLMs' generative and evaluative capabilities (Gu et al., 2024; Zelikman et al., 2022; Kumar et al., 2024), with models excelling at tasks like judging or verifying outputs. We leverage this by proposing a multi-step annotation aggregation process, where probability estimates are judged based on reasoning quality by another LLM – yielding consistent performance gains. We explored various ways of improving the pseudo-label quality, including one similar to the confidence ranking approach proposed by Shrivastava et al. (2025) with LLM-based pairwise comparison, which we detailed in Appendix A. Overall, we find the approach adopted here scales the most efficiently with compute.

**Reasoning-Augmented Prompting** The first step in our annotation process involves eliciting direct probability estimates from LLMs alongside their corresponding reasoning chains. Our prompting strategy instructs LLMs to leverage world knowledge to assess contextual factors before arriving at a final estimation, as incorporating reasoning helps models better approximate human decision-making (Chen et al., 2024). This approach encourages models to decompose ambiguous premises into plausible real-world scenarios and estimate the likelihood of each scenario's occurrence (Hou et al., 2024a). Intuitively, multiple rounds of annotation using different models and configurations can provide a richer understanding of each data point. We enforce reasoning chains for two main reasons: (1) they tend to enhance performance, and (2) they improve interpretability, enabling more rigorous evaluation and verification (Lightman et al., 2023), while also serving as useful input for subsequent steps.[4]

**Agreement-Based Filtering** After obtaining raw probability scores by directly prompting LLMs, we seek to identify low-quality estimates for further refinement. We observe that estimates with high agreement are more aligned with human annotations in the UNLI

[4] Please refer to subsection D.1 for the exact prompts we use.

validation set (Chen et al., 2020). This is supported by prior findings that ensemble agreement can be a reliable proxy for label quality (Deng et al., 2023; Baek et al., 2022), and that targeting uncertain labels can improve annotation quality (Gligorić et al., 2024). To quantify confidence, we define *discrepancy* as the difference between the maximum and minimum probability estimates across models. For low-discrepancy samples, we retain the original scores; high-discrepancy samples are flagged for further review.

**LLM-as-a-Judge** In the final step, we employ a large language model (LLM) as a judge to adjudicate among candidate reasoning chains (Zheng et al., 2023). Specifically, the LLM evaluates the quality of each reasoning process that leads to a conditional probability estimate (Du et al., 2024; Chiang & yi Lee, 2023). For reasoning traces produced by reinforcement learning-enhanced models, we first summarize the rationale to extract key steps influencing the probability assessment before presenting it to the judge. This preprocessing step is necessary because such reasoning traces are often lengthy and exhibit complex reasoning patterns (DeepSeek-AI et al., 2025), which may confuse the LLM judge. After analyzing the context, outcome, and candidate reasoning chains, the judge assigns a confidence score between 0 and 1 to reflect the reliability of each chain (Xu et al., 2023). These confidence scores are then used as supervision signals during model fine-tuning, as discussed below.

### 3.3 Model Fine-tuning

While there already exists a pretraining-based model for subjective probabilities (Chen et al., 2020), they are typically small-encoder-based and tuned naively to match human scalar annotation. In this section, we describe our training recipe for leveraging modern large-scale decoder-based language models, incorporating calibrated human annotations, utilizing synthetic labels for domain generalization, and integrating pairwise ranking consistency signals to enhance model performance.

**Decoder-based Regression** refers to the practice of training decoder-only language models to perform regression by outputting textual representations of numeric values (Song & Bahri, 2025). While LLM embeddings have been used as features for regression tasks, prior studies report unclear scaling trends (Tang et al., 2024), as these embeddings mostly capture semantic similarity rather than supporting precise numerical prediction (Devlin et al., 2019; Li et al., 2020). Besides, we prefer textual-based outputs as users will be able to directly get the outputs from common LLM services without additional processing (Kwon et al., 2023).

Specifically, we split the interval $[0, 1]$ into $N$ bins of equal width $\{b_0, b_1, \ldots, b_{n-1}\}$, and assign each bin $b_j$ a unique special token $t_j$. For an instance $(x_i, y_i) \in \mathcal{D}$ (where $\mathcal{D}$ denotes the dataset), we tune the decoder-only model to predict the token $t_j$ corresponding to the bin $b_j$ that contains the target probability $y_i$. However, this conversion $y \mapsto t$ is lossy, resulting in a coarse prediction. To recover fine-grained scalar prediction, we consider taking the expectation over all possible token predictions. Suppose we have a scoring function $f$ that converts each bin into a scalar $f := \mathcal{B} \to \mathbb{R}^+$, we construct our fine prediction as

$$\hat{y} = \sum_{j=0}^{N-1} f(b_j) \cdot p(t_j \mid x_i).$$

This is called the expected label scoring rule, introduced by Jiang et al. (2024a), shows that both the mean absolute error and the ranking risk regret lower-bound the calibration error of the underlying classifier. In our case, this implies that the better the model approximates the true token distribution, the more accurately it captures the underlying conditional probability. Inspired by their work, we convert the ground truth $y_i$ to a Gaussian distribution centered at $y_i$ with a small fixed variance $\sigma^2$. We then quantize this distribution into a discrete distribution with $N$ fixed supports at $\{f(b_0), \ldots, f(b_{N-1})\}$, as Jiang et al. (2024a) has shown that the Wasserstein-2 distance minimizing quantization is given by

$$q(t_j|x) = F\Big(\frac{f(b_{j+1}) + f(b_j)}{2}\Big) - F\Big(\frac{f(b_j) + f(b_{j-1})}{2}\Big),$$

where $F(\cdot)$ is the CDF function of the Gaussian distribution $\mathcal{N}(y_i, \sigma^2)$. Figure 2 shows some example quantization. We then use the quantized distribution as the target for our model training. This allows us to further discriminate targets that fall in the same bin. While prior work has been shown that enforcing distribution over a set of sequences can be achieved through sampling and reweighting (Zhang et al., 2024), our regression setting only requires a single decoding step. This enables us to directly optimize the model using a forward KL-divergence loss:

$$\mathcal{L}_{\text{Direct}}(\mathcal{D}) = \mathbb{E}_{(x,y)\sim\mathcal{D}} D_{\text{KL}}\Big(Q(t|x)||P_\theta(t|x)\Big).$$

We use forward KL-divergence as we observe more stable training, and it seems that the theoretical mode-seeking behavior of reverse-KL does not always hold for LLMs (Wu et al., 2025).

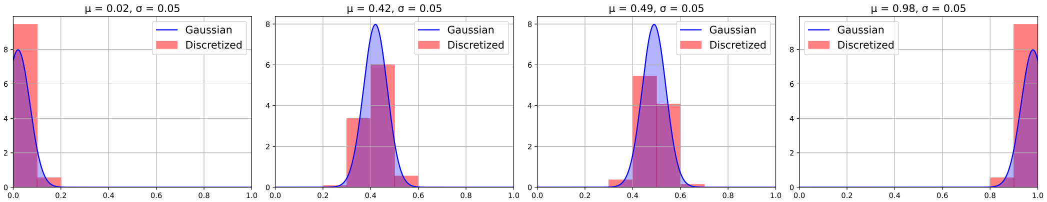


Figure 2: Illustration of our distribution quantization process. Notice that how the quantization preserves fine-grained label ordering and allows for better discrimination of targets that fall in the same bin.

**Utilizing Synthetic Data** While our approach to regression allows us to convert any target scalar label $y$ into a distribution over bin-associated tokens, our synthetic data is annotated with multiple LLMs to provide a more reliable estimate of the true conditional probability. On each synthetic data point $x \in \mathcal{X}$, our synthetic data creation process generates $K$ probability estimates $\{y^0, \ldots, y^{k-1}\}$, paired with $K$ confidence judgments $\{c^0, \ldots, c^{k-1}\}$. As through the previous discussion, each of the score can be mapped to a distribution over tokens. To increase sharpness of the overall prediction, we construct a mixture distribution over $\{Q^0(t|x), \ldots, Q^{k-1}(t|x)\}$, where each component is weighted by a power-normalized confidence score:

$$Q(t|x) = \sum_{k=0}^{K-1} \pi^k \cdot Q^k(t|x), \text{ where } \pi^k = \frac{(c^k)^\alpha}{\sum_{j=0}^{K-1}(c^j)^\alpha}.$$

**Rank Consistency Training** As ground truth probability estimates are only available on very limited data points, we can still leverage other forms of supervision to improve model performance. Similar to how we construct the evaluation suite from defeasible NLI (Rudinger et al., 2020) and choice of plausible alternatives (Dagan et al., 2005; Zellers et al., 2019), we can supervise rank-consistency using pairwise ranking labels. Given two instance $(x_1, y_1)$, $(x_2, y_2)$, we use the pairwise margin-loss (Weston et al., 1999) to enforce the order consistency:

$$\mathcal{L}_{\text{Rank}}(\mathcal{D}_{\text{Pairwise}}) = \mathbb{E}_{\mathcal{D}_{\text{Pairwise}}} \max\{0, \delta - \text{sgn}(y_1 - y_2) \cdot (\hat{y}_1 - \hat{y}_2)\},$$

as it has been shown to lead to less over-confident probability estimates (Li et al., 2019). Thus the final loss becomes

$$\mathcal{L}(\mathcal{D}) = \mathcal{L}_{\text{Direct}}(\mathcal{D}_{\text{Human}}) + \beta_1 \mathcal{L}_{\text{Direct}}(\mathcal{D}_{\text{Synthetic}}) + \beta_2 \mathcal{L}_{\text{Rank}}(\mathcal{D}_{\text{Pairwise}}).$$

Notice that while Jiang et al. (2024a) empirically show that optimizing for rank-consistency improves the calibration of the underlying classifier, their theoretical results depend on the assumption that the predictive distribution is single-peaked. For our case this does not always hold, and in fact our gradient analysis in Appendix B shows that the margin-loss as well as expected-scoring-rule-based regression loss alone can be under-specified, and does not guarantee that greedy decoding will yield a reasonable approximation of the target conditional probability. Therefore, we apply rank-consistency training only when the pairwise labels are available, to improve correspondence of the probability estimates.

## 4 Experiment Setup

### 4.1 Dataset

**Training** We construct our training dataset from a mixture of human-annotated, synthetic, and pairwise ranking data. As our tasks are closely related to Natural Language Inference (NLI), we heavily rely on NLI datasets for training. We use the following datasets: **UNLI** (Chen et al., 2020), which contains high-quality subjective probabilistic relabeling of a subset of SNLI (Bowman et al., 2015); **ANLI** (Nie et al., 2020a), which contains adversarially generated NLI examples that are often longer and involve more challenging reasoning patterns (Williams et al., 2020); and **WANLI** (Liu et al., 2022), which consists of automatically generated NLI pairs seeded from challenging instances in MultiNLI (Williams et al., 2018), identified through data maps (Swayamdipta et al., 2020).

Since UNLI applies a logistic transformation to its annotation scale, we invert that transformation to make our model more sensitive to probability differences near 0 and 1. As ANLI and WANLI do not provide probabilistic labels, we use our proposed synthetic data annotation pipeline, detailed in subsection 3.2, to generate pseudo-labels for these datasets.

Additionally, we leverage **$\delta$-NLI** (Rudinger et al., 2020) and **HellaSwag** (Zellers et al., 2019) for rank-consistency training. For $\delta$-NLI, we train our models' probability estimates to be consistent with the direction of defeasible updates, while for HellaSwag, we train the models to assign the highest probability to the most plausible completion. During training, we upsample UNLI and $\delta$-NLI by up to 4 times (Lee et al., 2022; Li et al., 2025).

**Intrinsic Evaluation** Besides evaluating our models on the **UNLI** dataset, we also sample from **EntailmentBank** (Dalvi et al., 2021) and **e-CARE** (Du et al., 2022) to directly evaluate the quality of our probability estimates. While EntailmentBank trees are predominantly entailment-focused, we rewrite hypotheses to introduce probabilistic uncertainty by removing modality, increasing vagueness, generating alternatives, perform existential instantiation, and generate abductions to convert the task into a probabilistic reasoning challenge. To evaluate the quality of probability estimations on other domains, we follow **GNLI** (Hosseini et al., 2024) to generate synthetic NLI pairs in more distant domains such as fans forum, blog post, medical texts, and others. We also cast the scalar annotation (Jiang et al., 2024a) on Circa (Louis et al., 2020) to the probability of an ambiguous response being affirmative. For all intrinsic evaluation, we report Spearman correlation.

**Comparison Evaluation** We evaluate accuracy of plausibility ranking on **$\delta$-NLI**, **HellaSwag** and **COPA** (Roemmele et al., 2011). Specifically, for COPA, we rank the two choices based on the conditional probability of the effect given the cause, regardless of the question type. For each instance, our model is required to predict conditional probability for each of the alternatives independently, which makes the task more challenging than the original multiple-choice setting. We report the accuracy of the model's top-1 ranking.

**Structural Evaluation** We evaluate our model on reasoning traces from two reasoning frameworks. The first is **Maieutic Prompting** (Jung et al., 2022), which iteratively expands a maieutic tree using abductive and recursive explanations, solving for the most consistent truth value assignment. While the original formulation incorporates multiple methods to estimate uncertainty and applies post-hoc filtering, we make the task more challenging by relying solely on our model to estimate all required conditional probabilities. Specifically, we estimate both *belief* and *consistency*, and we remove the NLI filtering constraint to further increase difficulty.

The second framework is **BIRD** (Feng et al., 2024), which performs Bayesian inference through LLM-generated abductions. Given a BIRD trace, we use our model to score both the conditional probabilities of each factor condition given the context, and the probabilities of each final outcome given those conditions. Unlike the original BIRD method, we do not apply any optimization for consistency, aside from renormalizing the aggregated probabilities over final outcomes.

For Maieutic Prompting, we evaluate on publicly available traces from **Com2Sense** (Singh et al., 2021), **CREAK** (Onoe et al., 2021), and **CSQA2** (Talmor et al., 2021)all binary QA benchmarks focused on commonsense reasoning and fact verification. For BIRD, we use their released dataset, including additional sentence generation for the comparison subset of **Com2Sense**, and evaluate on **Today** (Feng et al., 2023) for temporal reasoning. We report overall accuracy across all datasets. Example traces and scoring details can be found in Appendix E.

### 4.2 Model and Tuning Details

We use four open-source models to generate synthetic datasets: Qwen2.5-32B-Instruct and QwQ-32B (Qwen et al., 2025), DeepSeek-R1-Distill-Qwen-32B (DeepSeek-AI et al., 2025), and Llama-3.3-70B-Instruct (Grattafiori et al., 2024). For all models except QwQ, we apply greedy decoding; for QwQ, we follow the official recommendation of using temperature $= 0.6$ and MinP $= 0$ to mitigate repetition.[5]

For fine-tuning, we consider three base models: Llama-3-8B-Instruct, Qwen2.5-7B-Instruct, and Qwen2.5-14B-Instruct. We apply LoRA (Hu et al., 2022) with $r = 16$ and $\alpha = 32$, using a learning rate of $2 \times 10^{-5}$ and a linear learning rate scheduler. For each configuration, we perform a hyperparameter sweep over the number of levels $\in \{10, 20, 100\}$, $\beta_2 \in \{1, 10, 100\}$, and $\sigma \in \{0.01, 0.05, 0.1\}$.

We evaluate our proposed models against a diverse set of baselines to ensure a comprehensive comparison. These baselines are: 1) Encoder Model: The RoBERTa-L model (Chen et al., 2020), a fine-tuned encoder using UNLI dataset. 2) Zero-Shot LLMs: We prompt both a proprietary model (GPT4o (Achiam et al., 2023)) and an open-source model (DeepSeekR1DistillQwen32B (DeepSeek-AI et al., 2025)) in a zero-shot inference setting. Due to the slow inference speed of GPT-4o on structural reasoning tasks with the Maieutic Prompting method, we employ stratified sampling to create a representative mini-batch for its evaluation. 3) Probe Method: We adopt the true/false probing technique from Tian et al. (2023). This involves extracting probabilities from the ”true” and ”false” token logits of the Qwen2.5-14B-Instruct model, which are then calibrated with a general temperature scaling method (Shen et al., 2024).

## 5 Results and Discussion

Our comprehensive evaluation result is shown in Table 1. Overall, we observe noticeable improvements over previous generation models for subjective probability estimation (Chen et al., 2020), as well as over 0-shot prompting approaches that are widely considered. We discuss particular observations below.

**Model performance improves with scale.** Even when trained on the same dataset as prior work (Chen et al., 2020), our models built on top of modern LLM backend consistently outperform both smaller encoder-based approaches and zero-shot prompting baselines. Notably, Chen et al. (2020) report an aggregated human performance of Spearman's $\rho = 0.727$, whereas all of our models exceed this benchmark by a substantial margin. These results suggest that large language models, with their extensive world knowledge, can be effectively leveraged for probabilistic inference.

[5] https://huggingface.co/Qwen/QwQ-32B

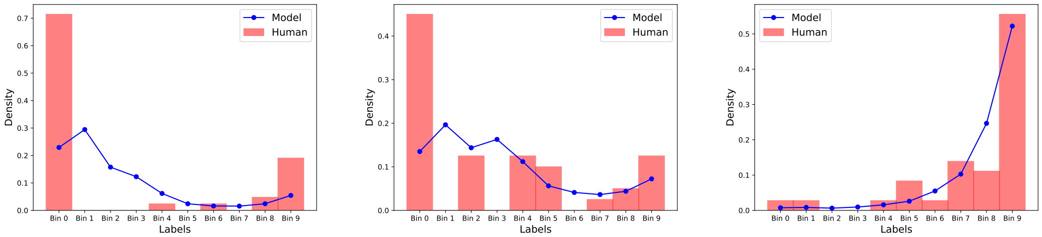
| Type | Tasks | Encoder | 0-Shot | | Probe | Ours | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | RoBERTa-L | 🐋 | ֍ | ⬡-14B | 🦙-8B | ⬡-7B | ⬡-14B | +Syn | +Syn+R |
| Intrinsic | UNLI | .707 | .629 | .699 | .681 | .804 | .802 | **.813** | .812 | .812 |
| | circa | .430 | .663 | .734 | .553 | .474 | .544 | .536 | .564 | **.747** |
| | GNLI | .586 | .755 | .796 | .843 | .789 | .811 | .814 | **.838** | .820 |
| | EntailmentBank | .503 | .783 | .760 | .558 | .659 | .687 | .735 | **.789** | .787 |
| | e-CARE | .738 | .871 | .898 | .856 | .855 | .870 | .888 | .884 | **.905** |
| Comp. | $\delta$-SNLI | 77.9 | 77.2 | 75.3 | 81.3 | 83.3 | 84.4 | 85.1 | 86.0 | **88.9** |
| | $\delta$-ATOMIC | 75.0 | 69.1 | 70.7 | 74.7 | 78.9 | 78.6 | 80.1 | 81.2 | **87.6** |
| | COPA | 83.0 | 87.7 | 81.0 | 86.8 | 85.1 | 87.2 | 86.5 | 87.9 | **89.3** |
| | HellaSwag | 42.0 | 57.8 | 75.4 | 74.9 | 67.2 | 70.2 | 75.3 | 75.5 | **95.7** |
| Structural | C2S-M | 50.6 | 68.5 | **77.9** | 49.4 | 65.5 | 68.4 | 73.5 | 75.2 | 75.6 |
| | CREAK-M | 62.5 | 68.0 | 81.1 | 42.1 | 83.2 | 82.8 | 84.8 | 85.6 | **86.5** |
| | CSQA2-M | 49.5 | 55.6 | 71.4 | 48.5 | 60.4 | 63.0 | 67.9 | 70.4 | **72.0** |
| | C2S-Sent-B | 65.0 | 73.0 | 76.8 | 51.3 | 72.7 | 76.3 | 71.3 | 73.3 | **89.8** |
| | TODAY-B | 56.0 | 64.0 | 63.0 | 65.0 | 65.0 | 64.0 | **68.0** | 67.0 | 66.3 |

Table 1: Evaluation of our models on the tasks listed out in subsection 4.1. 🐋 denotes the DeepSeek-R1-Distill-Qwen-32B (DeepSeek-AI et al., 2025) and ֍ GPT-4o (Achiam et al., 2023). 🦙 corresponds to Llama-3-Instruct series (Grattafiori et al., 2024), ⬡ the Qwen2.5-Instruct series (Qwen et al., 2025). **+Syn** indicates Qwen2.5-14B-Instruct augmented with synthetic data, **+R** indicates the same model with rank consistency training. -M and -B suffixes denote the Maieutic Prompting and Bird frameworks respectively. Best results are in **Bold**, second-best results are underlined.

**Synthetic data enhances domain generalization.** Comparing the results in column **+Syn** with those in column ⬡-14B, we find that incorporating synthetic data generally improves performance. These improvements are especially notable in probability estimation for more challenging reasoning tasks (e.g., EntailmentBank) and in generalization to distant domains (e.g., GNLI, Circa). We hypothesize that this is due to the broader range of inference types covered by ANLI and WANLI, as well as their longer contexts, which may help the model better understand complex scenarios and integrate multiple pieces of evidence.

**Rank-consistency training facilitates better decision making.** The results in column **+Syn+R** clearly demonstrate that rank-consistency training further improves performance across multiple tasks, particularly in plausibility ranking and structured reasoning. While this outcome is intuitive, it is noteworthy that on some taskssuch as $\delta$-SNLI and C2S-Sent-B our models perform competitively or even outperform previous results, of larger models or systems employing more complex processing pipelines (Srikanth & Rudinger, 2025; Feng et al., 2024). Although the improvements with Maieutic Prompting are less pronounced, this is likely due to its lack of a fully probabilistic interpretation (Jung et al., 2022).

**Our model captures some level of human uncertainty.** An additional benefit of our model is that, even without explicit training on human label distributions, it still exhibits human-like uncertainty on many data points, as shown in Figure 3. To show this, we compare the token-level distribution of our model with the distributed scalar judgments collected by Pavlick & Kwiatkowski (2019). We observe that our model produces multi-modal distributions when human opinion divides, and agree with human judgment when the opinion is more consistent. A more detailed analysis of our model's performance on the ChaosNLI (Nie et al., 2020b) and ProtoQA (Boratko et al., 2020) datasets is provided in Appendix C.



**P**: Ruth's 1927 single season record of 60 home runs stood unsurpassed until Roger Maris hit 61 in 1961. ⇝ **H**: Babe Ruth hit 60 home runs in his lifetime.

**P**: If no settlement is reached, a divided Cyprus will join the European Union on May 1, 2004. ⇝ **H**: Cyprus was divided into 2 parts on May 1, 2004.

**P**: Man in green vest directing traffic in snowy conditions, pedestrians standing on the sidelines. ⇝ **H**: A person remains in the conditions.

Figure 3: Comparison of token-level distribution between our model and human label distribution. The x-axis represents the probability bins, while the y-axis indicates the probability density. The distribution learned by our model reflects intrinsic human disagreements.

## 6 Conclusion

In this work, we present a series of strong and precise probability estimation models, developed through a novel pipeline that combines LLM-based synthetic data generation with decoder-based regression fine-tuning. Our models consistently outperform previous generation approaches for similar tasks, as well as commonly used zero-shot prompting methods with strong LLMs. Through comprehensive evaluation, we demonstrate that improving local probabilistic estimation can significantly enhance the performance of complex probabilistic reasoning systems. We hope our work inspires future advancements in building efficient, accurate models for real-world probability estimation and, more broadly, in developing general and versatile probabilistic reasoning systems.

## 7 Acknowledgments

This work is supported by U.S. National Science Foundation under grant No. 2204926, ONR grant (N0001424-1-2089) and Defense Advance Research Projects Agency (DARPA) under Contract No. HR001125C0304. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation and DARPA.

## References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

Afra Feyza Akyürek, Ekin Akyürek, Leshem Choshen, Derry Tanti Wijaya, and Jacob Andreas. Deductive closure training of language models for coherence, accuracy, and updatability. In Findings of the Association for Computational Linguistics ACL 2024, pp. 9802–9818, 2024.

Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernández. Stop measuring calibration when humans disagree. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 1892–1915, 2022.

Joris Baan, Raquel Fernández, Barbara Plank, and Wilker Aziz. Interpreting predictive probabilities: Model confidence or human label variation? In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 268–277, 2024.

Christina Baek, Yiding Jiang, Aditi Raghunathan, and J Zico Kolter. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), Advances in Neural Information Processing Systems, 2022. URL https://openreview.net/forum?id=EZZsnke1kt.

Michael Boratko, Xiang Li, Tim O'Gorman, Rajarshi Das, Dan Le, and Andrew McCallum. ProtoQA: A question answering dataset for prototypical common-sense reasoning. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1122–1136, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.85. URL https://aclanthology.org/2020.emnlp-main.85/.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In Lluís Màrquez, Chris Callison-Burch, and Jian Su (eds.), Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://aclanthology.org/D15-1075/.

Shulin Cao, Jiajie Zhang, Jiaxin Shi, Xin Lv, Zijun Yao, Qi Tian, Lei Hou, and Juanzi Li. Probabilistic tree-of-thought reasoning for answering knowledge-intensive complex questions. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 12541–12560, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.835. URL https://aclanthology.org/2023.findings-emnlp.835/.

Sky CH-Wang, Benjamin Van Durme, Jason Eisner, and Chris Kedzie. Do androids know they're only dreaming of electric sheep? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), Findings of the Association for Computational Linguistics: ACL 2024, pp. 4401–4420, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.260. URL https://aclanthology.org/2024.findings-acl.260/.

Beiduo Chen, Xinpeng Wang, Siyao Peng, Robert Litschko, Anna Korhonen, and Barbara Plank. "seeing the big through the small": Can LLMs approximate human judgment distributions on NLI from a few explanations? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 14396–14419, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.842. URL https://aclanthology.org/2024.findings-emnlp.842/.

Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. Uncertain natural language inference. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 8772–8779, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.774. URL https://aclanthology.org/2020.acl-main.774.

Yangyi Chen, Lifan Yuan, Ganqu Cui, Zhiyuan Liu, and Heng Ji. A close look into the calibration of pre-trained language models, 2023. URL https://arxiv.org/abs/2211.00151.

Qi Cheng, Michael Boratko, Pranay Kumar Yelugam, Tim OGorman, Nalini Singh, Andrew Mccallum, and Xiang Li. Every answer matters: Evaluating commonsense with probabilistic measures. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 493–506, 2024.

Cheng-Han Chiang and Hung yi Lee. A closer look into using large language models for automatic evaluation. In The 2023 Conference on Empirical Methods in Natural Language Processing, 2023. URL https://openreview.net/forum?id=RsK483IRuO.

André F Cruz, Moritz Hardt, and Celestine Mendler-Dünner. Evaluating language models as risk scores. In The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2024. URL https://openreview.net/forum?id=qrZxL3Bto9.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment, MLCW'05, pp. 177190, Berlin, Heidelberg, 2005. Springer-Verlag. ISBN 3540334270. doi: 10.1007/11736790_9. URL https://doi.org/10.1007/11736790_9.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. Explaining answers with entailment trees. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 7358–7370, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.585. URL https://aclanthology.org/2021.emnlp-main.585/.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Ailin Deng, Miao Xiong, and Bryan Hooi. Great models think alike: improving model reliability via inter-model latent agreement. In Proceedings of the 40th International Conference on Machine Learning, ICML'23. JMLR.org, 2023.

Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 295–302, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), pp. 4171–4186, 2019.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. e-CARE: a new dataset for exploring explainable causal reasoning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 432–446, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.33. URL https://aclanthology.org/2022.acl-long.33/.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In Proceedings of the 41st International Conference on Machine Learning, ICML'24. JMLR.org, 2024.

Yu Feng, Ben Zhou, Haoyu Wang, Helen Jin, and Dan Roth. Generic temporal reasoning with differential analysis and explanation. In The 61st Annual Meeting Of The Association For Computational Linguistics, 2023.

Yu Feng, Ben Zhou, Weidong Lin, and Dan Roth. Bird: A trustworthy bayesian inference framework for large language models, 2024. URL https://arxiv.org/abs/2404.12494.

Oren Glickman, Ido Dagan, and Moshe Koppel. A probabilistic classification approach for lexical textual entailment. In AAAI, pp. 1050–1055. Pittsburgh, PA, 2005.

Kristina Gligorić, Tijana Zrnic, Cinoo Lee, Emmanuel J Candès, and Dan Jurafsky. Can unconfident llm annotations be used for confident conclusions? arXiv preprint arXiv:2408.15204, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. arXiv preprint arXiv:2411.15594, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In International conference on machine learning, pp. 1321–1330. PMLR, 2017.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300, 2020.

Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill™: A bayesian skill rating system. In B. Schölkopf, J. Platt, and T. Hoffman (eds.), Advances in Neural Information Processing Systems, volume 19. MIT Press, 2006. URL https://proceedings.neurips.cc/paper_files/paper/2006/file/f44ee263952e65b3610b8ba51229d1f9-Paper.pdf.

Mohammad Javad Hosseini, Andrey Petrov, Alex Fabrikant, and Annie Louis. A synthetic data approach for domain generalization of nli models. arXiv preprint arXiv:2402.12368, 2024.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. Decomposing uncertainty for large language models through input clarification ensembling. In ICML, 2024a. URL https://openreview.net/forum?id=byxXa99PtF.

Bairu Hou, Yang Zhang, Jacob Andreas, and Shiyu Chang. A probabilistic framework for llm hallucination detection via belief tree propagation. arXiv preprint arXiv:2406.06950, 2024b.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. ICLR, 1(2):3, 2022.

Kevin G. Jamieson and Robert D. Nowak. Active ranking using pairwise comparisons. In Proceedings of the 25th International Conference on Neural Information Processing Systems, NIPS'11, pp. 22402248, Red Hook, NY, USA, 2011. Curran Associates Inc. ISBN 9781618395993.

Edwin T Jaynes. Probability theory: The logic of science. Cambridge university press, 2003.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. Investigating reasons for disagreement in natural language inference. Transactions of the Association for Computational Linguistics, 10:1357–1374, 2022.

Zheng Ping Jiang, Anqi Liu, and Benjamin Van Durme. Calibrating zero-shot cross-lingual (un-) structured predictions. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 2648–2674, 2022.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 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, 2021.

Zhengping Jiang, Anqi Liu, and Benjamnin Van Durme. Addressing the binning problem in calibration assessment through scalar annotations. Transactions of the Association for Computational Linguistics, 12:120–136, 2024a.

Zhengping Jiang, Jingyu Zhang, Nathaniel Weir, Seth Ebner, Miriam Wanner, Kate Sanders, Daniel Khashabi, Anqi Liu, and Benjamin Van Durme. Core: Robust factual precision with informative sub-claim identification. arXiv preprint arXiv:2407.03572, 2024b.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. Maieutic prompting: Logically consistent reasoning with recursive explanations. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 1266–1279, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.82. URL https://aclanthology.org/2022.emnlp-main.82/.

William Jurayj, Jeffrey Cheng, and Benjamin Van Durme. Is that your final answer? test-time scaling improves selective question answering. arXiv preprint arXiv:2502.13962, 2025.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. Advances in neural information processing systems, 35:22199–22213, 2022.

Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. Advances in neural information processing systems, 32, 2019.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. Training language models to self-correct via reinforcement learning. arXiv preprint arXiv:2409.12917, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In Proceedings of the 29th Symposium on Operating Systems Principles, pp. 611–626, 2023.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 8424–8445, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.577. URL https://aclanthology.org/2022.acl-long.577/.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. On the sentence embeddings from pre-trained language models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 9119–9130, 2020.

Tianjian Li, Haoran Xu, Weiting Tan, Kenton Murray, and Daniel Khashabi. Upsample or upweight? balanced training on heavily imbalanced datasets, 2025. URL https://arxiv.org/abs/2410.04579.

Xiang Lorraine Li, Adhiguna Kuncoro, Jordan Hoffmann, Cyprien de Masson d'Autume, Phil Blunsom, and Aida Nematzadeh. A systematic investigation of commonsense knowledge in large language models. arXiv preprint arXiv:2111.00607, 2021.

Zhongyang Li, Tongfei Chen, and Benjamin Van Durme. Learning to rank for plausible plausibility. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4818–4823, 2019.

Zongjie Li, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Tianxiang Li, Shuai Wang, Cuiyun Gao, and Yang Liu. Split and merge: Aligning position biases in large language model based evaluators, 2024. URL https://openreview.net/forum?id=1hLFLNu4uy.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In The Twelfth International Conference on Learning Representations, 2023.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. WANLI: Worker and AI collaboration for natural language inference dataset creation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), Findings of the Association for Computational Linguistics: EMNLP 2022, pp. 6826–6847, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.508. URL https://aclanthology.org/2022.findings-emnlp.508/.

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. Were afraid language models arent modeling ambiguity. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 790–807, 2023.

Annie Louis, Dan Roth, and Filip Radlinski. "I'd rather just go to bed": Understanding indirect answers. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 7411–7425, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.601. URL https://aclanthology.org/2020.emnlp-main.601/.

John McCarthy and Patrick J Hayes. Some philosophical problems from the standpoint of artificial intelligence. In Readings in artificial intelligence, pp. 431–450. Elsevier, 1981.

Johannes Mario Meissner, Napat Thumwanit, Saku Sugawara, and Akiko Aizawa. Embracing ambiguity: Shifting the training target of nli models. arXiv preprint arXiv:2106.03020, 2021.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. Reducing conversational agents' overconfidence through linguistic calibration. Transactions of the Association for Computational Linguistics, 10:857–872, 2022. doi: 10.1162/tacl_a_00494. URL https://aclanthology.org/2022.tacl-1.50.

Sarah Moss. Probabilistic knowledge. Oxford University Press, 2018.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. arXiv preprint arXiv:2501.19393, 2025.

Aliakbar Nafar, Kristen Brent Venable, and Parisa Kordjamshidi. Reasoning over uncertain text by generative large language models, 2024. URL https://arxiv.org/abs/2402.09614.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A new benchmark for natural language understanding. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4885–4901, Online, July 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.441. URL https://aclanthology.org/2020.acl-main.441/.

Yixin Nie, Xiang Zhou, and Mohit Bansal. What can we learn from collective human opinions on natural language inference data? In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 9131–9143, Online, November 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.734. URL https://aclanthology.org/2020.emnlp-main.734.

Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2019.

Yasumasa Onoe, Michael Zhang, Eunsol Choi, and Greg Durrett. Creak: A dataset for commonsense reasoning over entity knowledge. In J. Vanschoren and S. Yeung (eds.), Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/5737c6ec2e0716f3d8a7a5c4e0de0d9a-Paper-round2.pdf.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simn Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, ukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, ukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mly, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth

Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cern Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Batu Ozturkler, Nikolay Malkin, Zhen Wang, and Nebojsa Jojic. ThinkSum: Probabilistic reasoning over sets using large language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1216–1239, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.68. URL https://aclanthology.org/2023.acl-long.68.

Vishakh Padmakumar and He He. Does writing with language models reduce content diversity? arXiv preprint arXiv:2309.05196, 2023.

Akshay Paruchuri, Jake Garrison, Shun Liao, John Hernandez, Jacob Sunshine, Tim Althoff, Xin Liu, and Daniel McDuff. What are the odds? language models are capable of probabilistic reasoning. In Conference on Empirical Methods in Natural Language Processing, 2024. URL https://api.semanticscholar.org/CorpusID:270562235.

Ellie Pavlick and Tom Kwiatkowski. Inherent disagreements in human textual inferences. Transactions of the Association for Computational Linguistics, 7:677–694, 2019. doi: 10.1162/tacl_a_00293. URL https://aclanthology.org/Q19-1043.

R. L. Plackett. The analysis of permutations. Journal of the Royal Statistical Society. Series C (Applied Statistics), 24(2):193–202, 1975. ISSN 00359254, 14679876. URL http://www.jstor.org/stable/2346567.

Zhen Qin, Junru Wu, Jiaming Shen, Tianqi Liu, and Xuanhui Wang. LAMPO: Large language models as preference machines for few-shot ordinal classification. In First Conference on Language Modeling, 2024. URL https://openreview.net/forum?id=ig6NI9oPhD.

Linlu Qiu, Fei Sha, Kelsey R Allen, Yoon Kim, Tal Linzen, and Sjoerd van Steenkiste. Can language models perform implicit bayesian inference over user preference states? In The First Workshop on System-2 Reasoning at Scale, NeurIPS'24, 2024.

Linlu Qiu, Fei Sha, Kelsey Allen, Yoon Kim, Tal Linzen, and Sjoerd van Steenkiste. Bayesian teaching enables probabilistic reasoning in large language models. arXiv preprint arXiv:2503.17523, 2025.

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.

Yasaman Razeghi, Robert L Logan Iv, Matt Gardner, and Sameer Singh. Impact of pretraining term frequencies on few-shot numerical reasoning. In Findings of the Association for Computational Linguistics: EMNLP 2022, pp. 840–854, 2022.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In AAAI spring symposium: logical formalizations of commonsense reasoning, pp. 90–95, 2011.

Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. Thinking like a skeptic: Defeasible inference in natural language. In Trevor Cohn, Yulan He, and Yang Liu (eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 4661–4675, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.418. URL https://aclanthology.org/2020.findings-emnlp.418/.

Kate Sanders and Benjamin Van Durme. Bonsai: Interpretable tree-adaptive grounded reasoning, 2025. URL https://arxiv.org/abs/2504.03640.

Maohao Shen, Subhro Das, Kristjan Greenewald, Prasanna Sattigeri, Gregory W. Wornell, and Soumya Ghosh. Thermometer: Towards universal calibration for large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), Proceedings of the 41st International Conference on Machine Learning, volume 235 of Proceedings of Machine Learning Research, pp. 44687–44711. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/shen24c.html.

Lin Shi, Chiyu Ma, Wenhua Liang, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in llm-as-a-judge, 2024. URL https://arxiv.org/abs/2406.07791.

Vaishnavi Shrivastava, Ananya Kumar, and Percy Liang. Language models prefer what they know: Relative confidence estimation via confidence preferences. arXiv preprint arXiv:2502.01126, 2025.

Shikhar Singh, Nuan Wen, Yu Hou, Pegah Alipoormolabashi, Te-lin Wu, Xuezhe Ma, and Nanyun Peng. COM2SENSE: A commonsense reasoning benchmark with complementary sentences. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 883–898, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.78. URL https://aclanthology.org/2021.findings-acl.78/.

Xingyou Song and Dara Bahri. Decoding-based regression. arXiv preprint arXiv:2501.19383, 2025.

Xingyou Song, Oscar Li, Chansoo Lee, Bangding Yang, Daiyi Peng, Sagi Perel, and Yutian Chen. Omnipred: Language models as universal regressors. Transactions on Machine Learning Research, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=t9c3pfrR1X.

Neha Srikanth and Rachel Rudinger. Nli under the microscope: What atomic hypothesis decomposition reveals. arXiv preprint arXiv:2502.08080, 2025.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 9275–9293, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.746. URL https://aclanthology.org/2020.emnlp-main.746/.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. CommonsenseQA 2.0: Exposing the limits of AI through gamification. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1), 2021. URL https://openreview.net/forum?id=qF7FlUT5dxa.

Eric Tang, Bangding Yang, and Xingyou Song. Understanding llm embeddings for regression. arXiv preprint arXiv:2411.14708, 2024.

Sree Harsha Tanneru, Chirag Agarwal, and Himabindu Lakkaraju. Quantifying uncertainty in natural language explanations of large language models, 2023. URL https://arxiv.org/abs/2311.03533.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 5433–5442, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.330. URL https://aclanthology.org/2023.emnlp-main.330.

Robert Vacareanu, Vlad Andrei Negru, Vasile Suciu, and Mihai Surdeanu. From words to numbers: Your large language model is secretly a capable regressor when given in-context examples. In First Conference on Language Modeling, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In The Eleventh International Conference on Learning Representations, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Yuxia Wang, Minghan Wang, Yimeng Chen, Shimin Tao, Jiaxin Guo, Chang Su, Min Zhang, and Hao Yang. Capture human disagreement distributions by calibrated networks for natural language inference. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), Findings of the Association for Computational Linguistics: ACL 2022, pp. 1524–1535, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.120. URL https://aclanthology.org/2022.findings-acl.120/.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine de Marneffe, and Barbara Plank. Varierr nli: Separating annotation error from human label variation. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2256–2269, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), Advances in Neural Information Processing Systems, 2022. URL https://openreview.net/forum?id=_VjQlMeSB_J.

Jason Weston, Chris Watkins, et al. Support vector machines for multi-class pattern recognition. In Esann, volume 99, pp. 219–224, 1999.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 1112–1122, 2018.

Adina Williams, Tristan Thrush, and Douwe Kiela. Anlizing the adversarial natural language inference dataset. arXiv preprint arXiv:2010.12729, 2020.

Lionel Wong, Gabriel Grand, Alexander K Lew, Noah D Goodman, Vikash K Mansinghka, Jacob Andreas, and Joshua B Tenenbaum. From word models to world models: Translating from natural language to the probabilistic language of thought. CoRR, 2023.

Taiqiang Wu, Chaofan Tao, Jiahao Wang, Runming Yang, Zhe Zhao, and Ngai Wong. Rethinking kullback-leibler divergence in knowledge distillation for large language models. In Proceedings of the 31st International Conference on Computational Linguistics, pp. 5737–5755, 2025.

Shepard Xia, Brian Lu, and Jason Eisner. Let's think var-by-var: Large language models enable ad hoc probabilistic reasoning. arXiv preprint arXiv:2412.02081, 2024.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In The Twelfth International Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=gjeQKFxFpZ.

Zhenran Xu, Senbao Shi, Baotian Hu, Jindi Yu, Dongfang Li, Min Zhang, and Yuxiang Wu. Towards reasoning in large language models via multi-agent peer review collaboration, 2023. URL https://arxiv.org/abs/2311.08152.

Gal Yona, Shay Moran, Gal Elidan, and Amir Globerson. Active learning with label comparisons. In The 38th Conference on Uncertainty in Artificial Intelligence, 2022. URL https://openreview.net/forum?id=S2zMhPUi5xq.

Moy Yuan, Eric Chamoun, Rami Aly, Chenxi Whitehouse, and Andreas Vlachos. PRobELM: Plausibility ranking evaluation for language models. In First Conference on Language Modeling, 2024. URL https://openreview.net/forum?id=k8KS9Ps71d.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. Advances in Neural Information Processing Systems, 35:15476–15488, 2022.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

Yiming Zhang, Avi Schwarzschild, Nicholas Carlini, J Zico Kolter, and Daphne Ippolito. Forcing diffuse distributions out of language models. In First Conference on Language Modeling, 2024.

Shengjia Zhao, Michael Kim, Roshni Sahoo, Tengyu Ma, and Stefano Ermon. Calibrating predictions to decisions: A novel approach to multi-class calibration. Advances in Neural Information Processing Systems, 34:22313–22324, 2021a.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In International conference on machine learning, pp. 12697–12706. PMLR, 2021b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

Xiang Zhou, Yixin Nie, and Mohit Bansal. Distributed nli: Learning to predict human opinion distributions for language reasoning. arXiv preprint arXiv:2104.08676, 2021.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020. URL https://arxiv.org/abs/1909.08593.

# A The Other Probability Estimation Attempt

In this section, we introduce an alternative approach to probability estimation by scaling compute for synthetic data creation. Inspired by the preference data used in Reinforcement Learning from Human Feedback, we hypothesize that pairwise comparison is cognitively easier and more reliable than direct scoring on probability estimation task (Ziegler et al., 2020). Our method consists of three main steps: **pairwise comparison, results aggregation and score mapping**. We describe each component in detail below.

## A.1 Pairwise Comparison

The core of this probability synthesis pipeline is the pairwise comparison of NLI samples using LLMs. Given two given NLI samples $A$ and $B$, each consisting of a premise and hypothesis, the LLM is prompted to determine which sample has a higher probability of hypothesis conditioned on the premise. We adopt the prompt format logic introduced by Qin et al. (2024), for example:

*"Compare the probabilities of two NLI samples. Passage A: $\{A\}$, Passage B: $\{B\}$."*

For specific prompts, please refer to Appendix D.1. To mitigate potential positional biases (Zheng et al., 2023; Li et al., 2024; Shi et al., 2024), where an LLM might prefer either the first or second sample regardless of its content, we alternate the order of $A$ and $B$ in the prompt. A preference for $A$ over $B$, denoted $p(A \succ B)$ is considered valid only if the model consistently ranks $A$ higher across both orderings; otherwise, the comparison is treated as a draw (Qin et al., 2024).

## A.2 Results Aggregation

To construct a global ranking of the test dataset, we aggregate pairwise comparison results using the TrueSkill framework (Herbrich et al., 2006). In this setup, each NLI sample is modeled as a player with a Gaussian skill distribution characterized by mean $\mu$ and variance $\sigma^2$. For every comparison between two samples, TrueSkill updates their skill distributions accordingly.

The computational cost of this aggregation is efficient, operating near the theoretical minimum of $n \log(n)$ comparisons for a dataset of size $n$ Herbrich et al. (2006). To further reduce comparison overhead, we adopt a binning strategy. The test set is divided into $m$ bins based on scores from direct prompting, e.g., with $m = 5$, bins correspond to levels like "impossible","technically possible", "plausible", "likely" and "very likely". During ranking, comparisons are limited to samples within the same bin, effectively reducing comparison times to converge, as previous researches show that adaptive selection of comparisons can reduce the number of required assessments while maintaining ranking accuracy (Jamieson & Nowak, 2011; Yona et al., 2022). We iterative updates each samples Gaussian distribution until its variance $\sigma$ falls below a predefined threshold $\delta$, indicating a stable estimate.

## A.3 Score Mapping

In this final step, we map the relative rankings of NLI samples into scalar probability scores. While TrueSkill provides meaningful relative rankings, its outputs are not directly interpretable as probabilities. To bridge this gap, we adopt the PlackettLuce approximation (Plackett, 1975). Specifically, we use the ratio $\mu/\sigma$ from each samples distribution as a proxy "score", and apply a softmax transformation to obtain a calibrated probability distribution.

## A.4 Experiment Setup

We conduct our experiments on a subset of the UNLI test split (Chen et al., 2020), focusing exclusively on NLI examples labeled as "neutral". This is because samples labeled as "entailment" or "contradiction" typically have ground-truth probabilities close to 1 or 0,

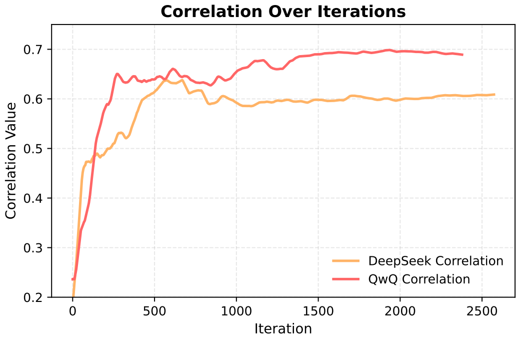


Figure 4: The Spearman Correlation over Pairwise Comparison Iterations.

respectively, whereas our interest lies in distinguishing uncertainty in the middle range. Since we are working with a small dataset, we only adopt the basic strategy to randomly sample compared NLI samples and update their scores using the TrueSkill framework. The random seed is fixed at 42 to ensure the same comparison process for different models. We initialize each player's mean randomly between 0 and 1 and set the initial variance $\sigma = 3$, where a larger variance reflects higher uncertainty in the probability estimate.

We evaluate performance using two primary metrics: **Spearman correlation** between human annotations and the synthesized probability scores, which reflects overall ranking quality; Pairwise comparison accuracy (**Compare Acc**), which measures agreement between model comparisons and the ground-truth ordering. We also track how the correlation evolves as the number of comparison iterations increases.

For the comparison model, we test the performance on DeepSeek-R1-Distill-Qwen-32B (DeepSeek-AI et al., 2025) and QwQ-32B (Qwen et al., 2025).

| Method | Correlation | Compare Acc |
|---|---|---|
| Pairwise Compare | 0.6420 | 0.7562 |
| Direct Prompting | 0.6872 | / |
| Pairwise Compare | 0.6991 | 0.7655 |
| Direct Prompting | 0.7414 | / |

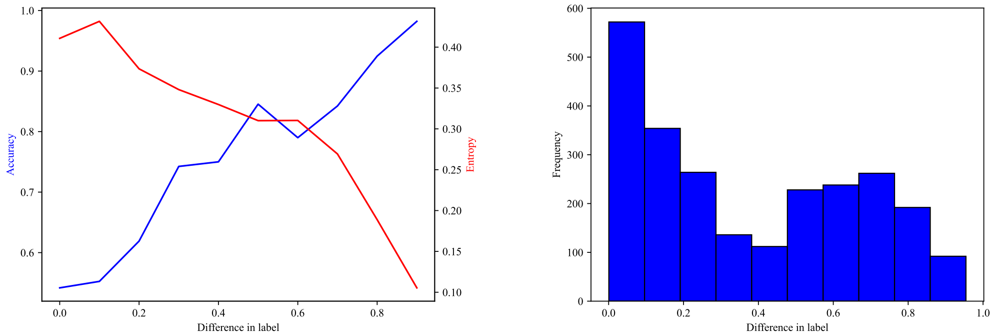
Table 2: Pairwise Compare Results

### A.5 Results and Discussion

**Main Result** Table 2 and Figure 4 presents the results of probability estimation using different methods and models. The final correlations from pairwise comparison process fall short of expectations and do not surpass the direct prompting method. We attribute this to the limited reliability of the pairwise comparison judgments, particularly due to the noisiness in comparing semantically unrelated NLI passages. While pairwise comparison is often considered cognitively easier in human evaluation settings, in our case, the LLMs require comparable amounts of reasoning and contextual understanding for both direct scoring and relative ranking. This undermines the presumed advantage of pairwise comparison in this task. These results suggest that for probabilistic NLI tasks, pairwise comparison may not offer a significant benefit over direct prompting, especially when comparisons are performed over contextually disjoint samples. More structured or semantically clustered comparisons may be necessary to fully realize the benefits of pairwise reasoning.

**Difficulty of NLI Comparison** We employ GPT-4o (OpenAI et al., 2024) to do pairwise comparison. The comparison accuracy is evaluated against ground-truth pairwise labels, while entropy, computed from the API logits, serves as a measure of the model's uncertainty.

As illustrated in Figure 5, the trends in accuracy and entropy align with intuitive expectations. Specifically, as the difference in label probabilities increases, the comparison accuracy improves, and the entropy decreases. This pattern suggests that the model becomes more confident in its decisions when the distinction between the samples is more significant. The results highlight GPT-4o can produce reliable judgments in scenarios with greater label separability while reflecting higher uncertainty in ambiguous cases.



(a) Accuracy and entropy of pairwise comparison as a function of label difference.

(b) Distribution of label differences in the sampled data.

Figure 5: Analysis of GPT-4.0 on NLI probability pairwise comparison tests. (a) Higher label differences lead to improved accuracy and reduced entropy. (b) Frequency distribution of label differences across the sampled data.

## B Analysis of Loss Function Gradients

We analyses the gradients of different losses we adopted to the class-wise probability after the softmax layer. We consider the following losses. Suppose that $b_i$ is the i-th bin, and $f(b_i)$ is the corresponding scalar score. Let $\mathbf{p}$ be the predictive distribution from the model and each $p_i$ be the predicted probability for the i-th bin, and $\boldsymbol{\pi}$ and $\pi_i$ the corresponding logits respectively. $s$ be the ground truth scalar score.

**Mean Square Error** takes the form

$$l_{\text{MSE}}(\boldsymbol{\pi}) = \frac{1}{2}\Big(\sum_{i=0}^{N-1} p_i \cdot f(b_i) - s\Big)^2, \text{ where the gradient is given by}$$

$$\frac{\partial l_{\text{MSE}}}{\partial \pi_i} = \Big(\sum_{i=0}^{N-1} p_i \cdot f(b_i) - s\Big) p_i \Big(f(b_i) - s\Big).$$

Therefore, if the model underpredict the probability, the logits for all bins above the true value will be increased, and the more the differences between the bin value and the true value, the more the logits will be increased. This is not ideal as we typically want to model to center its prediction around the true value.

**Margin Loss** For two instance $(x^j, s^j)$, $(x^k, s^k)$, the uncropped margin loss takes the form

$$
\begin{aligned}
l_{\text{Margin}}(\boldsymbol{\pi}^j, \boldsymbol{\pi}^k) &= -\text{sgn}(s^j - s^k) \cdot \Big( \sum_{i=0}^{N-1} p_i^j f(b_i) - \sum_{i=0}^{N-1} p_i^k f(b_i) \Big), \text{ where} \\
\frac{\partial l_{\text{Margin}}}{\partial \pi_i^j} &= \Big( f(b_i) - \sum_{i=0}^{N-1} p_i^j f(b_i) \Big) p_i^j,
\end{aligned}
$$

which similarly encourages the model to further updates the logits for the further away bins from the current prediction. To support greedy decoding of reasonable coarse labels, on human annotated data, we still train with the KL-divergence loss to match the quantized label distribution.

## C Aligning with Human Crowd Distributions

To evaluate the instance-level calibration of our model, we assess its ability to produce probability estimates that align with aggregated human judgments for the same instance. This is a crucial aspect of building trustworthy models whose confidence reflects human consensus. We conduct two experiments on datasets with rich human annotations to validate this capability.

First, we utilize the ChaosNLI dataset (Nie et al., 2020b), which provides 100-annotator voting distributions for NLI examples. Following a similar methodology to Jiang et al. (2024a), we map the distributions onto a single scalar probability score to serve as the ground truth. Specifically, we use the mapping rule:

$$p(\text{entailment}) = 1.0, p(\text{neutral}) = 0.2, p(\text{contradiction}) = 0.0.$$

As shown in Table 3, our model's predicted probabilities achieve a high Spearman's $\rho$ correlation with these human-derived scores and a correspondingly low ranking risk. This indicates a strong monotonic relationship between our model's confidence and the consensus of a human crowd.

| **Dataset** | **Total Examples** | **Spearman's $\rho \uparrow$** | **Ranking Risk $\downarrow$** |
|---|---|---|---|
| ChaosNLI-S | 1514 | 0.9029 | 0.1959 |
| ChaosNLI-M | 1599 | 0.8031 | 0.1959 |

Table 3: Correlation and ranking risk between our model's predictions and aggregated human judgments on ChaosNLI. High correlation and low risk demonstrate strong alignment.

To further validate these findings, we evaluate our model's calibration performance on the ProtoQA dataset (Boratko et al., 2020), a commonsense question-answering task where each question has multiple plausible answers annotated with human vote frequencies. For each question-answer pair $(q, a_i)$, we treat our model's predicted probability, $p(a_i|q)$, as the output of a probabilistic classifier. Table 4 presents the standard calibration metrics. The low Expected Calibration Error (ECE), Brier Score, and Jensen-Shannon Divergence (JSD) demonstrate that the model's predictive distribution is well-calibrated against the distribution of human judgments.

| **Metric** | ECE $\downarrow$ | Brier Score $\downarrow$ | JSD $\downarrow$ |
|---|---|---|---|
| **Value** | 0.0105 | 0.0159 | 0.0499 |

Table 4: Calibration metrics on ProtoQA. Low error values across all metrics confirm that the model's predictive distribution aligns well with human answer frequencies.

Taken together, the high rank correlation on ChaosNLI and the low calibration error on ProtoQA strongly suggest that our model's probability estimates effectively mirror human-annotated label distributions.

# D Prompt Template

Here are the prompts used in our experiments, with input data inserted into the curly brackets.

## D.1 Probability Extraction

```
system: You are a helpful assistant good at probabilistic reasoning in the real world
    setting,
human: Given a premise and a hypothesis, evaluate the probability of the hypothesis
    being true based on the information provided in the premise, supplemented by world
    knowledge and probabilistic reasoning.

Specifically:
1. Use relevant world knowledge to assess contextual factors (e.g., demographics, common
     practices, or statistical distributions) that may influence the likelihood of the
    hypothesis given the premise.
2. Perform the probabilistic reasoning to estimate the conditional probability P(
    Hypothesis | Premise).
3. Assign a probability score between [0, 1] that quantifies P(Hypothesis | Premise).
    Ensure this score reflects the strength of the connection between the premise and
    hypothesis based on probabilistic reasoning and world knowledge.

Premise: {premise}
Hypothesis: {hypothesis}

Your final probability estimate should be a value in the range [0,1], as fine-grained as
     possible, and formatted as follows: ```your_final_probability```.
For example, if the estimated probability is 0.0653, the output should be: ```0.0653```
```

## D.2 Reasoning Summary

```
system: You are a helpful assistant good at reasoning summary,
human: Simply **summarize** the reasoning process and final estimated probability (a
    float between 0 and 1). Include key factors, assumptions, and influences on the
    estimated probability, dismiss repetitive information.

Reasoning: {reasoning}

Begin your summary with ”Reasoning:” and end with ”Probability:”.
```

## D.3 0-shot Probability Scoring

```
system: You are a helpful assistant good at probabilistic reasoning in the real world
    setting.
human: Your task is to estimate the probability of a textual outcome $O$ given a
    description of the context $C$. Please respond with a probability in the range of
    [0, 1] that's your best estimate of the conditional probability P(O|C).

### Cotext
{context}

### Outcome
{outcome}

Wrap your final answer in triple quotes format. Please don't generate any other text.
```

## D.4 Judge with Confidence

```
system: You are a helpful assistant good at probabilistic reasoning in the real world
     setting
human: Define the Random Variables:
- Let \( H \) represent the hypothesis: {hypothesis}
- Let \( P \) represent the premise: {premise}

Below are the estimation of conditional probability P(H|P) based on the given premise
     and hypothesis from other agents using probabilistic reasoning and real-world
     knowledge.
Your task is to assign a confidence score (ranging from 0 to 1) to each agent's response.


1. {reasoning_1}

2. {reasoning_2}

3. {reasoning_3}

4. {reasoning_4}

Important Considerations:
- Think like a human, go beyond literal semantics by considering context, common sense,
     and real-world knowledge.
- Related premise and hypothesis do not necessarily cause high probability.
- Assign higher confidence to assumptions that are more commonly observed and reasoning
     processes that are logically sound, fully justified.
- If the premise and hypothesis both refer to an entity using an indefinite noun phrase
     (e.g., 'a person'), assume they refer to the same entity unless there is clear
     evidence suggesting otherwise.
- Errors may arise from disagreements in reasoning, differing assumptions, logical
     mistakes, overemphasis on corner cases, underconfidence or overconfidence, and
     inaccuracies in estimating P(H|P). Do not blindly trust other agents' probability
     estimation.

Here are some examples of probability extimation:
1. hypothesis: Three brothers pound on some drums
   premise: Three men dressed in white shirts and white hats, (two with baseball caps,
       the leader with a white construction helmet), pounding sticks on steel and
       plastic drums.
   probability: 0.00000027
2. hypothesis: There is a rock currently skipping down a pond.
   premise: A young african boy skipping rocks.
   probability: 0.058
3. hypothesis: The man is walking into a room.
   premise: A man is standing in the doorway of a building.
   probability: 0.2639
4. hypothesis: People are rollerblading for something to do.
   premise: At least six individuals are on a team wearing helmets and knee pads while
       rollerblading around a skating rink.
   probability: 0.5
5. hypothesis: A brown dog is outside and it's snowing
   premise: A brown dog plays in a deep pile of snow.
   probability: 0.7342
6. hypothesis: Two girls attend a convention.
   premise: Two girls in a crowd are dressed up, one as the cartoon character Wall-E.
   probability: 0.94
7. hypothesis: Some kids splash in the water and interact with each other.
   premise: many children play in the water.
   probability: 0.99

Output Format:
- The confidence score for other agents should be a decimal value between 0 and 1,
     formatted as: \\boxed{{confidence1, confidence2,confidence3,confidence4}}
- Example output: \\boxed{{0.1,0.5,0.8,0.2}}
```

### D.5 Pairwise Compare

```
system: You are a helpful assistant,
human: Given the premise and hypothesis of two natural inference passages, determine
    which hypothesis is more likely based on its premise.

Specifically:
1. Contextual Assessment with World Knowledge
Analyze each pair: Evaluate the premise and hypothesis using relevant world knowledge.
    Consider contextual factors such as demographics, common practices, or statistical
    distributions to estimate the likelihood of the hypothesis being true. State
    assumptions: Explicitly identify any assumptions or uncertainties introduced by
    missing information in the premise.
2. Comparison
Compare the likelihood of each hypothesis based on the alignment between the premise and
     hypothesis.
Justify your reasoning for why one hypothesis is more likely than the other, considering
     the degree of alignment and the assumptions made.
If the likelihoods of both hypotheses are sufficiently close or indistinguishable,
    return a None.

Passage A: {item1}
Passage B: {item2}

3. Output Format Example:
In your final decision, strictly output \boxed{{Passage A}}, \boxed{{Passage B}} or \
    boxed{{None}}”)])
```

### D.6 EntailmentBank, e-CARE Rewrite

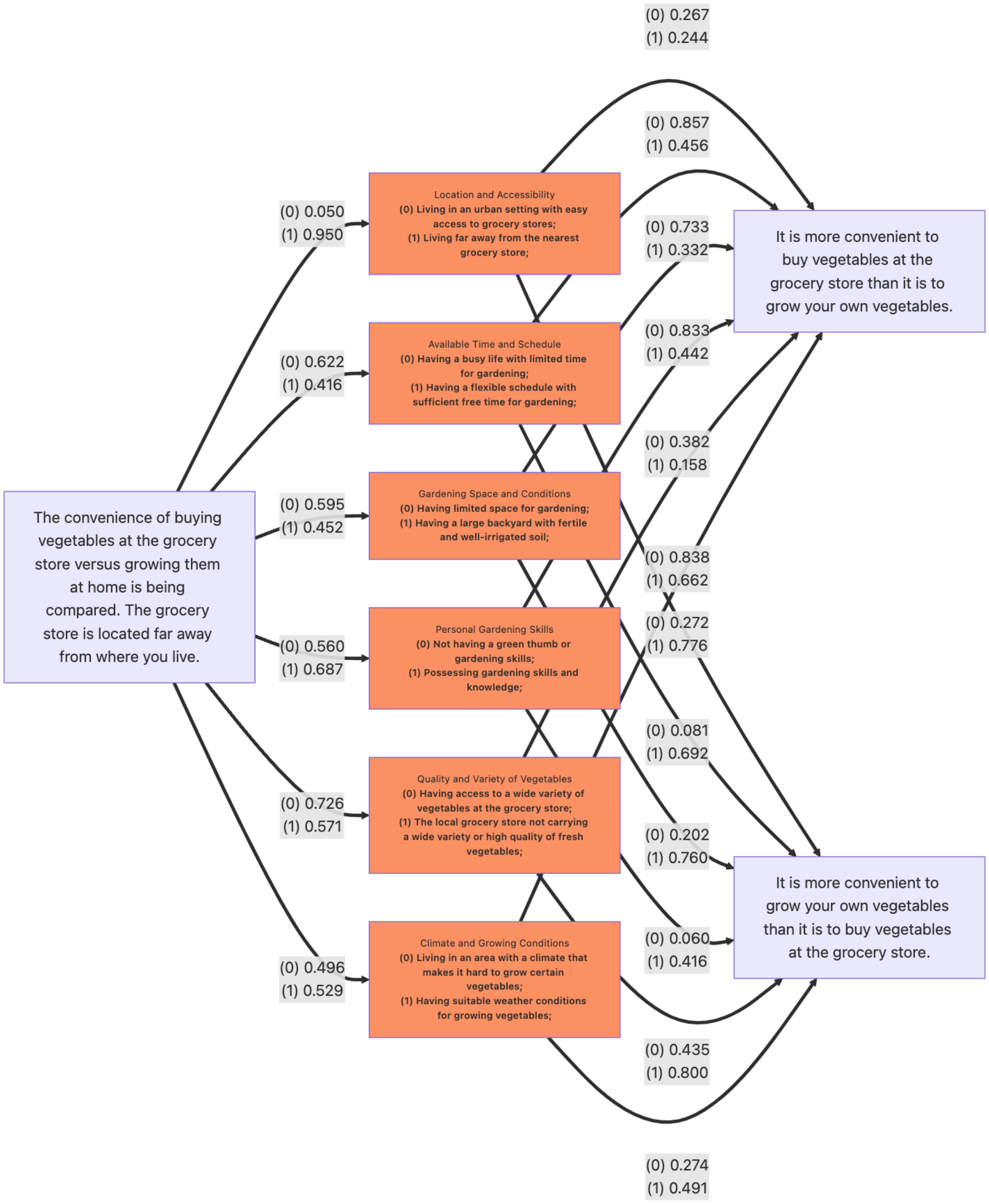
```
system: You are a helpful assistant”),
human: Given a natural language inferenc passage: {passage}
Your goal:
Rewrite the original premise and hypothesis
Generate 2 new premises related to the passage so that the probability of NLI P(
    hypothesis | new premise1, new premise2) ranges from **0.05 to 0.95**.

Steps to follow:
1. Rewrite the original premise and hypothesis for clarity and precision
- Ensure both the premise and hypothesis are clear, precise, and logically sound.
- Removed unnecessary modal verbs (e.g., ”can,” ”might”) and hedging language (e.g., ”
    possibly,” ”somewhat”).
- If needed, specify a concrete example for clarity.
2. Generate new premises that modify the likelihood of the hypothesis being inferred:
- Ensure all generated premises are factually correct and logically consistent.
- Here are strategies you may consider to adjust the probability of inference:
- Alternative Explanation (Misattribution): Provide a different cause for the phenomenon
     , weakening or shifting inference.
- Increase Vagueness: Make the premise more general, requiring additional inference.
- Observer-Dependent Effects: Frame the premise in a way that makes the inference more
    subjective or situational.
- Instantiation: Provide a specific example that either supports or challenges the
    inference.
3. Categorize Premises into Four Bins Based on Probability
- highly likely(probability~0.9): Premises that strongly support the hypothesis but may
    introduce slight variation or broader interpretations.
- moderately likely(probability~0.7): Premises that are related to the passage but are
    more general, potentially requiring additional context to confirm the hypothesis.
- neutral(probability~0.5): Balances support and doubt, making the inference uncertain.
- unlikely (probability~0.3): Premises that introduce alternative mechanisms or are only
     tangentially related to the hypothesis.
- contradict(probability~0.1): Challenges the hypothesis by shifting the explanation,
    observer perspective, or causal factorwithout introducing factual errors.
```

```
4. Format the output as a valid JSON object with the following structure:
{{
  ”premise”: ”Your revised premise here.”,
  ”hypothesis”: ”Your revised hypothesis here.”,
  ”highly likely”: [premise1, premise2],
  ”moderately likely”: [premise1, premise2],
  ”neutral”: [premise1, premise2],
  ”unlikely”: [premise1, premise2],
  ”contradict”: [premise1, premise2]
}}
5. Recheck that output statements are factually accurate and format is a valid JSON.
```

## E Example Structural Reasoning Traces

In this section, we provide example traces constructed to evaluate local scoring models. The example reasoning trace corresponds to an instance from the BIRD dataset for C2S-Sent-B (Feng et al., 2024). The block on the left describe a situation with an additional sentence supporting an outcome, and the block on the right describes potential outcomes. In this case BIRD infers the outcome the additional sentence support by reasoning over a range of factors and conditions probabilistically. The scores produced by our model are shown in gray blocks. Note that we do not perform NLI-based edge filtering, and the scores are not coherent.


The convenience of buying vegetables at the grocery store versus growing them at home is being compared. The grocery store is located far away from where you live.
(0) 0.050
(1) 0.950
(0) 0.622
(1) 0.416
(0) 0.595
(1) 0.452
(0) 0.560
(1) 0.687
(0) 0.726
(1) 0.571
(0) 0.496
(1) 0.529
Location and Accessibility
(0) Living in an urban setting with easy access to grocery stores;
(1) Living far away from the nearest grocery store;
Available Time and Schedule
(0) Having a busy life with limited time for gardening;
(1) Having a flexible schedule with sufficient free time for gardening;
Gardening Space and Conditions
(0) Having limited space for gardening;
(1) Having a large backyard with fertile and well-irrigated soil;
Personal Gardening Skills
(0) Not having a green thumb or gardening skills;
(1) Possessing gardening skills and knowledge;
Quality and Variety of Vegetables
(0) Having access to a wide variety of vegetables at the grocery store;
(1) The local grocery store not carrying a wide variety or high quality of fresh vegetables;
Climate and Growing Conditions
(0) Living in an area with a climate that makes it hard to grow certain vegetables;
(1) Having suitable weather conditions for growing vegetables;
(0) 0.267
(1) 0.244
(0) 0.857
(1) 0.456
(0) 0.733
(1) 0.332
(0) 0.833
(1) 0.442
(0) 0.382
(1) 0.158
(0) 0.838
(1) 0.662
(0) 0.272
(1) 0.776
(0) 0.081
(1) 0.692
(0) 0.202
(1) 0.760
(0) 0.060
(1) 0.416
(0) 0.435
(1) 0.800
(0) 0.274
(1) 0.491
It is more convenient to buy vegetables at the grocery store than it is to grow your own vegetables.
It is more convenient to grow your own vegetables than it is to buy vegetables at the grocery store.