

A Study on the Calibration of In-context Learning

Hanlin Zhang¹ Yi-Fan Zhang² Yaodong Yu³
 Dhruv Madeka⁴ Dean Foster⁴ Eric Xing^{5,6} Hima Lakkaraju⁴ Sham Kakade^{1,4}

¹Harvard University ²Chinese Academy of Sciences ³UC Berkeley ⁴Amazon
⁵Carnegie Mellon University ⁶Mohamed Bin Zayed University of Artificial Intelligence

Abstract

Modern auto-regressive language models are trained to minimize log loss on broad data by predicting the next token so they are expected to get calibrated answers when framing a problem as next-token prediction task. We study this for in-context learning (ICL), a widely used way to adapt frozen large language models (LLMs) via crafting prompts, and investigate the trade-offs between performance and calibration on a wide range of natural language understanding and reasoning tasks. We conduct extensive experiments to show that such trade-offs may get worse as we increase model size, incorporate more ICL examples, and fine-tune models using instruction, dialog, or reinforcement learning from human feedback (RLHF) on carefully curated datasets. Furthermore, we find that common recalibration techniques that are widely effective such as temperature scaling provide limited gains in calibration errors, suggesting that new methods may be required for settings where models are expected to be reliable.

1 Introduction

Language models (LMs) that encompass transformer-based architectures (Brown et al., 2020; Chowdhery et al., 2023; OpenAI, 2023) can generate coherent and contextually relevant texts for various use cases. Despite their impressive performance, these models occasionally produce erroneous or overconfident outputs, leading to concerns about their calibration (Dawid, 1982; DeGroot and Fienberg, 1983) which measures how faithful a model’s prediction uncertainty is. Such a problem is pressing as users adapt them using a recent paradigm called in-context learning (Brown et al., 2020) to construct performant predictors, especially for applications in safety-critical domains (Bhatt et al., 2021; Kadavath et al., 2022; Pan et al., 2023).

We provide an in-depth evaluation and analysis

of how well these models are calibrated - that is, the alignment between the model’s confidence in its predictions and the actual correctness of those predictions. This token-level calibration assessment will enable us to measure the discrepancy between the model’s perceived and actual performance through a Bayesian uncertainty lens, providing a valuable metric for assessing the model’s accuracy and reliability.

We find that LMs including GPT-2 (Radford et al., 2019) and LLaMA (Touvron et al., 2023a) are poorly calibrated and there exists a calibration-accuracy trade-off (Fig. 1), i.e. as we increase the amount of in-context samples, the prediction accuracy and calibration error both increase. Crucially, this calibration degradation worsens as the model size increases or when fine-tuning occurs using specialized data, such as curated instructions (Dubois et al., 2023), dialogues (Zheng et al., 2023), or human preference data (Ziegler et al., 2019). Though previous work (Braverman et al., 2020) shows the entropy of each generation step is drifting and can be recalibrated via scaling techniques (Platt et al., 1999) such as temperature scaling (Guo et al., 2017), we show that the miscalibration issue in ICL can not be easily addressed using such well-established recalibration approaches that rely on additional validation data.

Moreover, we study the trade-off in reasoning tasks that involve the generation of explanations (Camburu et al., 2018; Nye et al., 2021; Wei et al., 2022) before the answer, showing that the model can produce confidently wrong answers (using confidence histograms and reliability plots) when prompted with explanations on Strategy QA (Geva et al., 2021), Commonsense QA (Talmor et al., 2018), OpenBook QA (Mihaylov et al., 2018), World Tree (Jansen et al., 2018). We carefully design our human assessment to observe that, with the increase in model sizes and the quantity of ICL examples, there is a corresponding rise in the pro-

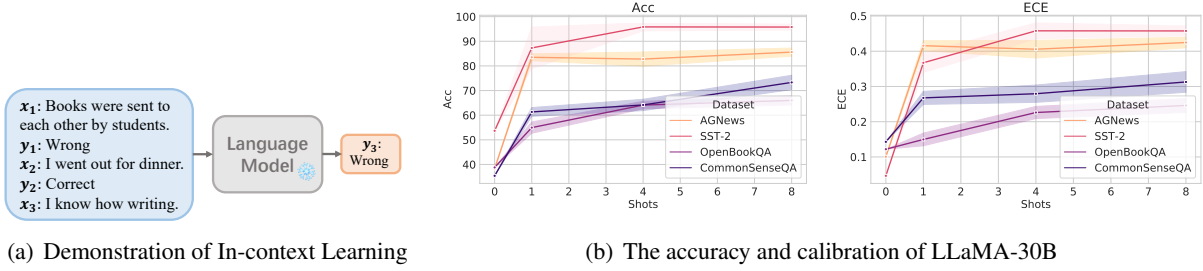


Figure 1: **The accuracy-calibration trade-off of in-context learning.** (a) ICL concerns taking task-specific examples as the prompt to adapt a frozen LLM to predict the answer. (b) Classification accuracy and expected calibration error of ICL. As the number of ICL samples increases, the prediction accuracy improves (**Left**); at the same time, the calibration gets worse (**Right**).

portion of confidently predicted examples among those incorrectly forecasted. Moreover, we find that a high proportion of wrong predictions are of high confidence and showcase those typical confidently wrong examples of LLMs.

In-context learning has been expected to learn models by gradient descent in their forward pass (Von Oswald et al., 2023), which might hopefully yield calibrated predictions (Błasiok et al., 2023) if the models are getting close to local optimality with respect to test loss through meta-optimization. However, the fact that choosing ICL samples from the validation set does not naturally lead to calibrated predictions shows that ICL learns in a fairly different way than SGD. We design controlled experiments to illustrate task learning properties of ICL, showing that when examples in the prompt demonstrate consistent task properties, the learning performance, and calibration would be improved.

2 Related Work

Uncertainty quantification in NLP. Uncertainty quantification in NLP, which often adopts the Bayesian principle to sophisticated methods tailored for neural networks, aims to enhance the reliability of model predictions. This may involve non-trivial designs as directly interpreting language model predictions via probabilities (Kadavath et al., 2022) and linguistic expressions (Lin et al., 2022; Mielke et al., 2022; Zhou et al., 2023) may inadvertently lead to over-reliance on the model’s uncertainties (Si et al., 2023), thus complicating the establishment of trustworthy common ground between humans and models (Bućinca et al., 2021). Notable recent advancements include employing model confidence as a critical factor in various applications like dialogue generation (Mielke et al.,

2022), cascading prediction (Schuster et al., 2021), open-domain QA (Fisch et al., 2020; Angelopoulos et al., 2022), summarization (Laban et al., 2022), language modeling (Schuster et al., 2022), image captioning (Petryk et al., 2023).

Calibration of LLMs. Calibration is a safety property to measure the faithfulness of machine learning models’ uncertainty, especially for error-prone tasks using LLMs. Previous works find that pre-training (Desai and Durrett, 2020) and explanation (Zhang et al., 2020; González et al., 2021) improves calibration. Models can be very poorly calibrated when we prompt LMs (Jiang et al., 2021), while calibration can also depend on model size (Kadavath et al., 2022). (Braverman et al., 2020) assesses the long-term dependencies in a language model’s generations compared to those of the underlying language and finds that entropy drifts as models such as GPT-2 generate text. The intricacy of explanations on complementary team performance poses additional challenges due to the over-reliance on explanations of users regardless of their correctness (Bansal et al., 2021). (Mielke et al., 2022) gives a framework for *linguistic calibration*, a concept that emphasizes the alignment of a model’s expressed confidence or doubt with the actual accuracy of its responses. The process involves annotating generations with <DK>, <LO>, <HI> for confidence levels, then training the confidence-controlled model by appending the control token <DK/LO/HI> at the start of the output, followed by training a calibrator to predict these confidence levels, and finally predicting confidence when generating new examples. (Tian et al., 2023) finds that asking LLMs for their probabilities can be better than using conditional probabilities in a traditional way. (Shih et al., 2023) proposes a simple amor-

tized inference trick for temperature-scaled sampling from LMs and diffusion models. To enhance the estimation of uncertainty in language models, (Kuhn et al., 2023) developed a method that aggregates log probabilities across semantically equivalent outputs. This approach utilizes bidirectional entailment through a model to identify outputs that are semantically similar, thereby refining the uncertainty estimation process. (Cole et al., 2023) identifies the calibration challenge in ambiguous QA and distinguishes uncertainty about the answer (epistemic uncertainty) from uncertainty about the meaning of the question (denotational uncertainty), proposing sampling and self-verification methods. (Kamath et al., 2020) trains a calibrator to identify inputs on which the QA model errs and abstains when it predicts an error is likely. (Zhao et al., 2023) proposes the Pareto optimal learning assessed risk score for calibration and error correction but requires additional training. (Kalai and Vempala, 2023) shows the trade-off between calibration and hallucination but they didn’t study it in an ICL setting and how the predicted answer’s accuracy would impact those two safety aspects.

3 Background

Setting. Given a pre-trained language model $\mathcal{P}_\theta(w_t|w_{<t})$, we seek to adapt it using the prompt $w_0 = [x_1, y_1, x_2, y_2, \dots, x_{n-1}, y_{n-1}, x_n]$ to generate a predicted answer $y_n = \mathcal{P}_\theta(w_0)$. In the context of reasoning, a popular approach is to hand-craft some explanations/rationales/chain-of-thoughts e in the prompt $w_0 = [x_1, e_1, y_1, x_2, e_2, y_2, \dots, x_{n-1}, e_{n-1}, y_{n-1}, x_n]$ to generate explanation e_n and answer y_n , for the test sample: $\overbrace{w_1, w_2, \dots, w_k}^{e_n}, y_n = \mathcal{P}_\theta(w_0)$.

We extract token-level answer probabilities of LLMs,¹ e.g. for binary classification tasks, we filter and extract probabilities $P(\text{“Yes”})$ and $P(\text{“No”})$, based on which we calculate the following statistics for studying the confidence and calibration of LMs:

Confidence and feature norm. We record the maximum probability of the answer token as its confidence $\text{Conf} = \mathcal{P}_\theta(y_n|w_{<n})$ and the feature norm z_n as the hidden states of the answer token from the output of the last layer of the model.

Entropy rate. We denote the entropy of a token w_t at position t as $H(w_t|w_{<t}) =$

¹We also normalize the probability $\mathcal{P}_\theta(y_n | w_{<n}) \in \Delta^K$ for classification problems with K choices

$-\mathbb{E}_{w_t \sim \mathcal{P}_\theta(\cdot|w_{<t})}[\log \mathcal{P}_\theta(w_t|w_{<t})]$. We typically measure it based on the answer token via setting $w_t = y_n$. Note that auto-regressive LLMs are trained via maximizing the negative log-likelihood objective $\mathcal{L} = -\mathbb{E}_t[\log \mathcal{P}_\theta(w_t|w_{<t})]$ on massive corpora.

Empirical estimate of the expected calibration error (ECE) In the realm of probabilistic classifiers, calibration is a crucial concept. A classifier, denoted as \mathcal{P}_θ with parameters θ and operating over C classes, is said to be "canonically calibrated" when, for every probability distribution p over the C classes and for every label y , the probability that the label is y given the classifier’s prediction is p matches the component of p corresponding to y . This is mathematically represented as:

$$\forall p \in \Delta^{C-1}, \forall y \in Y : P(Y = y | \mathcal{P}_\theta(X) = p) = p_y. \quad (1)$$

Here, Δ^{C-1} symbolizes the $(C - 1)$ -dimensional simplex, which encompasses all potential probability distributions over the C classes.

A simpler calibration criterion is the "top-label calibration." In this case, a classifier is deemed calibrated if, for every top predicted probability p^* , the probability that the true label belongs to the class with the highest predicted probability, given that this maximum predicted probability is p^* , equals p^* . Formally:

$$\forall p^* \in [0, 1] : P(Y \in \arg \max p | \max \mathcal{P}_\theta(X) = p^*) = p^*. \quad (2)$$

To gauge the calibration of a model, we adopt Expected Calibration Error (ECE) defined as:

$$\mathbb{E}[|p^* - \mathbb{E}[Y \in \arg \max \mathcal{P}_\theta(X) | \max \mathcal{P}_\theta(X) = p^*]|]. \quad (3)$$

In real-world applications, this quantity cannot be computed without quantization. So, the ECE is approximated by segmenting predicted confidences into M distinct bins, B_1, \dots, B_M . The approximation is then computed as:

$$\widehat{\text{ECE}} = \sum_{m=1}^M \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)|.$$

Here, $\text{acc}(B_m)$ is the accuracy within bin B_m , and $\text{conf}(B_m)$ is the average confidence of predictions in bin B_m . The total number of samples is represented by n , and the dataset consists of n

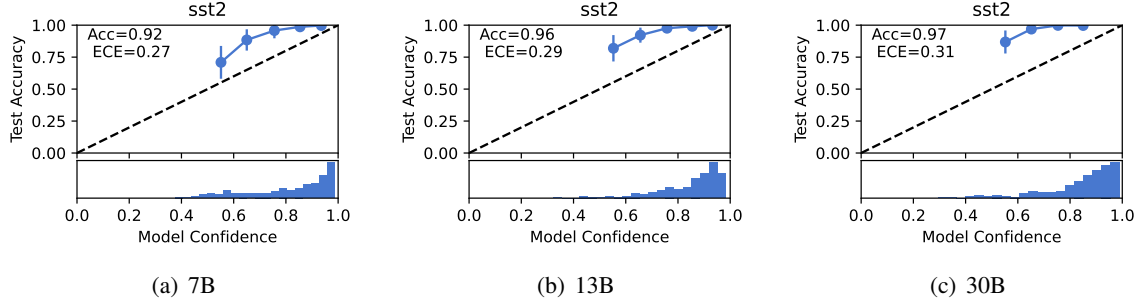


Figure 2: Reliability plots of LLaMA models.

Table 1: **Accuracy and Calibration** of LLaMA-30B model with three sizes across four text classification datasets and four reasoning datasets

| Dataset | LLaMA-30B | | | | | | | |
|---------------------------|-----------|-------|--------|-------|--------|-------|--------|-------|
| | 0-shot | | 1-shot | | 4-shot | | 8-shot | |
| | Acc | ECE | Acc | ECE | Acc | ECE | Acc | ECE |
| Text Classification | | | | | | | | |
| AGNews | 0.383 | 0.10 | 0.835 | 0.416 | 0.828 | 0.406 | 0.856 | 0.425 |
| TREC | 0.651 | 0.392 | 0.70 | 0.442 | 0.76 | 0.492 | 0.777 | 0.542 |
| CB | 0.50 | 0.143 | 0.696 | 0.409 | 0.821 | 0.383 | 0.798 | 0.359 |
| SST-2 | 0.537 | 0.047 | 0.873 | 0.367 | 0.958 | 0.458 | 0.958 | 0.458 |
| DBPdia | 0.363 | 0.287 | 0.792 | 0.669 | 0.83 | 0.69 | 0.782 | 0.646 |
| Reasoning with Scratchpad | | | | | | | | |
| Strategy QA | 0.452 | 0.039 | 0.617 | 0.047 | 0.679 | 0.045 | 0.678 | 0.088 |
| Commonsense QA | 0.354 | 0.143 | 0.613 | 0.268 | 0.642 | 0.279 | 0.733 | 0.313 |
| World Tree | 0.53 | 0.253 | 0.594 | 0.276 | 0.655 | 0.31 | 0.24 | 0.230 |
| OpenBook QA | 0.388 | 0.122 | 0.55 | 0.15 | 0.641 | 0.226 | 0.66 | 0.246 |

independent and identically distributed samples, $\{(x_i, y_i)\}_{i=1}^n$. In our work, we use this estimator to approximate the ECE.

4 Experimental Results

4.1 Experimental Settings

Models. We study decoder-only autoregressive LMs involving GPT-2, LLaMA (Touvron et al., 2023a), and their variants fine-tuned with instruction, dialog, or RLHF like Alpaca (Dubois et al., 2023), Vicuna (Zheng et al., 2023), and LLaMA2-Chat (Touvron et al., 2023b).

Datasets and tasks. We used both traditional NLU tasks such as AGNews (Zhang et al., 2015), TREC (Voorhees and Tice, 2000), CB (Schick and Schütze, 2021), SST-2 (Socher et al., 2013), DBPe-dia (Zhang et al., 2015), as well as reasoning question answering tasks like Strategy QA (Geva et al., 2021), Commonsense QA (Talmor et al., 2018), OpenBook QA (Mihaylov et al., 2018), World Tree

(Jansen et al., 2018). Notably, the reasoning task performance can be greatly improved in general via prompting methods like scratchpad (Nye et al., 2021; Wei et al., 2022) that enables models to generate natural language explanations before predicting an answer.

In-context learning settings. We prompt the model via sampling k examples from the training set for each test example in the k -shot setting. Each experiment is repeated 10 times to reduce variance and we report the mean results. We use $M = 10$ bins for calculating calibration errors.

4.2 Numerical Results

The performance of LLaMA. We seek to characterize the calibration-accuracy trade-off in both simple and realistic settings (Tab. 1). We record the performance and calibration errors in both miscalibrated and recalibrated settings. Moreover, we take a close look at the prompting approaches that explicitly include explanations in reasoning tasks

such as scratchpad (Nye et al., 2021) or chain-of-thought (Wei et al., 2022), showing that the calibration degrades after generating a long context for reasoning and explaining the final answer.

The effect of temperature scaling. We experiment with three strategies in applying temperature scaling methods (Guo et al., 2017) to fix miscalibration:

1. We learn one temperature for each n -shot ICL, i.e., we learn different temperatures for different shot numbers in ICL;
2. Learn a temperature from the training split (zero-shot) and apply it to all test samples with different shot numbers;
3. For each experiment, we fix the prompt and learn the temperature for the fixed prompt. That is, for every possible ICL prompt, we learn a corresponding temperature for calibration.

Looking into Fig. 6, none of the above strategies achieves satisfactory calibration performance, which is in contrast to the well-studied supervised learning setting where scaling the confidence scores (via temperature scaling) can effectively reduce calibration errors (Guo et al., 2017). The fact that applying a post-processing calibration method, such as temperature scaling, as used in most previous work, cannot directly resolve the miscalibration issue suggests that ICL might have substantially different properties compared to making predictions via classical supervised learning models, thus future investigations are needed to address such miscalibration issues.

The effect of finetuning. We show that vicuna and alpaca are both more accurate but less calibrated than their LLaMA counterpart backbones, the margin is especially large for reasoning tasks and vicuna. Thus we compare those models’ accuracy and ECE in Fig. 3, showing that finetuning might significantly degrade calibration, corroborating the evidence shown in (OpenAI, 2023), albeit it can improve the reasoning accuracy dramatically. Our results provide evidence that though finetuned on carefully curated datasets can greatly improve question-answering performance, especially for hard tasks like reasoning problems, attention may need to be paid when assessing the calibration of those models’ predictions.

The accuracy is high for 0-shot ICL but has not increased much as we include more in-context examples. We also note that the pattern of zero-shot performance is totally different for two fine-tuned models, i.e. vicuna, and alpaca.

The effect of prompt repetition. In our study investigating the impact of various prompt strategies, we employ three distinct approaches: **Repeat-context:** In this strategy, we construct the prompt as $w_0 = [x_1, x_1, \dots, x_1, y_1]$, where we repetitively include only the context x_1 a total of n times, excluding the label y_1 from repetition. **Repeat-prompt:** Here, we shape the prompt as $w_0 = [x_1, y_1, \dots, x_1, y_1]$, repeating both the context x_1 and the label y_1 n times within the prompt. **Normal:** In this strategy, we construct the prompt as $w_0 = [x_1, y_1, x_2, y_2, \dots, x_{n-1}, y_{n-1}, x_n, y_n]$, where distinct context-label pairs are systematically chosen to form the prompt. The results presented in Table 3 unveil essential insights: (1) The inclusion of labels within the prompt contributes to a reduction in uncertainty and facilitates more effective reasoning. Conversely, merely repeating the context without incorporating labels fails to yield improved performance. (2) Notably, the diversity inherent in the prompt’s construction significantly impacts performance, particularly concerning larger language models.

4.3 Qualitative Results

Reliability diagram and confidence histogram.

A reliability diagram is a graphical tool used to evaluate the calibration of probabilistic predictions of a model across multiple classes; it compares the predicted probabilities of each class against the actual outcomes, with a perfectly calibrated model having its values lie on the diagonal $y = x$ line. A confidence histogram, on the other hand, displays the distribution of the model’s prediction confidences across all classes, showing how often the model predicts certain probabilities.

We showcase that in SST-2 (4-shot), showing that both ACC and ECE of LLaMA increase as the model size increases (Fig. 2). We can observe that confidence scores tend to concentrate on values above 0.8 as we enlarge model sizes.

4.4 Ablation Studies

For case studies, we research how miscalibration can impact the selective classification of LLMs, where models are supposed to abstain from uncer-

Table 2: **Norm of representation, entropy, and confidence** of LLaMA-30B model across three text classification datasets.

| Dataset | LLaMA-30B | | | | | | | | | | | |
|---------|-----------|--------|--------|--------|---------|--------|--------|--------|------------|--------|--------|--------|
| | Norm | | | | Entropy | | | | Confidence | | | |
| | 0-shot | 1-shot | 4-shot | 8-shot | 0-shot | 1-shot | 4-shot | 8-shot | 0-shot | 1-shot | 4-shot | 8-shot |
| AGNews | 78.8 | 92.3 | 92.1 | 92.2 | 3.920 | 0.650 | 0.595 | 0.444 | 0.214 | 0.821 | 0.819 | 0.865 |
| CB | 88.4 | 91.7 | 89.2 | 87.9 | 3.857 | 1.266 | 0.935 | 0.823 | 0.193 | 0.566 | 0.629 | 0.577 |
| DBPdia | 77.9 | 89.5 | 91.0 | 90.1 | 4.105 | 1.438 | 0.848 | 0.718 | 0.078 | 0.578 | 0.705 | 0.671 |

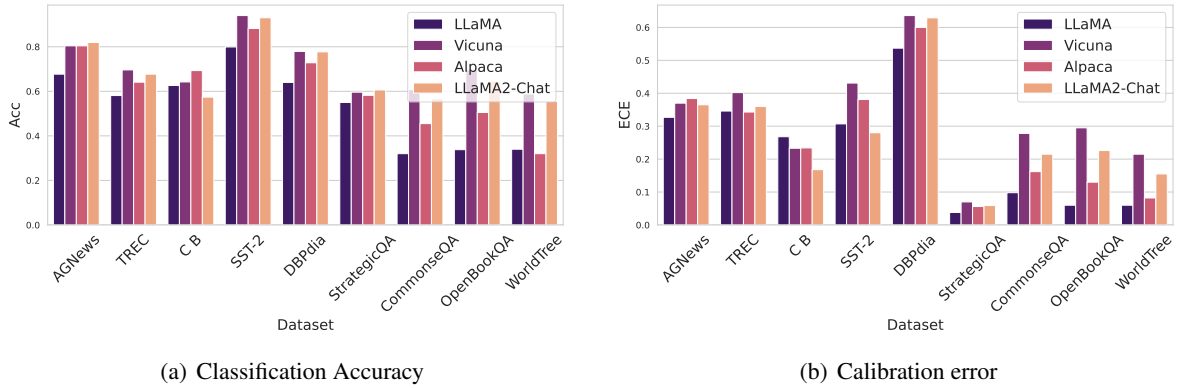


Figure 3: Accuracy and calibration errors of LLaMA and its finetuned variants.

tain predictions in high-stakes settings.

Ablation with model sizes. As we enlarge the sizes of models, they will become more confident and accurate (Fig. 2). As a result, the entropy decreases and ECE increases, showing that token-level calibration might have an inverse scaling relationship with model sizes.

A closer look at the hidden state and confidence score. To better understand the miscalibration issue of ICL, we conduct fined-grained experiments to take a closer look at ICL properties: we measure the norm of the representation vectors² for different number of shots in ICL, to better understand how the representation vectors are changing when increasing the number of shots in ICL. Meanwhile, we also measure the confidence and entropy of the prediction for y_n , and the results are summarized in Table 2. When switching from 0-shot to 1-shot, all three measurements (representation norm, entropy, and confidence) drastically change. Meanwhile, more ICL samples lead to smaller entropy and higher confidence in most cases.

Confidence and wrongly classified reasoning examples. To take a closer look at the failure modes of LMs, we randomly sample 100 reasoning exam-

ples of LLaMA and plot the distribution of wrongly predicted samples and the confidence scores via thresholding. Similar to previous observations, as model sizes and the number of ICL examples scale up, LMs would generate more confident samples (Fig. 4 (c, d)). Note, that we observe "inverse scaling" behaviors where models with larger sizes are more error-prone and tend to generate more confidently wrong samples (Fig. 5).

Examples of hallucinated explanations for highly confident predictions. Next, we showcase in Table 4 that models generate both wrong explanations and incorrect predictions with high confidence. We also observe that most of the wrong predictions are highly confident. We manually examine the correctness of explanations on common-sense QA, and found its high correlations with predicted answer accuracy, which is the opposite of token-level explainability that tends to get worse when the accuracy improves.

5 Discussion and Concluding Remarks

In our investigation of the token-level calibration of in-context learning in contemporary Large Language Models (LLMs), we have delineated the intricate balance between ICL performance and calibration. Our findings underscore the importance of

²The representation vector refers to the intermediate output before the linear prediction layer.

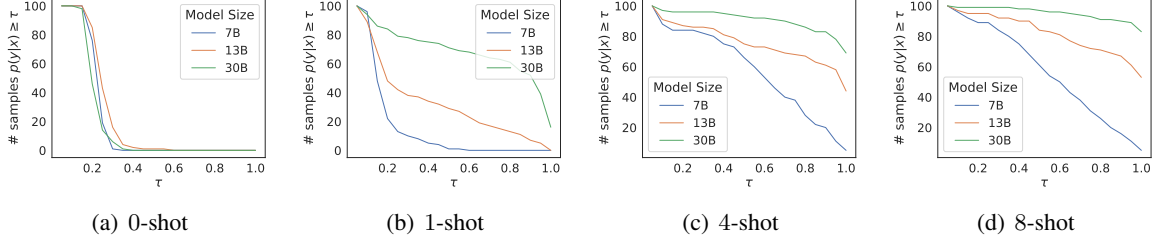


Figure 4: **Illustration of confidence distribution:** The number of samples whose confidence is greater than a threshold on Commonsense QA.

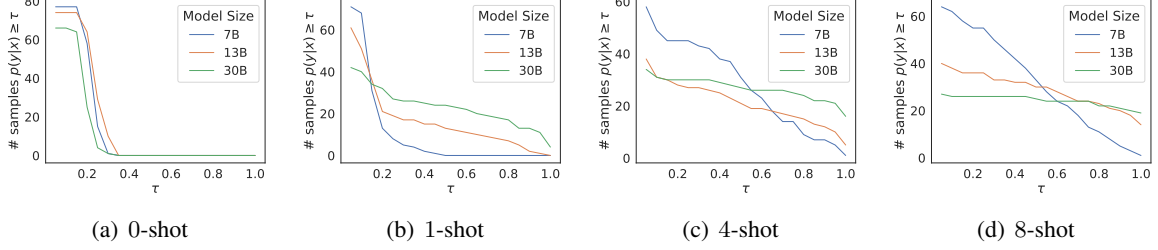


Figure 5: The number of **wrongly classified** examples whose confidence is above a threshold with different numbers of shots on Commonsense QA.

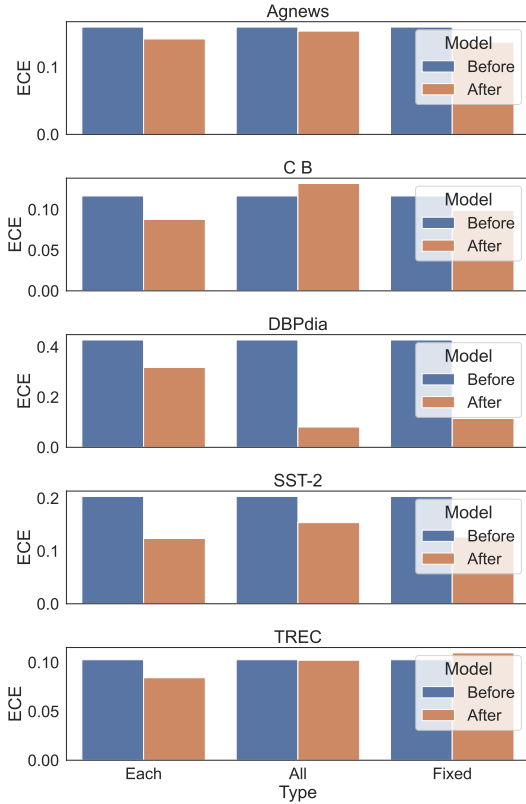


Figure 6: The comparison of calibration errors **before and after** applying different recalibration strategies.

being circumspect in model deployment, as maximizing ICL performance does not invariably translate to improved calibration. As LMs continue to evolve and gain more capabilities such as having long enough context windows that can include the whole training set as in-context examples for some downstream tasks, our result can be pedagogical when users would like to examine their uncertainty through prediction probabilities. Moreover, the work suggests the following future directions:

Understanding the internal mechanism of ICL for calibration. In this work, we observe that existing scaling recalibration methods cannot fully resolve the miscalibration issues of ICL, so better understanding and mitigation strategies are needed. A potential approach can be leveraging transparency tools and studying whether predictable errors exist during text generation.

Calibration beyond classification regimes. Our findings indicate that in multi-choice or multi-class classification tasks, even though the calibration of answer tokens may deteriorate in high-performance settings, there may be a positive correlation between accuracy and the correctness of explanations in reasoning tasks. This suggests potential avenues for future research such as exploring strategies such as the use of hedging words to express uncertainty and examining their relationship with predictive performance.

Table 3: **Accuracy and Calibration** of LLaMA-7B model and GPT-2 with different prompt repetition strategies.

| Model | Strategy | 1-shot | | 4-shot | | 8-shot | | AVG ACC | AVG ECE |
|----------|----------------|--------|-------|--------|-------|--------|-------|---------|---------|
| | | Acc | ECE | Acc | ECE | Acc | ECE | | |
| LLaMA-7B | Repeat-context | 78.200 | 0.281 | 78.370 | 0.283 | 77.470 | 0.274 | 78.013 | 0.279 |
| | Repeat-Prompt | 78.200 | 0.281 | 89.470 | 0.395 | 90.330 | 0.403 | 86.000 | 0.360 |
| | Normal | 78.200 | 0.281 | 95.200 | 0.452 | 96.130 | 0.462 | 89.843 | 0.398 |
| GPT-2 | Repeat-context | 65.970 | 0.287 | 55.300 | 0.230 | 54.330 | 0.230 | 58.533 | 0.249 |
| | Repeat-Prompt | 65.970 | 0.287 | 62.500 | 0.147 | 57.970 | 0.176 | 62.147 | 0.204 |
| | Normal | 65.970 | 0.287 | 63.030 | 0.212 | 56.270 | 0.220 | 61.757 | 0.240 |

Table 4: Qualitative Results of LLaMA on Commonsense and OpenBook

| Dataset | Case | Reasoning | \hat{y} | p |
|-------------|--|--|-----------|--------|
| Commonsense | Question: What would you do if you have curiosity but are blind and paralyzed? | | | |
| | Options: A) hear news, B) analyse, C) go somewhere, D) examine thing, E) see favorite show | | | |
| | Ground Truth: B | | | |
| | 7B | If you have curiosity but are blind and paralyzed, you would go somewhere to see things. You cannot hear news, analyse, go somewhere, examine thing or see favorite show. | C | 0.5046 |
| | 13B | If you are blind and paralyzed, you can't see anything. So, you can't go somewhere. You can't examine thing. You can't see favorite show. You can't analyse. So, you can hear news. | A | 0.8684 |
| | 30B | If you are blind and paralyzed, you cannot see your favorite show. You cannot hear news as you are blind. You cannot examine thing as you are paralyzed. You can go somewhere to see things. | C | 0.9720 |
| Open Book | Question: What would you do if you have curiosity but are blind and paralyzed? | | | |
| | Options: A) hear news, B) analyse, C) go somewhere, D) examine thing, E) see favorite show | | | |
| | Ground Truth: B | | | |
| | 7B | A girl and her mom have the same hair length. | D | 0.6365 |
| | 13B | A girl and her mom have the same date of birth. | A | 0.9782 |
| | 30B | A girl and her mom have the same genes. | A | 0.9831 |

Acknowledgment

We thank Jean-Stanislas Denain for valuable feedback. Kempner Institute computing resources enabled this work.

References

- Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. 2022. Conformal risk control. *arXiv preprint arXiv:2208.02814*.
- Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16.
- Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, et al. 2021. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 401–413.
- Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. 2023. When does optimizing a proper loss yield calibration? *arXiv preprint arXiv:2305.18764*.
- Mark Braverman, Xinyi Chen, Sham Kakade, Karthik Narasimhan, Cyril Zhang, and Yi Zhang. 2020. Calibration, entropy rates, and memory in language models. In *International Conference on Machine Learning*, pages 1089–1099. PMLR.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Zana Bućinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21.

- Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.
- Jeremy R Cole, Michael JQ Zhang, Daniel Gillick, Julian Martin Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. Selectively answering ambiguous questions. *arXiv preprint arXiv:2305.14613*.
- A Philip Dawid. 1982. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610.
- Morris H DeGroot and Stephen E Fienberg. 1983. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22.
- Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. *arXiv preprint arXiv:2003.07892*.
- Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. [Alpaca-farm: A simulation framework for methods that learn from human feedback](#).
- Adam Fisch, Tal Schuster, Tommi Jaakkola, and Regina Barzilay. 2020. Efficient conformal prediction via cascaded inference with expanded admission. *arXiv preprint arXiv:2007.03114*.
- Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.
- Ana Valeria González, Gagan Bansal, Angela Fan, Yashar Mehdad, Robin Jia, and Srinivasan Iyer. 2021. Do explanations help users detect errors in open-domain qa? an evaluation of spoken vs. visual explanations. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1103–1116.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.
- Peter A Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton T Morrison. 2018. Worldtree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. *arXiv preprint arXiv:1802.03052*.
- Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.
- Adam Tauman Kalai and Santosh S. Vempala. 2023. [Calibrated language models must hallucinate](#).
- Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. *arXiv preprint arXiv:2006.09462*.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.
- Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.
- Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents’ overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*.
- Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. 2021. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.
- OpenAI. 2023. Gpt-4 technical report. <https://cdn.openai.com/papers/gpt-4.pdf>.
- Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. 2023. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. In *International Conference on Machine Learning*, pages 26837–26867. PMLR.

- Suzanne Petryk, Spencer Whitehead, Joseph E Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. 2023. Simple token-level confidence improves caption correctness. *arXiv preprint arXiv:2305.07021*.
- John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269.
- Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q Tran, Yi Tay, and Donald Metzler. 2022. Confident adaptive language modeling. *arXiv preprint arXiv:2207.07061*.
- Tal Schuster, Adam Fisch, Tommi Jaakkola, and Regina Barzilay. 2021. Consistent accelerated inference via confident adaptive transformers. *arXiv preprint arXiv:2104.08803*.
- Andy Shih, Dorsa Sadigh, and Stefano Ermon. 2023. Long horizon temperature scaling. *arXiv preprint arXiv:2302.03686*.
- Chenglei Si, Navita Goyal, Sherry Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daumé III, and Jordan Boyd-Graber. 2023. Large language models help humans verify truthfulness—except when they are convincingly wrong. *arXiv preprint arXiv:2310.12558*.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.
- Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR.
- Ellen M Voorhees and Dawn M Tice. 2000. Building a question answering test collection. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 200–207.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.
- Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.
- Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 295–305.
- Theodore Zhao, Mu Wei, J Samuel Preston, and Hoi-fung Poon. 2023. Automatic calibration and error correction for large language models via pareto optimal self-supervision. *arXiv preprint arXiv:2306.16564*.
- 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. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#).
- Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: Expressions of overconfidence and uncertainty in language models. *arXiv preprint arXiv:2302.13439*.
- Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.