In-Context Quality Scoring: A Method to Evaluate LLMs without Humans

Wenwen Si [1] **Kaustubh Sridhar** [1] **Insup Lee** [1]

Abstract

Evaluating large language models (LLMs) is a critical yet challenging task due to the open-ended nature of their outputs. While human evaluation remains the gold standard, its cost and scalability limitations have driven increased reliance on automated LLM-based assessments. In this work, we introduce In-Context Quality Scoring (ICQS), a novel method that leverages structured in-context learning to rank language model outputs with minimal supervision. By framing the evaluation process as an approximation of Bayesian posterior inference, ICQS efficiently estimates the quality of LLM-generated outputs without relying on extensive labeled data. We demonstrate that ICQS exhibits semantic generalization, surpassing standard in-context learning and LLM verbalized scoring baselines in quality ranking. ICQS' epistemic variant further enhances stability and consistency. Empirical results on diverse evaluation benchmarks–sentiment analysis, natural language inference, and creative writing–validate ICQS as an effective and scalable solution for LLM evaluation.

1. Introduction

Evaluation plays a critical role in machine learning, serving as a key element in both model development and selection. This is especially true in the era of large language models (LLMs), where model architectures have become increasingly homogeneous, making evaluation—including both quantitative benchmarks and qualitative assessments—a central driver in developing more capable and aligned models. However, evaluating LLMs presents unique challenges, primarily due to the open-ended nature of their outputs. While human evaluation remains the gold standard, its high cost and time-intensive nature limit its scalability in fast-paced development environments. As a result, advanced LLMs like ChatGPT have increasingly been used as viable alternatives to human annotators (Zheng et al., 2023), alongside web-based information retrieval.

In-context learning (ICL) (Brown, 2020; Garg et al., 2022; Fu et al., 2023; Lee et al., 2024; Wies et al., 2024; Agarwal et al., 2024) is a remarkable phenomenon that has emerged in transformer models with billions of parameters. Without modifying model weights, large transformers such as LLMs can adapt to new downstream tasks in a training-free manner by incorporating prefix demonstration examples directly into their inputs. ICL has demonstrated its effectiveness across various applications, including uncertainty quantification (Hou et al., 2023; Tanneru et al., 2024; Yadkori et al., 2024; Liu et al., 2024a), jailbreak scenarios (Wei et al., 2023), reinforcement learning (Lee et al., 2024; Kirsch et al., 2023; Dai et al., 2024; Grigsby et al., 2023), and imitation learning (Sridhar et al., 2024; Raparthy et al., 2023).

In this work, we propose a general quality scoring method based on in-context learning (ICL), designed for scenarios where no external information, such as retrieval, is available. Our method is able to evaluate a given set of answers by ranking their quality, requiring only a few in-context examples. This lightweight approach makes it applicable to a wide range of tasks without reliance on additional data labelers. Moreover, our method demonstrates strong generalization capabilities, both in terms of semantic understanding and across different tasks. This approach highlights a promising application of ICL, providing a novel avenue for evaluating language model generations.

Several prior studies have examined the learnability (Wies et al., 2024) and theoretical behavior (Falck et al., 2024) of ICL, often framing it as an (approximate) Bayesian inference problem. In this view, ICL is typically interpreted as inferring latent structure from observed examples, akin to Bayesian posterior updating. However, recent work by (Bigelow et al., 2023) suggests that the dynamics of ICL align more closely with model selection rather than model averaging, indicating that ICL tends to favor a single hypothesis consistent with the observed data rather than integrating multiple possible explanations. This observation provides a simplified modeling of in-context learning and helps us simplify our framework.

[1]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, Pennsylvania. Correspondence to: Wenwen Si <wenwens@seas.upenn.edu>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

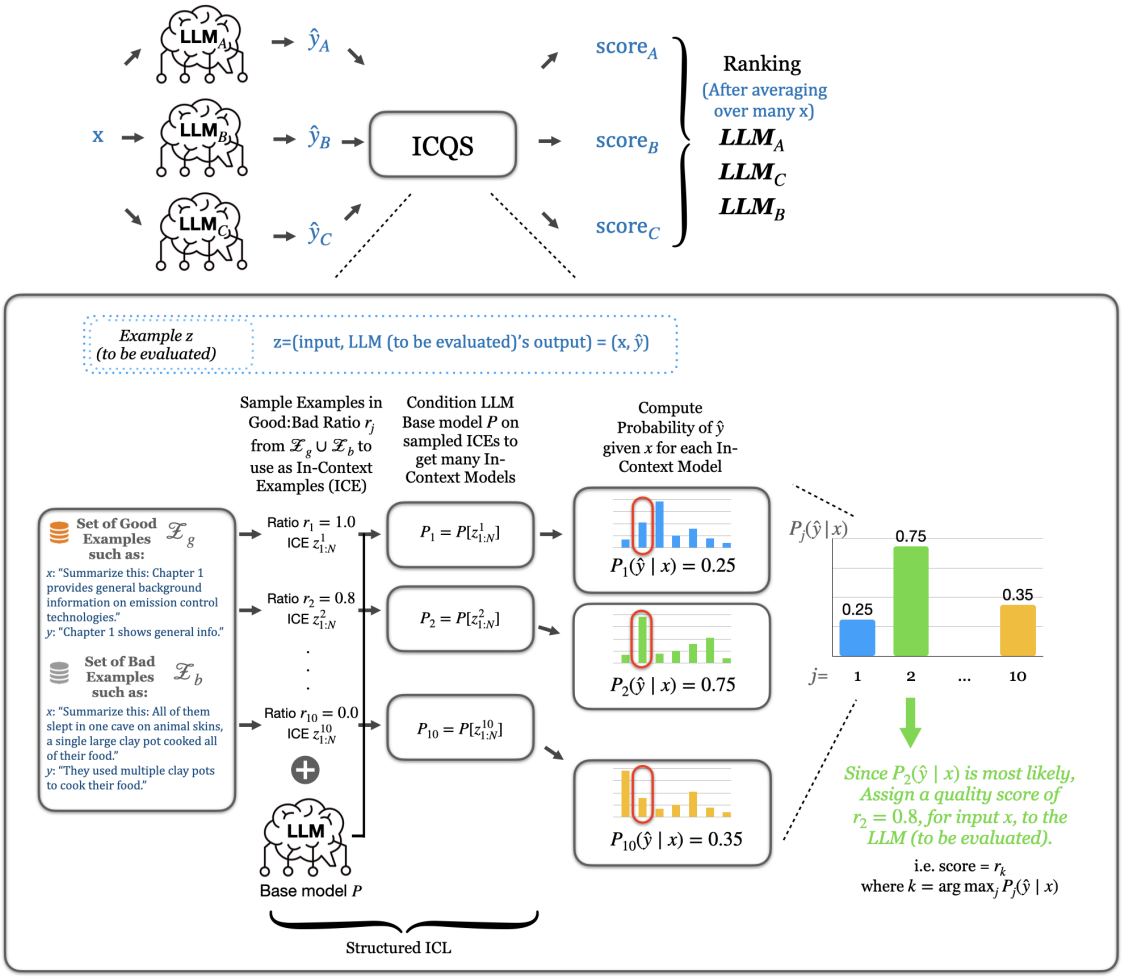


Figure 1. Framework of In-Context Quality Scoring (ICQS), which consists of: (1) constructing M in-context demonstration sets with varying mixture ratios from $\mathcal{Z}_g \cup \mathcal{Z}_b$; (2) conditioning an ICL base model P_{base} on each set to form M induced distributions P_j; (3) computing the log-likelihood of the evaluation instance under each P_j; (4) selecting the distribution P_k that maximizes likelihood and assigning its mixture ratio r_k as the final quality score.

Furthermore, (Liu et al., 2024b) investigates the geometric properties of in-context learning and verifies that Llama-series LLMs follow structured, low-dimensional semantic trajectories. Their findings provide a formal characterization of the geometric structure underlying in-context learning-induced probabilities. This suggests that ICL models constructed with different configurations (e.g., size, mixture ratios) of ICEs differ mainly in semantic quality rather than in irrelevant noise. We leverage this observation for constructing in-context hypothesis space in Bayesian inference.

Several LLM evaluation benchmarks have been developed to assess LLM performance, particularly in terms of alignment with human preferences. Compared to traditional ground truth-based NLP datasets, these benchmarks focus on evaluating how well models align with human judgment, especially in tasks involving free-form generation. A recent trend involves leveraging GPT-4 as a proxy for human judgment, as demonstrated by notable benchmarks such as MT-Bench (Zheng et al., 2023) and AlpacaEval (Li et al., 2023). On the other hand, Chatbot Arena (Chiang et al., 2024) stands out as the first open, large-scale, crowdsourced benchmark platform that utilizes real-time human interaction for evaluation. This shift toward more dynamic and interactive evaluation methods facilitates more effective and scalable human-aligned assessments in model development.

In this paper, we frame the evaluation of language modeling as measuring the distance between a probabilistic model of interest and the underlying ground truth distribution of natural language. However, traditional statistical metrics, such as KL divergence, are often computationally intensive for complex probabilistic models used in natural language processing and typically require large amounts of labeled ground truth data. To address this, we propose In-Context Quality Scoring (ICQS), which leverages structural in-context learning to derive tractable, representative heuristic statistics.

Unlike traditional i.i.d. sampling of good in-context examples, our structural method samples in-context demonstrations from a mixture of high- and low-quality examples in varying proportions, thereby creating a hypothesis space of projected probabilistic distributions across different quality levels. In this way, a generation's quality can be captured by the most likely hypothesis within this space—which is effectively reflected by the quality of the chosen demonstrations (Xie et al., 2022). Due to the inherent randomness of in-context learning, ICQS naturally favors aggregated evaluation, making it well-suited for tasks such as LLM evaluation across multiple questions. Additionally, we extend ICQS with an epistemic enhancement to mitigate the variance introduced by in-context example selection, ensuring more stable performance. We call this ICQS-Epistemic. ICQS enables quality rankings that align better with human preference while requiring only a few labeled in-context examples. Furthermore, it demonstrates promising semantic and task generalization capabilities, making it an effective tool for evaluating large language models. Our contributions are outlined as follows.

Contributions We introduce In-Context Quality Scoring (ICQS), a novel method for automatically evaluating language model quality that leverages structural in-context learning. We demonstrate its effectiveness across diverse tasks, including sentiment analysis, fine-grained natural language inference, and creative writing within the LMSYS benchmark. Our results indicate ICQS' quality ranking aligns better with human preference, compared to standard in-context learning and verbalized LLM scoring (Zheng et al., 2023). Further, ICQS with an epistemic enhancement ensures notable stability. In addition, our results highlight the semantic generalization properties of structured in-context learning, showing how mixed in-context examples from the extremes form the complete intermediate spectrum of in-context models with varying quality.

2. Problem Formulation

We address the general problem of evaluating the quality of specific language model generations. For a task Q, we define two probability distributions: P_g, representing the desired "good" distribution of correct or high-quality outputs, and P_b, representing the undesired "bad" language distribution, both aligned with human natural language. For example, P_g may correspond to a set of gold-standard ground truth labels, while P_b includes undesired outputs, such as flipped answers in classification tasks or creative writing samples that fail to adhere to the given instructions.

Consider a probabilistic model of interest with distribution P_z, such as a large language model (LLM). We sample responses $z \sim P_z$, and and seek to evaluate their quality effectively.

Definition 2.1 (α-Quality Distribution)**.** Let $(\Omega, \mathcal{F})$ be a measurable space, and let $\mathcal{P}(\Omega)$ denote the space of all probability measures defined on $(\Omega, \mathcal{F})$. Consider a metric

$$d : \mathcal{P}(\Omega) \times \mathcal{P}(\Omega) \to \mathbb{R}_{\geq 0}$$

that measures the distance between probability distributions. Let the distance between the two predefined extreme distributions $P_b \in \mathcal{P}(\Omega)$ and $P_g \in \mathcal{P}(\Omega)$ be denoted by

$$D = d(P_b, P_g).$$

For any distribution $P_z \in \mathcal{P}(\Omega)$, we say that P_z is an α-*quality distribution* if for a scalar $\alpha \in [0, 1]$,

$$d(P_b, P_z) = \alpha D.$$

In this formulation, We quantify the quality of P_z using a scalar $\alpha \in [0, 1]$ such that:

- $\alpha = 1$ corresponds to the highest-quality distribution ($P_z = P_g$).
- $\alpha = 0$ corresponds to the lowest-quality distribution ($P_z = P_b$).
- Intermediate values of α represent distributions of varying quality between P_b and P_g, where $d(P_b, P_z)$ measures how close P_z is to the high-quality distribution P_g.

Formally, our objective is to construct a quality scoring function $r(z)$ for responses z such that it preserves the ordering of α under some reasonable distance metrics d. That is, for any two distributions P_{z_1} and P_{z_2} with corresponding quality parameters α_1 and α_2, we require:

$$\alpha_1 \geq \alpha_2 \Rightarrow r(z_1) \geq r(z_2).$$

Secondly, we aim to achieve this ranking while relying on a limited number of labeled examples. In practice, we do not have direct access to P_g and P_b but instead observe small sample sets:

$$\mathcal{Z}_g = \{z_i^g \sim P_g \mid i = 1, \ldots, m\},$$
$$\mathcal{Z}_b = \{z_i^b \sim P_b \mid i = 1, \ldots, n\}.$$

In this way, we estimate α empirically by constructing a rank-preserving function r. The challenge is to design $r(z)$ so that it effectively ranks responses while leveraging the limited labeled data available and generalizing beyond it.

3. In-Context Quality Scoring

In this section, we describe our approach to designing the quality scoring function r. Classical probabilistic distances, such as KL divergence, are computationally expensive (Yadkori et al., 2024), making the direct estimation of α infeasible without a large amount of ground truth labeling.

To address this challenge, we propose using the *in-context sample mixture ratio* as a proxy for constructing r. Intuitively, we first present the quality evaluation problem, which estimates how much a generating distribution differs from the extreme distributions (P_g, P_b). We then show that this problem can be approximately formulated as a in-context mixture ratio problem under the Bayesian inference framework. In the following sections, we will systematically demonstrate why this approach could be a plausible solution. The validity of the proposed mixture ratio will be verified through empirical experiments.

3.1. In-Context Learning as Bayesian Inference

Recent work (Falck et al., 2024; Bigelow et al., 2023) has framed in-context learning as Bayesian inference over the probabilistic space. Given in-context examples $z_{1:N}$ and a new sample z, the predictive distribution $p(z|z_{1:N})$ follows a Bayesian posterior predictive distribution over the probability space $\mathcal{P}_\alpha = \{P_\alpha : d(P_b, \cdot) = \alpha D\}$:

$$p(z|z_{1:N}) = \int_{\alpha\in[0,1]} p(z|P_\alpha)p(P_\alpha|z_{1:N})d\alpha. \quad (1)$$

This formulation suggests that in-context learning implicitly performs Bayesian inference within $\mathcal{P}_\alpha$ by averaging over possible model parameters.

Interestingly, (Bigelow et al., 2023) provided evidence that ICL on large transformers often perform Bayesian model selection rather than model averaging. That is, instead of marginalizing over α, they tend to select a single latent parameter that best explains the observed context:

$$P_{\alpha_0} = \arg\max_\alpha p(P_\alpha|z_{1:N}), \quad p(z|z_{1:N}) = p(z|P_{\alpha_0}). \quad (2)$$

We adopt Eq. (2) in our framework for its simplicity, allowing us to estimate a single α_0 that best represents the quality of P_z.

3.2. α as the Oracle Quality

We define the oracle quality α based on the performance risk (Wei et al., 2023), which quantifies the effectiveness of a model P_α. Given a prompt $x \sim Q$, the expected reward (e.g. quality) $\mathcal{R}$ of a model P is given by:

$$\mathcal{R}_P(x) = \mathbb{E}_{y\sim P(\cdot|x)}R(y).$$

For $z = (x, a_z)$ collected from model P_z, we approximate its quality by considering its likelihood under a proxy model space $P_{\alpha_j} \in \mathcal{P}_\alpha$, realized via a finite set of in-context samples $\{z^j_{1:N}\}_{j=1}^M$. By Bayes' rule:

$$p(P_{\alpha_j}|z) \propto p(z \mid P_{\alpha_j})p(P_{\alpha_j}).$$

Thus, we estimate $\hat{P}_z$ as:

$$\hat{P}_z = \arg\max_j p(z \mid P_{\alpha_j}), \quad (3)$$

Algorithm 1 ICQS: In-context quality scoring for a single evaluation sequence z using base LLM P_{base} given sets of good and bad answers $\mathcal{Z}_g, \mathcal{Z}_b$.

function ICQS($\mathcal{Z}_g, \mathcal{Z}_b, P_{\text{base}}, z$)
  for $j \in \{0, 1, \ldots, M\}$ **do**
    $r_j \leftarrow \frac{j}{M}$
    $P_j \leftarrow$ ConstructICLModel($\mathcal{Z}_g, \mathcal{Z}_b, r_j$) {*Obtain an in-context learned model using mixed demonstration sets.*}
    Compute $\ell_j \leftarrow \log P_j(\hat{y} \mid x)$ {*Compute the likelihood of the j-th model generating the answer $\hat{y}$.*}
  end for
  $k \leftarrow \arg\min_k \ell_k(x, \hat{y})$ {*Find the most likely ICL model*}
  Output r_k {*Return its sampling ratio as the score.*}
end function
function ConstructICLModel($\mathcal{Z}_g, \mathcal{Z}_b, r_j$)
  Sampling in-context examples with mixed ratio:
  for $i \in \{1, \ldots, N\}$ **do**
    $z_i^j \sim \mathcal{Z}_g$ with probability $1 - r_j$, else $z_i^j \sim \mathcal{Z}_b$
  end for
  $P_j \leftarrow P_{\text{base}}[z^j_{1:N}]$
  Output P_j
end function

with the corresponding α_{j_0} providing the best likelihood estimate of z. By definition, the probabilistic distance is:

$$d(P_b, P_z) = \sup_{x\sim Q} |\mathcal{R}_{P_b}(x) - \mathcal{R}_{P_z}(x)| = \alpha_{j_0} D. \quad (4)$$

3.3. Mixed Sampling Ratio Preserves the Order of α

To realizable the models in the hypothesis space, we construct candidate model distributions using different sampling ratios from $\mathcal{Z}_b \cup \mathcal{Z}_g$. The in-context sampling ratio r_0 then serves as a proxy for quality metrics.

Formally, assume in-context examples are drawn as:

$$\text{ICE}_r = \{x_1, y_1, \ldots, x_N, y_N \mid (x_i, y_i) \sim rP_g + (1-r)P_b\}.$$

Under reasonable assumptions from (Wei et al., 2023) (Section E), we derive:

$$\begin{aligned}|\mathcal{R}_{P_{\alpha(x)}} - \mathcal{R}_{P_b}(x)| &= |\mathcal{R}_P([\text{ICE}_r, x]) - \mathcal{R}_{P_b}(x)| \\ &\le C \cdot \frac{P_g([\text{ICE}_r, x])}{P_b([\text{ICE}_r, x])} = \alpha D.\end{aligned}$$

Now, denote sample sets:

$$S_g = \{i \mid (y_i, x_i) \overset{\text{i.i.d.}}{\sim} \mathcal{Z}_g\}, \quad S_b = \{i \mid (y_i, x_i) \overset{\text{i.i.d.}}{\sim} \mathcal{Z}_b\}.$$

Then, there exists some constant $\Delta > 1$ such that for all $i \in S_g$, we have $\frac{P_g(y_i|x_i)}{P_b(y_i|x_i)} > \Delta$. By expanding the likelihood

ratio:

$$\frac{P_g([\text{ICE}_r, x])}{P_b([\text{ICE}_r, x])} \propto \prod_{i=1}^{k} \frac{P_g(y_i \mid x_i)}{P_b(y_i \mid x_i)} \tag{5}$$
$$= \prod_{i \in S_g} \frac{P_g(y_i \mid x_i)}{P_b(y_i \mid x_i)} \cdot \prod_{i \in S_b} \frac{P_g(y_i \mid x_i)}{P_b(y_i \mid x_i)}.$$

Thus, for two different sampling ratios $r_i > r_j$, we obtain:

$$\frac{\alpha_i}{\alpha_j} = \frac{P_g([\text{ICE}_{r_i}, x])}{P_b([\text{ICE}_{r_i}, x])} \Big/ \frac{P_g([\text{ICE}_{r_j}, x])}{P_b([\text{ICE}_{r_j}, x])}$$
$$= \prod_{i \in S_{g_i} \setminus S_{g_j}} \frac{P_g(y_i \mid x_i)}{P_b(y_i \mid x_i)} \Big/ \prod_{i \in S_{b_j} \setminus S_{b_i}} \frac{P_g(y_i \mid x_i)}{P_b(y_i \mid x_i)}$$
$$> \Delta^{|S_{g_i} \setminus S_{g_j}|} > 1.$$

This result implies that $r_i \geq r_j \iff \alpha_i \geq \alpha_j$.

Thus, the ordering of r preserves the ordering of α, making it a valid quality score.

Remark 3.1. *In sample-mixed LLM in-context learning, the sampling ratio r serves as an effective approximate statistic for model quality evaluation.*

3.4. Overall Algorithm

We apply an ICL base model P_{base} for the evaluation. We then construct N in-context examples from the dataset $\mathcal{Z}_g \cup \mathcal{Z}_b$ using a mixture ratio r_j, forming the sets $\{z^j_{1:N}\}$ for $j = 1, \ldots, M$.[1]

Next, we use the M sets to create M distributions conditioning on P_{base} denoted as

$$P_j := P_{\text{base}}[z^j_1, ..., z^j_N].$$

For $z = (x, \hat{y})$, we compute the log likelihood

$$\ell_j(x, \hat{y}) = \log P_j(\hat{y} \mid x).$$

Finally, we find P_k that obtains the maximum likelihood, where $k = \arg\max_k \ell_k(x, \hat{y})$. In this case, we assign r_k as the score of the eval sequence z. The details of the ICQS algorithm are formally outlined in Algorithm 1.

4. Epistemic In-Context Quality Scoring

In the previous section, we introduced the ICQS method based on the recent in-context learning theories. However, in practice, the performance of in-context learning can exhibit significant variance across different demonstration sets. This inherent randomness can degrade the effectiveness of the ICQS method.

[1]One may question the number of mixture ratios M required to obtain reliable results. We discuss this further in Appendix D.

Algorithm 2 ICQS-Epistemic: In-context quality scoring with epistemic uncertainty mitigation.

function ICQS-Epistemic($\mathcal{Z}_g, \mathcal{Z}_b, P_{\text{base}}, z, L$)
  for $j \in \{0, 1, \ldots, M\}$ **do**
    $r_j \leftarrow \frac{j}{M}$
    Iterate over L different demonstration sets:
    for $l \in \{1, \ldots, L\}$ **do**
      $P^l_j \leftarrow$ ConstructICLModel($\mathcal{Z}_g, \mathcal{Z}_b, r_j$)
      Compute $\ell^l_j \leftarrow \log P^l_j(\hat{y} \mid x)$
    end for
    Compute epistemic-smoothed score:
    $\bar{\ell}_j \leftarrow \frac{1}{L} \sum_{l=1}^{L} \ell^l_j$
  end for
  $k \leftarrow \arg\max_k \bar{\ell}_k(x, \hat{y})$ {*Obtain quality score on the epistemic likelihood.*}
  Output r_k
end function

In particular, the likelihood should be decomposed into two components: the epistemic component that captures the model's inherent uncertainty in generating specific answers, and aleatoric components, which arises from variability due to different demonstration set selections. To address this issue, we propose an epistemic-enhanced version of our algorithm following a previous work in ICL uncertainty decomposition (Ling et al., 2024).

4.1. Uncertainty Decomposition of In-context Learning

In ICL, the predictive distribution for generating $\hat{y}$ given a set of few-shot demonstrations $z_{1:N}$ and a test input x can be denoted as: $p(\hat{y} \mid \Theta, z_{1:N}, x)$, where we explicitly denote the base model parameters as Θ.

Let $\ell(\Theta) = P(\hat{y} \mid x, z_{1:N}, \Theta)$ represents the overall likelihood of the predictive distribution. To estimate the epistemic component, we condition on a fixed base model, effectively omitting Θ as $\ell = p(\hat{y} \mid z_{1:N}, x)$. The expectation of ℓ over L set of demonstrations then serves as a metric for the epistemic components of the likelihood.

Formally, (Ling et al., 2024) constructs a matrix $M \in \mathbb{R}^{|\mathcal{Y}| \times L}$ for close-form tasks with $|\mathcal{Y}|$ predefined answer options. The epistemic component is estimated as:

$$\text{EU} = \frac{1}{L} \sum_{l} P(\sigma(M_{:,l})),$$

where $\sigma(\cdot)$ denotes the softmax function, and $M_{:,l}$ represents the l-th column of the matrix M. This approximation encapsulates the variability introduced by different demonstration configurations.

4.2. ICQS Epistemic Approximation

In the ICQS setting, we generate M sets of mixed in-context examples (ICE) with varying mixture ratios r_j, denoted as $z^j_{1:N}, j = 1, \ldots, M$. Specifically, each set $z^j_{1:N}$ with ratio the r_j undergoes an ICL epistemic decomposition process. In this way, for each r_j, we proceed L independent iterations of ICE sampling from $\mathcal{Z}_g \cup \mathcal{Z}_b$, acquiring matrix M^j for epistemic estimation.

For the open-ended generation tasks, we estimate the epistemic component by direct averaging of the logit values over L iterations. A comprehensive breakdown of ICQS-Epistemic algorithm is provided in Algorithm 2.

5. Experiments

5.1. Experimental Setup

We study the performance of our proposed method on four natural language question-answering tasks: sentiment analysis (SST-2 (Socher et al., 2013), Financial-PhraseBank (Malo et al., 2014)), fine-grained natural language inference (ChaosNLI (Nie et al., 2020)), and creative writing (LMSYS Jokes (Zheng et al., 2024)). We compare our method with two baselines that also do not rely on information retrieval or iterative search:

- *Verbalized LLM Evaluation (or LLM-as-a-judge (Zheng et al., 2023)):* where we directly request a verbalized rating or the predicted labelings from the base LLM.
- *In-context learning:* where we apply naive in-context learning on the base model, then evaluate on the the predicted labelings.

Next, we briefly describe our problem domains. The tasks are presented in increasing order of difficulty.

- **Sentiment Analysis (SST-2):** Binary sentiment classification on sentences extracted from movie reviews. *Task:* Evaluate the semantic accuracy of the sentiment. *Data:* (sentence, label) pairs. *Metric:* Accuracy with respect to the ground-truth sentiment class. *Base LM:* LLaMA-2-7B-8bit (Touvron et al., 2023). *Format:* $\{x : \textit{movie review}, y : \textit{sentiment}\}$. This dataset is mainly used in primary verification for its simplicity.
- **Sentiment Analysis (Financial-PhraseBank or FB):** Polar sentiment classification of sentences from financial news, predicting their influence on the financial market. *Task:* Evaluate the semantic accuracy of the sentiment. *Data:* (sentence, label) pairs. *Metric:* Sentiment distance with respect to the ground-truth sentiment class. *Base LM:* LLaMA-2-7B-8bit. *Format:* $\{x : \textit{financial news}, y : \textit{sentiment}\}$.

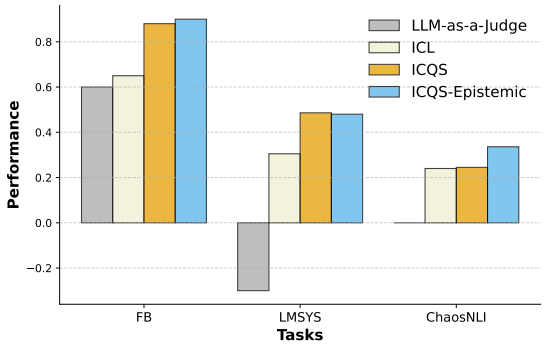


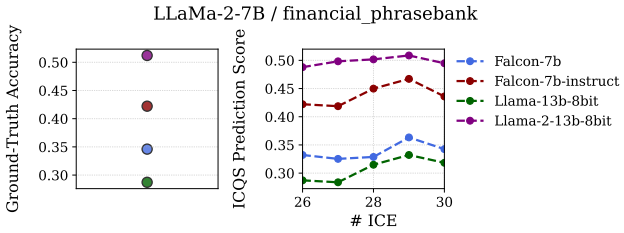
Figure 2. Overall comparison of ICQS and baseline methods across tasks and metric scores. We present the best performance of each method across tasks.

- **Creative Writing (LMSYS Jokes):** Generate creative writings based on human requests for jokes. *Task:* We filter a subset of the LMSYS dataset in which the prompt includes "joke". Evaluate LLM's generation performance by aggregating human preference scores. *Data:* A tuple of (question, model_a_answer, model_b_answer, winner). *Metric:* Relative rankings of all models involved in the LMSYS leaderboard. *Base LM:* LLaMA-3-70B (Dubey et al., 2024) . *Format:* $\{x : \textit{question}, A : \textit{model\_a\_answer}; B : \textit{model\_b\_answer}, y : \textit{winner}\}$.
- **Natural Language Inference (ChaosNLI):** A fine-grained NLI problem that compares model distribution outputs with human label distributions. *Task:* Predict distributional labels by evaluating the quality of the hypothesis. *Data:* A tuple of (premise, hypothesis, percentage of contradiction, neutral, and entailment ϕ). *Metric:* Mean Absolute Error (MAE) and relative ranking on the labeling distributions. *Base LM:* LLaMA-3-70B. *Format:* $\{x : \textit{premise}, y : \textit{hypothesis}\}$.

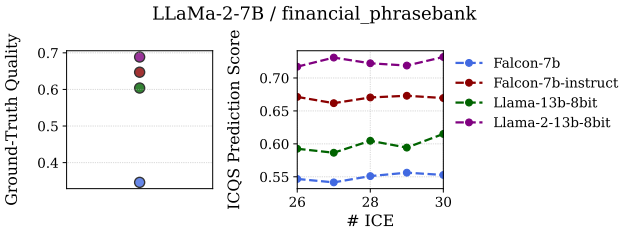
For all tasks, fewer than 60 good/bad examples were provided, from which we independently sample in-context examples. All examples belong to the two extreme ground-truth labels (e.g., positive and negative). To evaluate the semantic generalization properties of ICQS, we introduce additional quality categories in the test set, such as the neutral category in sentiment analysis.[2] We provide prompt examples in Appendix C.

Is r a valid statistic for estimating the quality of the answers? First, we verify the connection between the sampling ratio r and quality. The experiments are conducted with synthetic labels on two sentiment analysis

[2]For the ChaosNLI dataset, we provide polarity-labeled examples but evaluate on all distributional-labeled examples. For the LMSYS benchmark, we include the tie category in the test set.



(a) ICQS successfully recovered the ground truth ranking of the binary scoring. Left: Ground truth accuracy of the four LLMs. Right: ICQS prediction scores for the four LLMs, along with the trend across different demonstration sizes.



(b) ICQS successfully recovered the ground truth ranking of the semantic scoring. Left: Ground truth semantic quality of the four LLMs. Right: ICQS prediction scores for the four LLMs, along with the trend across different demonstration sizes.

Figure 3. ICQS perfectly recovered the ground truth ranking of LLMs on the financial phrasebank dataset.

datasets—SST-2 and Financial Phrasebank.

We provide the results in Appendix A. In particular, Figures 5-6 illustrate the connection between r and extreme-quality answers. Figure 7-8 demonstrates ICQS's semantic generalization ability for middle-quality answers. Moreover, Table 4 presents quantitative results on quality accuracy, showing that ICQS outperforms baseline methods as a general quality scoring approach.

Can this performance be scaled to real LLM generations and complex tasks? Next, we aim to extend our evaluation to real-world quality assessment across different tasks. Based on Figure 2, we answer this question in the affirmative. Figure 2 shows an overall comparison of ICQS and baseline methods across tasks. Across different evaluation and metrics, we observe that ICQS (and its epistemic variant) consistently outperforms the baselines. We provide detailed results and discussions in the following paragraphs.

Can ICQS evaluate real LLM generations? We extend our experiments in the financial sentiment analysis task from synthetic label quality ranking to real LLM quality ranking. Four real LLMs are evaluated: Falcon-7B, Falcon-7B-Instruct, LLaMA-13B-8bit, and LLaMA-2-13B-8bit, each generating responses for 30 Financial PhraseBank questions. In this part, we define quality based on the exact label matching score (see an example in the following table).

GT Label	Prediction	Score
Positive	Positive	1
Positive	Neutral or Negative	0

The results in Figure 3a demonstrate that ICQS perfectly recovers the ground truth accuracy ranking of the four models.

Beyond label accuracy: can ICQS capture semantic qualities? In practice, answer quality is rarely binary. Instead, people assess quality based on semantic discrepancy, assigning a *semantic quality score*. Therefore, we follow the previous setting but introduce a semantic quality scoring system for the answers (see the following table).

GT Label	Prediction	Score
Positive	Positive	1
Positive	Neutral	0.5
Positive	Negative	0

Figure 3b shows that ICQS successfully recovers the ground truth LLM ranking under the semantic quality scores. Notably, we observe that the ranking of LLaMA-13B-8bit and Falcon-7B is flipped when switching to the semantic score, a change that is also perfectly captured by ICQS.

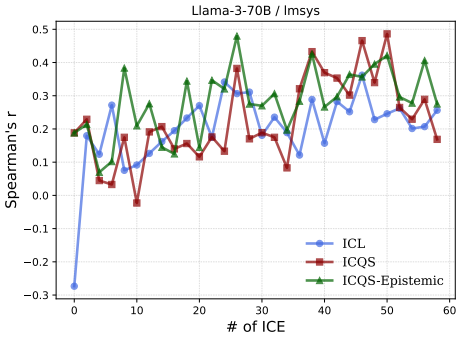
Can ICQS reflect human preferences on the noisy crowdsourced LMSYS benchmark? Pre-collected datasets are typically carefully processed. However, we are particularly interested in cases where there is insufficient labeled data for ground truth evaluation. The LMSYS benchmark serves as a good example, as it is a crowdsourced dataset containing noisy human prompts and preferences. Specifically, we applied our method to the creative writing subset of the LMSYS dataset—joke generation. This task is not only open-ended but also implicitly exhibits varying quality levels.

We extracted all joke-related questions from LMSYS, consisting of 480 samples with responses from 57 LLMs. We then split the dataset into 60 ICE sets and a 420-sample test set.

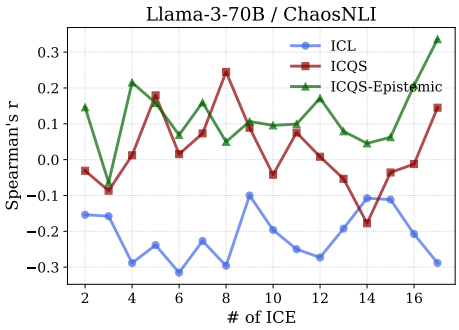
Due to the large number of LLMs, obtaining a perfect ranking is infeasible. Therefore, we apply a quantitative metric to evaluate model rankings: Spearman's rank correlation:

$$r_s = 1 - \frac{6 \sum d_i^2}{n\left(n^2 - 1\right)}. \tag{6}$$

Table 1 demonstrates that ICQS achieves an improved quality ranking of the models compared to ICL. The complete ranking list (of the 57 LLMs) output by ICQS, the ICL baseline, along with the ground truth from the LMSYS dataset is provided in Table 5 in Appendix B.



(a) Spearman's r for ICL, ICQS, and ICQS-Epistemic on LMSYS jokes. ICQS and ICQS-Epistemic stably outperform on larger ICEs.



(b) Spearman's r for ICL, ICQS and ICQS-Epistemic on ChaosNLI dataset.

Figure 4. ICQS achieves higher spearman's r for quality ranking.

Table 1. Comparison of Spearman's r for LMSYS dataset.

	Spearman's r ↑		
	ICL	ICQS	ICQS-Epistemic
Spearman's r ↑	0.305	**0.486**	0.480

Can ICQS maintain good performance on open-form generations? Finally, we demonstrate that ICQS has the potential to produce fine-grained distributional scoring for complex open-form generations, such as those in the ChaosNLI dataset. ChaosNLI is a natural language inference dataset that gives distributional labeling of entailment (E), neutral (N), and contradiction (C). A typical ground-truth label ϕ might be E 21%, N 12%, and C 67%, reflecting the varied human interpretations in each category. Our goal is to recover the distributional labels with ICQS. We randomly sample a test set of 60 samples.

We evaluate our performance on three metrics. 1) The Mean Absolute Error (MAE) as $\left|\phi - \hat{\phi}\right|_1$, 2) the cross-entropy (CE) as $-\sum_{i=1}^{n} \phi^i \log(\hat{\phi}^i)$, and 3) the Spearman correlation as defined in Equation (6)[3].

Table 2. Best performance of ICQS, ICQS-Epistemic and ICL on different metrics, on the ChaosNLI dataset.

	MAE ↓	CE ↓	Spearman's r ↑
Ask LLM	NA	NA	NA
ICL	0.333	0.525	-0.12
ICQS	0.246	0.372	0.245
ICQS-Epistemic	**0.240**	**0.371**	**0.336**

We present the performance curves, against number of ICEs, of ICQS and baseline results, for the spearman's coefficient metric in Figure 4b and for the MAE/CE metrics in Figures 10, 11 in Appendix G. For all three metrics, ICQS consistently outperforms ICL, as ICL fails to capture the distributional labeling. We summarize the best results in Table 2. Additionally, we provide the recovered average distribution for ICQS's best-performing ICE numbers. The results demonstrate that ICQS effectively recovers the distribution of each majority category, indicating its capability to learn the true semantic meaning.

Table 3. ICQS is able to reproduce the label prediction with 8 ICEs.

Sample's top label	Average Scores
E	[0.393 0.322 0.286]
N	[0.275 0.372 0.352]
C	[0.266 0.284 0.450]

How does ICQS progress with the number of in-context examples? Figures 4a and 4b both show the trend of ICQS performance with respect to the number of in-context examples. ICQS and ICQS-Epistemic outperform ICL, especially with a higher number of ICEs. Overall, ICQS performance gradually improves but may plateau or fluctuate after exceeding a certain number of ICEs, similar to ICL. Among all methods, ICQS-Epistemic achieves the most stable performance.

Is epistemic ICQS a stable improvement over ICQS? Yes, results across all tasks (see Figures 4a, 4b,10, 11) demonstrate that epistemic ICQS enhances the stability of ICQS, making it more practically useful.

6. Conclusion

In this paper, we introduce a general method for quality scoring of language generations that is efficient in terms of LLM evaluation requirements. Our approach applies to a wide range of probabilistic discrepancy evaluation problems, including the assessment of large language models (LLMs). The core idea is to construct a representative quality ranking function, r, based on structured in-context learning with a mixture of demonstrations of varying quality. We validate the effectiveness of our method through experiments on sentiment analysis, LMSYS model ranking, and fine-grained labeling tasks. These results validate the effectiveness of our

[3]Unlike previous experiments where we ranked individual labels, this time we rank the quality of the entire hypothesis sentence.

approach and provide insights into the structural properties of in-context learning, particularly its semantic generalization capability. Our findings suggest exploring semantically structured demonstrations as a potential direction for further investigating the statistical performance of in-context learning.

A limitation of our work is the challenge of achieving stable improvements as the number of prompts increases, due to the discrete nature of in-context learning.

7. Impact Statement

The proposed method for evaluating large language models (LLMs) contributes to the development of more reliable, transparent, and fair AI systems. By providing a rigorous and systematic evaluation framework, our approach enhances the ability to assess model performance across diverse tasks, mitigating biases and improving generalization. This can lead to better decision-making in high-stakes applications such as healthcare, legal analysis, and content moderation. Our work may have various societal implications, but none that require specific emphasis here.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Agarwal, R., Singh, A., Zhang, L. M., Bohnet, B., Chan, S., Anand, A., Abbas, Z., Nova, A., Co-Reyes, J. D., Chu, E., et al. Many-shot in-context learning. *arXiv preprint arXiv:2404.11018*, 2024.

Bigelow, E. J., Lubana, E. S., Dick, R. P., Tanaka, H., and Ullman, T. D. In-context learning dynamics with random binary sequences. *arXiv preprint arXiv:2310.17639*, 2023.

Brown, T. B. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.

Dai, Z., Tomasi, F., and Ghiassian, S. In-context exploration-exploitation for reinforcement learning. *arXiv preprint arXiv:2403.06826*, 2024.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Tan, X. E., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Grattafiori, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Vaughan, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Franco, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B.,

Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Wyatt, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Ozgenel, F., Caggioni, F., Guzmán, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Thattai, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Damlaj, I., Molybog, I., Tufanov, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Prasad, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Huang, K., Chawla, K., Lakhotia, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Tsimpoukelli, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Laptev, N. P., Dong, N., Zhang, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Li, R., Hogan, R., Battey, R., Wang, R., Maheswari, R., Howes, R., Rinott, R., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Kohler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Albiero, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wang, X., Wu, X., Wang, X., Xia, X., Wu, X., Gao, X., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Hao, Y., Qian, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., and Zhao, Z. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Dubois, Y., Galambosi, B., Liang, P., and Hashimoto, T. B. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Falck, F., Wang, Z., and Holmes, C. Is in-context learning in large language models bayesian? a martingale perspective. *arXiv preprint arXiv:2406.00793*, 2024.

Fu, D., Chen, T.-Q., Jia, R., and Sharan, V. Transformers learn higher-order optimization methods for in-context learning: A study with linear models. *arXiv preprint arXiv:2310.17086*, 2023.

Garg, S., Tsipras, D., Liang, P. S., and Valiant, G. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.

Grigsby, J., Fan, L., and Zhu, Y. Amago: Scalable in-context reinforcement learning for adaptive agents. *arXiv preprint arXiv:2310.09971*, 2023.

Hou, B., Liu, Y., Qian, K., Andreas, J., Chang, S., and Zhang, Y. Decomposing uncertainty for large language models through input clarification ensembling. *arXiv preprint arXiv:2311.08718*, 2023.

Kalai, A. T. and Vempala, S. S. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pp. 160–171, 2024.

Kim, S., Shin, J., Cho, Y., Jang, J., Longpre, S., Lee, H., Yun, S., Shin, S., Kim, S., Thorne, J., et al. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*, 2023.

Kim, S., Suk, J., Longpre, S., Lin, B. Y., Shin, J., Welleck, S., Neubig, G., Lee, M., Lee, K., and Seo, M. Prometheus 2: An open source language model specialized in evaluating other language models. *arXiv preprint arXiv:2405.01535*, 2024.

Kirsch, L., Harrison, J., Freeman, D., Sohl-Dickstein, J., and Schmidhuber, J. Towards general-purpose in-context learning agents. Workshop on Distribution Shifts, 37th Conference on Neural Information . . . , 2023.

Lee, J., Xie, A., Pacchiano, A., Chandak, Y., Finn, C., Nachum, O., and Brunskill, E. Supervised pretraining can learn in-context reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Li, X., Zhang, T., Dubois, Y., Taori, R., Gulrajani, I., Guestrin, C., Liang, P., and Hashimoto, T. B. Alpacaeval: An automatic evaluator of instruction-following models, 2023.

Lin, B. Y., Deng, Y., Chandu, K., Brahman, F., Ravichander, A., Pyatkin, V., Dziri, N., Bras, R. L., and Choi, Y. Wildbench: Benchmarking llms with challenging tasks from real users in the wild. *arXiv preprint arXiv:2406.04770*, 2024.

Ling, C., Zhao, X., Zhang, X., Cheng, W., Liu, Y., Sun, Y., Oishi, M., Osaki, T., Matsuda, K., Ji, J., et al. Uncertainty quantification for in-context learning of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3357–3370, 2024.

Liu, S., Cai, Z., Chen, G., and Li, X. Towards better understanding of in-context learning ability from in-context uncertainty quantification. *arXiv preprint arXiv:2405.15115*, 2024a.

Liu, T. J., Boullé, N., Sarfati, R., and Earls, C. J. Density estimation with llms: a geometric investigation of in-context learning trajectories. *arXiv preprint arXiv:2410.05218*, 2024b.

Malo, P., Sinha, A., Korhonen, P., Wallenius, J., and Takala, P. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796, 2014.

Nie, Y., Zhou, X., and Bansal, M. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020.

Raparthy, S. C., Hambro, E., Kirk, R., Henaff, M., and Raileanu, R. Generalization to new sequential decision making tasks with in-context learning. *arXiv preprint arXiv:2312.03801*, 2023.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.

Sridhar, K., Dutta, S., Jayaraman, D., and Lee, I. Regent: A retrieval-augmented generalist agent that can act in-context in new environments. *arXiv preprint arXiv:2412.04759*, 2024.

Tanneru, S. H., Agarwal, C., and Lakkaraju, H. Quantifying uncertainty in natural language explanations of large language models. In *International Conference on Artificial Intelligence and Statistics*, pp. 1072–1080. PMLR, 2024.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Wei, Z., Wang, Y., and Wang, Y. Jailbreak and guard aligned language models with only few in-context demonstrations. *arXiv preprint arXiv:2310.06387*, 2023.

Wies, N., Levine, Y., and Shashua, A. The learnability of in-context learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2111.02080.

Yadkori, Y. A., Kuzborskij, I., György, A., and Szepesvári, C. To believe or not to believe your llm. *arXiv preprint arXiv:2406.02543*, 2024.

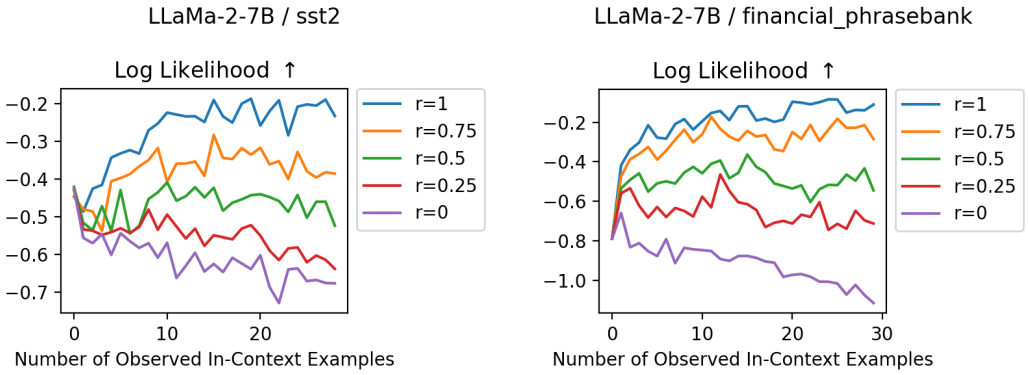
Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623, 2023.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.

A. Primary Verification

In this section, we evaluate the effectiveness of our proposed method along with its semantic generalization ability. All experiments are conducted using the base model LLaMA-2-7B-8bit with at most 30 in-context examples (ICE).

Correct labels $\Rightarrow r = 1$. For correct evaluation sequences (i.e., correctly labeled sentiment), ICQS assigns a quality score of 1, as the $r = 1$ curve is positioned at the top.

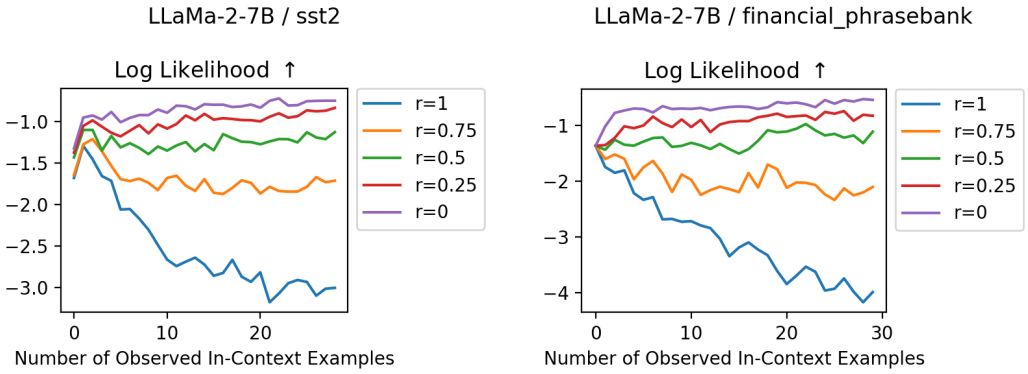


(a) SST-2 sentiment analysis

(b) Financial PhraseBank.

Figure 5. Since the log-likelihood of the $r = 1$ mixture ICE model is the highest, we assign $r = 1$ as the quality score for the evaluation sequences. Likelihoods are averaged over 30 samples.

Incorrect labels $\Rightarrow r = 0$ For incorrect evaluation sequences (i.e., those with wrongly labeled sentiment), ICQS assigns a score of $r = 0$ to the evaluation sequences.



(a) SST2 sentiment analysis

(b) Financial phrasebank

Figure 6. Since the log likelihood of the $r = 0$ mixture ICE model is on top, we assign $r = 0$ as the quality score of the eval sequences. Likelihood are averaged over 30 samples.

Semantic generalization ability Next, we investigate the semantic generalization ability of ICQS, an emerging property observed in our experiments. Despite being provided with only positive and negative sentences and their corresponding labels in the in-context examples (ICEs), ICQS is able to assign appropriate scores to unseen neutral sentences and neutral labels. Moreover, ICQS can generalize to labelings that use synonyms, demonstrating its generalization ability.

Figure 7 shows that ICQS assigns scores ranging from 0.75 to 1 for unseen neutral sentences. Table 4 demonstrates that ICQS achieves the best overall performance when dealing with correctly and incorrectly labeled positive and negative samples, neutrally labeled positive and negative samples, and semantically neutral examples.

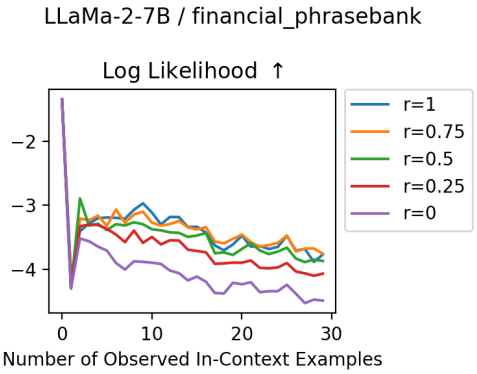
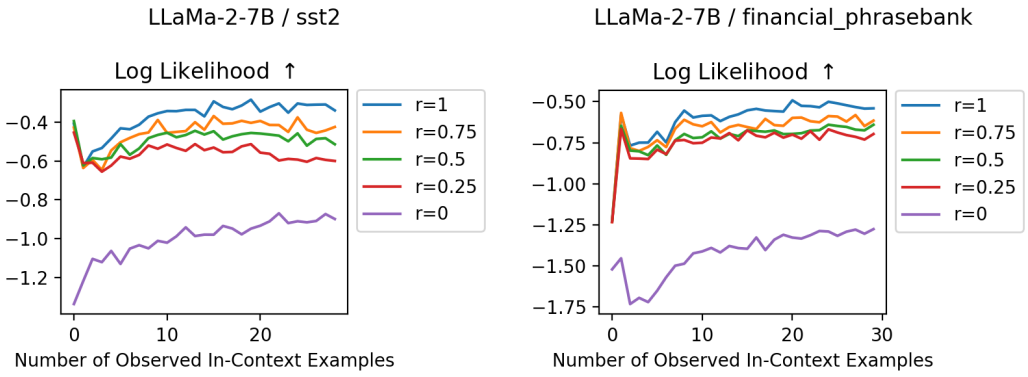


Figure 7. Financial phrasebank. Base model is conditioned on neg/pos sentences; but test on neutral sentences.

Table 4. Quantitative results on financial phrasebank dataset. In-context prompts are all neg/pos examples.

Accuracy on	Ask LLM	ICL	ICQS
Correct	85%	100%	**100%**
Incorrect	48.67%	96.7%	**96.7%**
Labeled neutral	**100%**	0	80.0%
Semantic neutral	46.7%	0	**67.7%**

Moreover, ICQS can accurately score unseen synonymous sentiment labels. As shown in Figure 8, ICQS assigns the correct quality score of $r = 1$ to the evaluation sequences. These results further validate the semantic generalization ability of ICQS in scoring.



(a) SST2 sentiment analysis

(b) Financial phrasebank

Figure 8. ICQS Synonym Scoring: The ICE labels are negative/positive, with SST-2 labels as bad/good and FB labels as sad/happy. Despite these variations, ICQS assigns a quality score of $r = 1$ to the evaluation sequences, demonstrating its robust synonym handling capability.

B. Leaderboard

We present the complete model rankings for the LMSYS joke subset in Table 5.

Table 5. Full LMSYS joke ranking. Tie models are in the same grid.

Ground Truth	ICQS	ICL
tripedhyena-nous-7b	mpt-30b-chat	gpt-4-0125-preview
stablelm-tuned-alpha-7b	qwen1.5-72b-chat	mpt-30b-chat
llama2-70b-steerlm-chat	zephyr-7b-alpha	llama2-70b-steerlm-chat
chatglm3-6b	stablelm-tuned-alpha-7b	openhermes-2.5-mistral-7b
gpt-3.5-turbo-0314	wizardlm-70b	wizardlm-13b
palm-2	llama2-70b-steerlm-chat	wizardlm-70b
gpt4all-13b-snoozy	llama-2-13b-chat	stablelm-tuned-alpha-7b
pplx-70b-online	gemini-pro	solar-10.7b-instruct-v1.0
tulu-2-dpo-70b	oasst-pythia-12b	gpt-4-1106-preview
gpt-4-1106-preview	vicuna-7b	mistral-medium
vicuna-33b	mistral-medium	llama-2-13b-chat
llama-2-70b-chat	gpt-4-1106-preview	oasst-pythia-12b
llama-2-13b-chat	claude-1	dolly-v2-12b
starling-lm-7b-alpha	dolly-v2-12b	mpt-7b-chat
codellama-34b-instruct	starling-lm-7b-alpha	gpt-3.5-turbo-0314
gemini-pro	qwen-14b-chat	claude-1
dolly-v2-12b	chatglm3-6b	gemini-pro
gpt-4-0314	zephyr-7b-beta	gpt-4-0314
gpt-3.5-turbo-0613	alpaca-13b	chatglm3-6b
claude-2.1	palm-2	vicuna-33b
claude-1	gpt-4-0314	RWKV-4-Raven-14B
gpt-3.5-turbo-1106	vicuna-33b	llama-2-70b-chat
guanaco-33b	gpt-3.5-turbo-0314	gpt4all-13b-snoozy
openhermes-2.5-mistral-7b	tulu-2-dpo-70b	fastchat-t5-3b
qwen1.5-72b-chat	stripedhyena-nous-7b	zephyr-7b-alpha
fastchat-t5-3b	chatglm-6b	stripedhyena-nous-7b
mpt-30b-chat	gpt-3.5-turbo-0613	gpt-3.5-turbo-0613
oasst-pythia-12b	llama-2-70b-chat	codellama-34b-instruct
zephyr-7b-alpha	claude-2.1	zephyr-7b-beta
vicuna-7b	openchat-3.5	gpt-3.5-turbo-1106
gpt-4-0613	gpt-3.5-turbo-1106	claude-instant-1
mistral-medium	mixtral-8x7b-instruct-v0.1	chatglm-6b
alpaca-13b	claude-instant-1	pplx-7b-online
zephyr-7b-beta	codellama-34b-instruct	gpt-4-0613
wizardlm-13b	gpt-4-0613	alpaca-13b
claude-instant-1	fastchat-t5-3b	mixtral-8x7b-instruct-v0.1
solar-10.7b-instruct-v1.0	pplx-70b-online	mistral-7b-instruct
deepseek-llm-67b-chat	mistral-7b-instruct	starling-lm-7b-alpha
koala-13b	solar-10.7b-instruct-v1.0	claude-2.0
openchat-3.5	pplx-7b-online	tulu-2-dpo-70b
RWKV-4-Raven-14B	llama-2-7b-chat	claude-2.1
pplx-7b-online	RWKV-4-Raven-14B	deepseek-llm-67b-chat
vicuna-13b	claude-2.0	palm-2
mpt-7b-chat	vicuna-13b	vicuna-13b
gemini-pro-dev-api	deepseek-llm-67b-chat	yi-34b-chat
claude-2.0	koala-13b	llama-2-7b-chat
wizardlm-70b	wizardlm-13b	koala-13b
chatglm-6b	yi-34b-chat	openchat-3.5
mixtral-8x7b-instruct-v0.1	chatglm2-6b	vicuna-7b
mistral-7b-instruct	llama-13b	qwen-14b-chat
yi-34b-chat	gemini-pro-dev-api	pplx-70b-online
llama-2-7b-chat	mpt-7b-chat	gemini-pro-dev-api
chatglm2-6b	gpt-3.5-turbo-0125	gpt-3.5-turbo-0125
gpt-3.5-turbo-0125	openhermes-2.5-mistral-7b	guanaco-33b
qwen-14b-chat	gpt4all-13b-snoozy	llama-13b
llama-13b	guanaco-33b	chatglm2-6b
gpt-4-0125-preview	gpt-4-0125-preview	qwen1.5-72b-chat

C. Prompt templates

We follow the prompt template of (Zheng et al., 2024) for the ask LLM baseline.

```
[Instruction]
Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question displayed below. Please rate the correctness of the response on a scale of 1 to 5, where 1 means incorrect and 5 means correct. Do not explain your evaluation.
[Question]
Is the sentiment of the following sentence 'negative', 'neutral', or 'positive': Finnish construction company YIT Oyj said on November 13 , 2007 it won a 70 mln euro $ 102.8 mln contract to construct the new office building for local property company Tapiola Real Estate Oy .
Respond with 'negative', 'neutral', or 'positive'.
[The Start of Assistant's Answer]
neutral
[The End of Assistant's Answer]
Rating:
```

Listing 1. Prompt template for financial phrasebank LLM evaluation.

```
Sentence: The pretax profit of the group 's life insurance business increased to EUR36m from EUR27m.
Answer: Positive.
```

Listing 2. In-context template for financial phrasebank LLM evaluation.

```
model_a: llama-2-13b-chat
model_b: gpt-4-0613
prompt: "tell me a joke containing Tiger and Mobile phone?"
response_a: "Sure, here's a joke for you:\n\nWhy did the tiger take his mobile phone to the party?\n\nBecause he heard it was a paw-some way to meet new people!\n\n(Get it? Paw-some? Like \"awesome\" but with \"paw\" because it's a tiger? Okay, maybe it's not that funny... but I hope it made you smile!)"
response_b: "Why don't tigers use mobile phones? \n\nBecause they're afraid they might press the \"paws\" button too often!"
winner: 1
```

Listing 3. Example format of LMSYS dataset

```
Premise: uh-huh and is it true i mean is it um.
Hypothesis: It is absolutely correct.
```

Listing 4. Example format of ChaosNLI dataset

D. Discussion on Scoring Granularity

One possible concern with our algorithm may be its computational complexity, as we require different ratios r of in-context sampling for finer scoring. Interestingly, some recent research (Kalai & Vempala, 2024) shows that large models should favor a more concentrated distribution of score likelihood to mitigate hallucination (Figure 9).

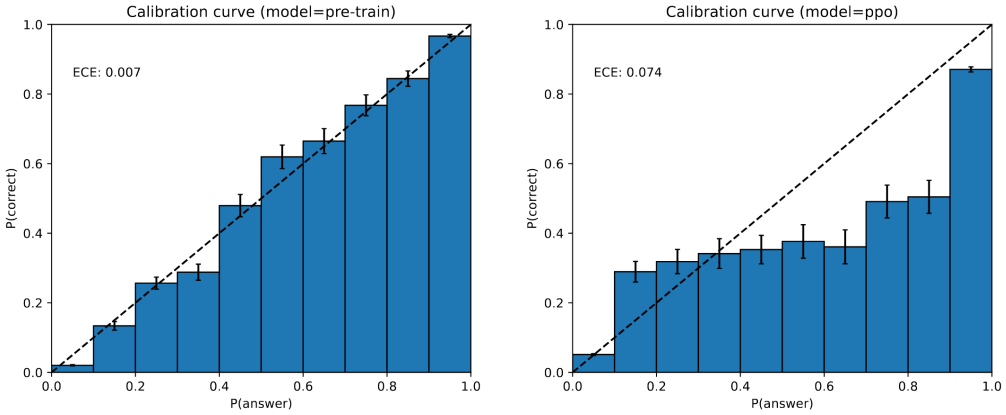


Figure 9. GPT-4 calibration curves before (left) and after (right) reinforcement learning. (Achiam et al., 2023; Kalai & Vempala, 2024) revealed the fact that in order to avoid hallucination, LLMs should obtain answer distributions akin the right figure.

E. Assumptions from (Wei et al., 2023)

We follow the similar weak assumptions to (Wei et al., 2023) in our framework. For self-containment, we list the assumptions below.

Assumption E.1 (Independence on requests)**.** For any question $\forall x \sim Q$ and its prefix prompt p^*, we have $P_b(x \mid p^*) = P_g(x \mid p^*)$. This assumes that the probability of each question is the same for the two distributions.

Assumption E.2 (Robustness of a single distribution)**.** For any demonstration set ICE_r and request x, we have $P_b(y \mid [\text{ICE}_r, x]) = P_b(y \mid x)$ and $P_g(y \mid [\text{ICE}_r, x]) = P_g(y \mid x)$. The distribution P_b (or P_g) is robust to context; that is, the output of the current question is unaffected by the preceding conversation.

Assumption E.3 (Distinguishability between the distributions)**.** There exists $\Delta > 0$ such that for any $\forall(x, y) \sim P_g$. Then, $\frac{P_g(y|x)}{P_b(y|x)} > \Delta$. Similarly is for any $\forall(x, y) \sim P_b$, $\frac{P_b(y|x)}{P_g(y|x)} > \Delta$.

Next, we provide the proof of Eq. (5) from Wei et al. (2023, Proof for Theorem 4.5).

Proof of Eq. (5). For k ICEs $\{(x_i, y_i) \sim P_z\}_{i=1}^k$ and a new test case $x \sim Q$, note that

$$
\begin{aligned}
&\frac{P_g([x_1, y_1, \cdots, x_k, y_k, x])}{P_b([x_1, y_1, \cdots, x_k, y_k, x])} \\
=&\frac{P_g(x \mid [x_1, y_1, \cdots, x_k, y_k])}{P_b(x \mid [x_1, y_1, \cdots, x_k, y_k])} \cdot \frac{P_g([x_1, y_1, \cdots, x_k, y_k])}{P_b([x_1, y_1, \cdots, x_k, y_k])} \\
=&\frac{P_g([x_1, y_1, \cdots, x_k, y_k])}{P_b([x_1, y_1, \cdots, x_k, y_k])} \quad \text{(Assumption E.1)} \\
=&\frac{P_g(y_k \mid [x_1, y_1, \cdots, x_k])}{P_b(y_k \mid [x_1, y_1, \cdots, x_k])} \cdot \frac{P_g([x_1, y_1, \cdots, x_k])}{P_b([x_1, y_1, \cdots, x_k])} \\
=&\frac{P_g(y_k \mid x_k)}{P_b(y_k \mid x_k)} \cdot \frac{P_g([x_1, y_1, \cdots, x_k])}{P_b([x_1, y_1, \cdots, x_k])} \quad \text{(Assumption E.2)} \\
=&\frac{P_g(y_k \mid x_k)}{P_b(y_k \mid x_k)} \cdot \frac{P_g(x_k \mid [x_1, y_1, \cdots, x_{k-1}, y_{k-1}])}{P_b(x_k \mid [x_1, y_1, \cdots, x_{k-1}, y_{k-1}])} \cdot \frac{P_g([x_1, y_1, \cdots, x_{k-1}, y_{k-1}])}{P_b([x_1, y_1, \cdots, x_{k-1}, y_{k-1}])}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{P_g(y_k \mid x_k)}{P_b(y_k \mid x_k)} \cdot \frac{P_g([x_1, y_1, \cdots, x_{k-1}, y_{k-1}])}{P_b([x_1, y_1, \cdots, x_{k-1}, y_{k-1}])} \quad \text{(Assumption E.1)} \\
&= \frac{P_g(y_k \mid x_k)}{P_b(y_k \mid x_k)} \cdot \frac{P_g(y_{k-1} \mid x_{k-1})}{P_b(y_{k-1} \mid x_{k-1})} \cdot \frac{P_g([x_1, y_1, \cdots, x_{k-2}, y_{k-2}])}{P_b([x_1, y_1, \cdots, x_{k-2}, y_{k-2}])} \\
&= \cdots \\
&= \prod_{i=1}^{k} \frac{P_g(y_i \mid x_i)}{P_b(y_i \mid x_i)}
\end{aligned}
$$

□

F. Discussions on LLM-as-a-Judge Baselines

There are multiple baselines for the LLM-as-a-Judge task. For example, the AlpacaEval (Li et al., 2023; Dubois et al., 2024) and WILDBENCH (Lin et al., 2024) mainly provide coarse datasets that mimic Chatbot Arena (Chiang et al., 2024) to evaluate the overall performance of LLMs. The difficulty in comparing these two papers with our algorithm is that ICL is specifically designed for task-wise settings. In this case, it is not straightforward for us to adapt the AlpacaEval and WILDBENCH datasets into task-specific subsets.

On the other hand, the two PROMETHEUS works (Kim et al., 2023; 2024) trained evaluators specifically for fine-grained evaluation tasks using a large number (20k) of human annotations. It is not very fair to compare those evaluators with our ICQS method, which requires fewer than 100 examples and no fine-grained human annotations.

G. Additional Results on the ChaosNLI Dataset

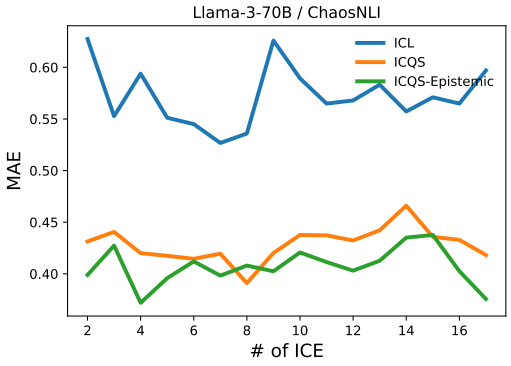


Figure 10. ICL, ICQS and ICQS-Epistemic cross entropy results ↓.

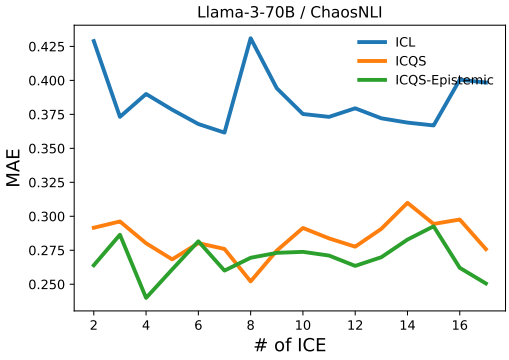


Figure 11. ICL, ICQS and ICQS-Epistemic MAE results ↓.