Learning from Past Experience: Confidence Expression Calibration in Language Models via Historical Evaluation

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have exhibited remarkable performance across various downstream tasks, but they still generate inaccurate or false information with a confident tone. One potential solution is to improve LLMs' ability to express calibrated confidence, ensuring that the confidence scores align well with the true probability of the generated answers being correct. However, leveraging the intrinsic abilities of LLMs or the output logits has proven ineffective in capturing response uncertainty. Therefore, drawing inspiration from cognitive diagnostics, we propose Learning from Past experience (LePe) to enhance the capability for confidence expression. We first identify three key problems: **(1) How to capture the inherent confidence of the LLMs? (2) How to teach the LLMs to express confidence? (3) How to verify the confidence expression of the LLMs?** To address these challenges, we design a three-phase framework within LePe. In addition, to accurately capture the confidence of an LLM when constructing the training data, we design a complete pipeline including question preparation and answer sampling. Experimental results across multiple datasets demonstrate that our proposed method consistently enables LLMs to provide reliable confidence scores.

1 Introduction

While large language models (LLMs) have achieved exceptional success across diverse domains (Guo et al., 2023; Han et al., 2024; Achiam et al., 2023), their lack of a reliable mechanism to measure confidence in their outputs marks a key contrast with human cognition (Wang et al., 2022; Shuster et al., 2021).

In human cognition, the calibration of confidence serves dual purposes: it facilitates calibrated decision-making and enables self-reflective awareness of knowledge boundaries through quantified uncertainty articulation (Gutierrez de Blume and Schraw, 2014; Stoten, 2019). Similarly, for LLMs, accurate uncertainty estimation can mitigate hallucination risks in generation tasks and provide actionable insights into response reliability for end users. Moreover, it can enable diagnostic feedback loops that help identify systemic model weaknesses for targeted optimization (Liu et al., 2024; Yang et al., 2023).

Accurate estimation of the uncertainty in the LLM text generation process is crucial (Xu et al., 2023; Zhao et al., 2024; Abbasi-Yadkori et al., 2024). Existing approaches to confidence estimation in LLMs can be broadly categorized into three types. The first approach is **prompt-based verbal confidence** (Zhang et al., 2024; Lin et al., 2022), which directly requires the model to give a confidence level after generating the text. The second approach focuses on indirectly estimating confidence by quantifying **properties of the probability distribution** associated with the content generated by the model(Kadavath et al., 2022; Li et al., 2024b; Xu et al., 2024). The third approach empowers the model confidence estimation through a **fine-tuning strategy**(Liu et al., 2023; Lin et al., 2022; Abbasi-Yadkori et al., 2024).

However, these confidence estimation approaches exhibit notable limitations. The first approach requires the model to generate outputs without addressing its inherent lack of internal confidence calibration. As a result, the predicted confidence often diverges significantly from the actual correctness of the output. This leads to issues such as overconfidence, where LLMs assign high confidence to incorrect outputs. Moreover, this approach relies on task-specific prompt engineering, which limits scalability. The second, more widely adopted approach estimates confidence based on statistical patterns in the model's internal representations. However, the mapping between these statistical signals and true semantic correctness is

not deterministic. Misjudgments may arise due to calibration bias, logical inconsistencies, or interference from out-of-domain knowledge. The third approach, which relies on fine-tuning, also faces notable limitations. These methods either require extensively annotated data or are restricted to specific domains, hindering their ability to generalize across diverse application scenarios.

In this paper, inspired by the **Cognitive Diagnostics** (Tatsuoka, 1983) approach for assessing students' ability levels, we propose a method of Learning from Past experience (LePe) to enhance the LLM's capability of confidence expression. In Cognitive Diagnosis, student knowledge mastery is modeled by analyzing their performance on prior experiences, thereby enabling an accurate assessment of their degree of knowledge acquisition and their potential performance on similar problems in the future. Similarly, LePe estimates the model's capability level by leveraging its historical performance, thereby guiding the model to generate confidence scores in its outputs. LePe consists of three main phases: the testing phase, the learning phase, and the verification phase. In the **testing phase**, we aim to capture the inherent confidence of LLM by assessing its performance across a predefined set of questions. In the **learning phase**, we transform historical performance data into interpretable calibration signals that can be used to fine-tune the LLM so that it learns how to express its confidence. Finally, in the **verification phase**, we evaluate the model's ability to generalize its calibrated confidence estimates to previously unseen problems.

However, during the data construction process in the testing phase, context sensitivity (Giallanza and Campbell, 2024) led to inconsistent outputs for identical questions presented in varying contexts. This poses a significant challenge to obtaining reliable confidence estimates from LLMs. To address this, we designed a comprehensive pipeline incorporating multi-faceted mitigation strategies, including *question mutation* and *hybrid sampling*. The question mutation method applies various transformations to questions and answer options—without altering their underlying semantics—to test the robustness of LLM-generated responses. The hybrid sampling strategy employs multiple sampling techniques to derive more representative intrinsic confidence estimates from the model.

Our main contributions are summarized as follows:

- Inspired by Cognitive Diagnostics, we propose a novel method, LePe, to generate calibrated confidence scores for answers produced by LLMs.
- We develop a comprehensive data construction and training pipeline to capture the LLM's underlying confidence patterns and mitigate contextual generation bias.
- Through extensive experiments on multiple datasets with various open-source LLMs, we demonstrate that our method yields confidence scores that are well-aligned with answer correctness. Notably, on the GSM8K dataset, selecting the top 10% of responses based on confidence leads to an accuracy improvement of nearly 50% for Llama2-13B.

2 Related Work

2.1 Self-awareness of LLMs

Although LLMs are equipped with extensive parametric knowledge, several studies highlight their lack of self-awareness in recognizing the boundaries of their competence (Wang et al., 2024). Prior studies on LLM self-awareness primarily focus on mapping their knowledge boundaries (Yin et al., 2023; Ren et al., 2023). These methods aim to help LLMs better leverage their intrinsic knowledge, thereby reducing hallucinations when encountering unfamiliar questions. The Inference-Time Intervention (ITI) method (Li et al., 2024a) achieves this by adjusting model activations along factuality-related heads during inference, thus promoting the generation of more truthful responses. FactTune (Tian et al., 2023a) employs Direct Preference Optimization (DPO) (Rafailov et al., 2024) to guide LLMs toward generating responses that better align with external knowledge. Similarly, Srivastava et al. (2022) assess LLMs' ability to delineate their knowledge boundaries using 23 pairs of answerable and unanswerable multiple-choice questions. However, these methods often enforce overly conservative behavior: when confronted with uncertain questions, LLMs frequently choose not to respond at all, rather than attempting to reason from known information or offer a speculative answer—thereby limiting their practical utility.

2.2 Confidence elicitation in LLMs

Confidence elicitation refers to the process of estimating the level of confidence in an LLM's re-

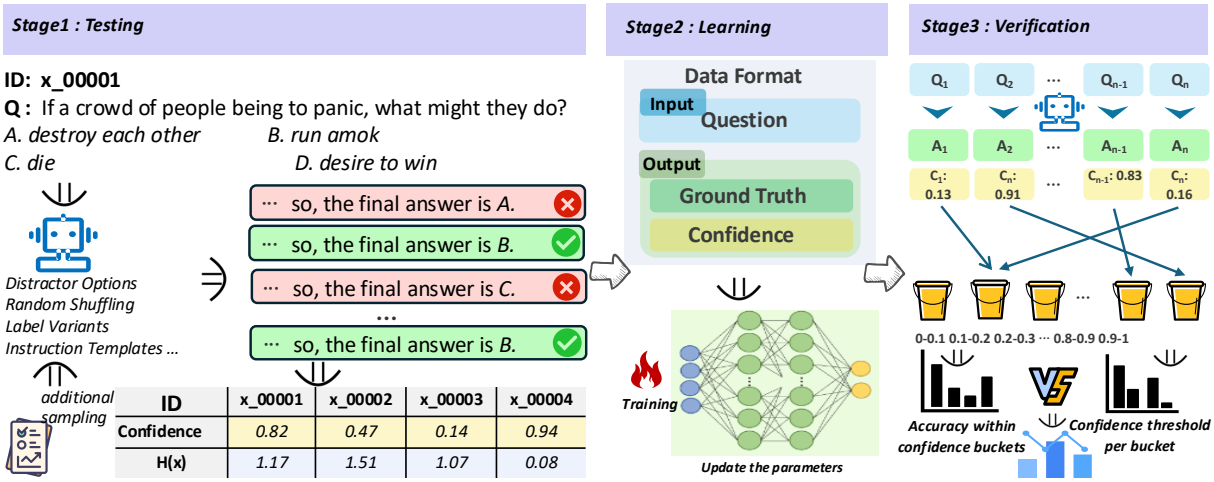

Figure 1: The pipeline of our proposed method LePe. **We determine the confidence levels of LLMs in categorical problem-solving through multiple independent sampling tests conducted in prior experimental evaluations**, rather than relying on intrinsic signals derived from the models' own output mechanisms.

sponse without requiring fine-tuning or access to proprietary model internals (Xiong et al., 2023). Existing methods are broadly classified into two categories. The first category employs carefully designed prompts to simultaneously guide answer generation and elicit verbalized confidence. Branwen (2020) demonstrates GPT-3's ability to express uncertainty on basic factual queries using few-shot prompting, marking the beginning of prompt-based confidence elicitation methods. Lin et al. (2022) formalize this idea by introducing the concept of *verbalized confidence*, where LLMs are explicitly prompted to express their certainty. Building on this, Xiong et al. (2023) expand the design space by proposing consistency-based and hybrid prompting methods. Zhou et al. (2023) attempt to inject uncertainty-related language into prompts, hoping that models mirror this in their responses. However, this often leads to reduced answer accuracy, particularly for complex tasks.

The second category of methods indirectly estimates confidence by analyzing the statistical properties of the output probability distribution. Kadavath et al. (2022) propose incorporating a dedicated *Value Head* to probe the self-assessed confidence of LLMs. Similarly, Kuhn et al. (2023) find that higher semantic diversity in outputs correlates with lower model confidence.

Overall, current approaches rely heavily on the inherent capabilities and internal signals of LLMs to elicit confidence. While effective to some extent, they are constrained by task complexity and model limitations—especially in reasoning-intensive scenarios where LLMs often exhibit overconfidence. In contrast, we treat the ability to express confidence as a meta-capability that should be explicitly trained and calibrated within the LLM through targeted learning from historical performance.

3 Methods

3.1 Task Formalization

To evaluate the capability of LLMs to internalize knowledge through prior experience, we propose a quantitative evaluation framework grounded in cognitive diagnostic theory. Given a set of questions $Q = \{q_1, q_2, \ldots, q_n\}$, we query the model k times independently for each question $q \in Q$ and compute its empirical accuracy as:

$$c(q) = \frac{1}{k} \sum_{i=1}^{k} I(a_i = a^*), \tag{1}$$

where a^* denotes the ground truth answer and a_i is the model's i-th generated answer. The indicator function $I(\cdot)$ returns 1 if the predicted answer matches the ground truth, and 0 otherwise. We interpret $c(q)$ as a proxy for the model's inherent confidence on question q.

Based on this, we construct a supervised learning dataset:

$$D = \{(q, a \oplus c(q)) \mid q \in Q\}, \tag{2}$$

where $\oplus$ denotes the concatenation of the answer and its associated confidence score into a joint representation. This formulation treats the confidence score $c(q)$ as an explicit signal reflecting the model's prior knowledge, enabling a

meta-learning evaluation framework that assesses the model's efficiency in knowledge acquisition over D.

3.2 Learning from Past Experience (LePe)

To enhance the ability of LLMs to express confidence, we raise three core research questions corresponding to the three phases of our proposed approach: (1) How can the intrinsic confidence of LLMs be effectively captured? (2) How can LLMs be trained to express confidence appropriately? (3) How can the accuracy of LLMs' expressed confidence be evaluated?

Testing Phase. In this phase, we aim to estimate the model's inherent confidence by evaluating its historical performance across multiple contexts. Given a set of questions $Q = \{q_1, \ldots, q_n\}$, each question q_i is posed to the model M for k independent trials, resulting in a set of answers $\{a_{i1}, a_{i2}, \ldots, a_{ik}\}$. Each answer a_{ij} corresponds to a model output, denoted as $M(q_i) \rightarrow a_{ij}$. We associate each answer with a correctness label $p_{ij} \in \{0, 1\}$, where $p_{ij} = 1$ if the answer is correct and 0 otherwise. Consequently, each instance is represented as a triplet (q_i, a_{ij}, p_{ij}). These records form the basis for constructing a training dataset used in the subsequent learning phase.

Learning Phase. We utilize the collected confidence signals to construct instruction-style training data and apply instruction fine-tuning (Stiennon et al., 2020; Ouyang et al., 2022) to guide the model in expressing confidence. This step aims to bridge the gap between intrinsic confidence and verbalized confidence output, enabling the model to articulate its uncertainty more faithfully.

Verification Phase. In this phase, we evaluate the calibration performance of the fine-tuned model on unseen questions. Specifically, we assess how closely the model's predicted confidence aligns with the empirical probability of correctness. A model is considered well-calibrated if its confidence estimates match the true likelihood of being correct. Formally, this is defined as:

$$P(y = \hat{y} \mid \text{conf} = z) = z, \quad \forall z \in [0, 1], \qquad (3)$$

where y denotes the model's predicted answer, $\hat{y}$ is the ground truth, and conf is the model's expressed confidence score. Calibration quality is evaluated by comparing the predicted confidence to the actual accuracy across confidence intervals.

3.3 Training Data Construction

To ensure the consistency of LLM responses, mitigate context sensitivity, and better capture intrinsic model uncertainty, we adopt two key strategies: **question mutation** and **answer sampling**.

Question Mutation. To improve robustness and reduce generation variance, we implement a *two-step sampling strategy*: In the first step, we sample each question k times, resulting in $k \times n$ outputs across n questions. We then measure the consistency of responses for each question. If the model's responses are relatively consistent, we assume high certainty and no further sampling is performed. Otherwise, questions showing high variation are flagged as ambiguous. In the second step, we select a proportion α (where $0 < \alpha < 1$) of questions with the highest entropy for each question type. These questions are considered to exhibit high ambiguity, and we perform k additional sampling iterations for each of them to better capture the distribution of possible model responses.

We quantify ambiguity using entropy over the model's answer distribution:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i), \qquad (4)$$

where X denotes the answer space, x_i is a distinct answer, and $p(x_i)$ is its empirical probability. A higher entropy indicates more uncertainty.

To reduce the influence of probabilistic fluctuations and prompt variations, we introduce controlled mutations at both the question and answer levels:

- **Distractor Options (DisO)**: We introduce misleading options such as *"None of the above"* and *"All of the above"* to evaluate robustness.
- **Random Shuffling (RS)**: Answer choices are randomly reordered each time the question is presented to mitigate position bias.
- **Label Variants**: We vary answer labels across formats—uppercase letters (A–D), lowercase letters (a–d), Arabic numerals (1–4), and Roman numerals (i–iv or I–IV).
- **Instruction Templates**: Multiple prompt templates, including few-shot examples and *Chain-of-Thought Prompts (COTP)* (Wei et al., 2022), are used to guide model reasoning.

The original question: Sammy wanted to go to where the people were. Where might he go?
A. race track B. populated areas C. the desert D. apartment
The varied question: *(TaskP)* Examine the following options carefully and select the correct one.
(COTP) Please select the correct option from the provided choices and offer a comprehensive problem-solving process.
(question) Sammy wanted to go to where the people were. Where might he go?
(Shuffle options and change option label)
1. populated areas 2. apartment 3. the desert 4. race track
(DisO) 5. All of the above / None of the above

Input: *(confidence expression prompt)* For the following question, please select the correct option, and provide your confidence in this answer.
Google Maps and other highway and street GPS services have replaced what?
A. united states B. Mexico C. countryside D. atlas E. None of the above
Output: The correct answer is D. atlas. *(Conf)* ***My confidence is 61.5%.***

Figure 2: Structure of instruction-format training data for confidence calibration. The top shows a sample input with an appended confidence score. The bottom illustrates the data components: question, answer, and confidence statement.

An example of a mutated question presented to the LLM is shown in Figure 2(top).

Answer Sampling. We adopt a random sampling decoding strategy augmented with a hybrid approach that integrates both Top-k and Top-p sampling. This combination enables the model to generate a diverse set of responses, thereby capturing a wider range of confidence levels. Specifically, for each input question $q_i \in Q$, the model produces k distinct responses, which collectively constitute the dataset of answer records.

To fine-tune the model's ability to express confidence, we use the collected answer records as training data. Crucially, not all model responses are correct. Including incorrect answers—especially those with low confidence—can help the model learn to appropriately signal uncertainty. However, excessive inclusion of such responses may exacerbate hallucination risks (Huang et al., 2023).

To balance this, we compute a normalized confidence score:

$$Conf = \frac{f_{q_i}}{k}, \tag{5}$$

where f_{q_i} denotes the number of correct answers out of k attempts for question q_i. During training, we format each sample as:

$$\langle \text{Question}, \text{Answer} + \text{Confidence} \rangle,$$

where confidence is expressed as: *"My confidence is [Conf $\times$ 100]%."* This unified format enables the model to learn proper confidence calibration across a range of correct and incorrect outputs while minimizing overfitting to flawed data.

An example of the final instruction format is shown in Figure 2(bottom).

4 Experiments

4.1 Experiment settings

Dataset. We evaluate the quality of confidence estimates across six datasets on three distinct domains, including mathematical domains: *GSM8K*(Cobbe et al., 2021), AIME24[1], commonsense reasoning: *CommonsenseQA*(CSQA)(Talmor et al., 2018), *MMLU* (Hendrycks et al., 2020), and open questions: *TriviaQA*(Joshi et al., 2017), NQ-Open(Kwiatkowski et al., 2019).

Models and Baselines. We consider three models, Llama2-13B(Touvron et al., 2023), Llama3.1-8B(Dubey et al., 2024) and Qwen2.5-7B(Yang et al., 2024). For comprehensive comparison, we consider two categories of baseline methods for confidence estimation: (1) Training-free approaches: ***First-Prob***(Santurkar et al., 2023), ***Verb***(Tian et al., 2023b), ***Multi-Step***(Xiong et al., 2023), ***SE***(Kuhn et al., 2023). (2) Training-based approaches: ***SuC***(Lin et al., 2022), ***P(IK)***(Kadavath et al., 2022).

Metrics. To evaluate the accuracy of generated answers, we employ a string-matching approach to extract the model's final answer and compare it with the ground truth. We use four evaluation metrics to assess the model's performance.

ACC. Represents the average accuracy of the LLM's responses.

AUROC. Area Under the Receiver Operating Characteristic Curve.Measures model's ability to distinguish correct vs. incorrect answers using confidence scores.

r. Pearson Correlation Coefficient. Quantifies linear correlation between confidence scores and

[1]https://huggingface.co/datasets/math-ai/aime24

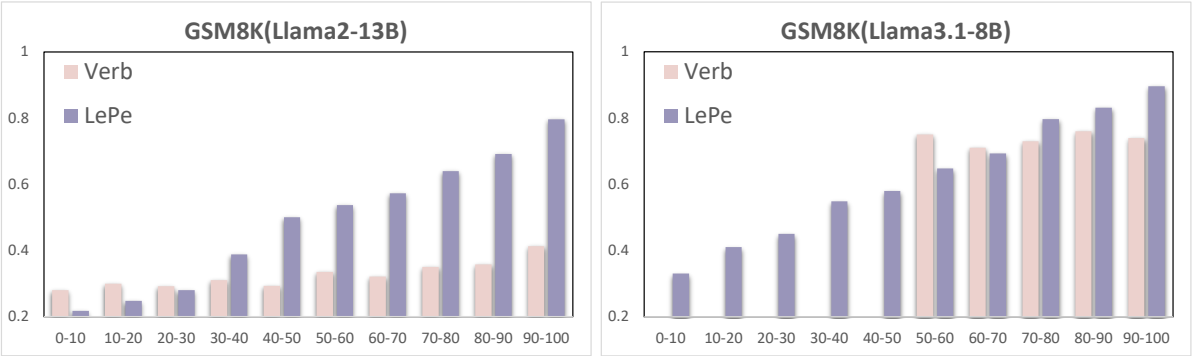

Figure 3: Calibration results on the GSM8K dataset using Llama2-13B and Llama3.1-8B. The horizontal axis shows predicted confidence, and the vertical axis shows actual accuracy. LePe demonstrates strong alignment between predicted and actual confidence, while the Verb method exhibits poor calibration.

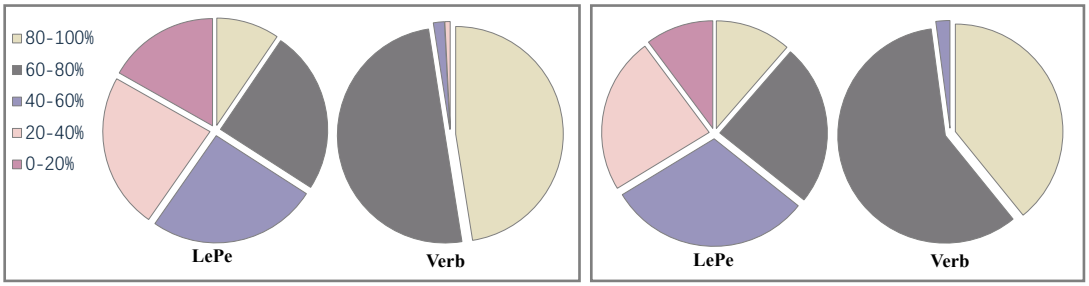

Figure 4: The detailed confidence statistics for LePe and Verb Using Llama2-13B on the GSM8K(left) and CSQA(right) Datasets.

answer correctness.

ECE. Expected Calibration Error. We partition the inference results into B disjoint bins based on the confidence scores, compute the average confidence score for each bin, and compare it with the average true accuracy of the answers within that bin. The ECE is calculated by:

$$ECE = \sum_{b=1}^{B} \frac{s_b}{S} \left| acc(b) - conf(b) \right|, \qquad (6)$$

where b is the b-th bin, B is the total number of bins, s_b is the number of questions in the b-th bin, S is the total number of questions in the test set, $acc(b)$ is the true correctness of the answers in the b-th bin, and $conf(b)$ is the average of the LLM confidence in the b-th bin. The smaller the value, the better.

Implementation Details. The three datasets used in the experiments were trained independently.For the hyperparameters used in our paper, we set $k = 30$ and $\alpha = 0.5$. Our optimizer is AdamW (Loshchilov and Hutter, 2017) with β_1 and β_2 values of 0.98 and 0.99. During training, we set the initial learning rate to 2e-5, the final learning rate to 5e-5. We conduct all our experiments using the NVIDIA A800.

4.2 Main results analysis

As shown in Table 1, although most methods retain the basic functionality of confidence estimation, their actual calibration performance varies significantly. This is evident from the substantial differences in empirical results. For instance, LePe and P(IK) demonstrate strong overall performance. On the Llama2-13B model with the GSM8K dataset, their ECE reach 8.9 and 14.5, respectively. In contrast, methods such as SuC and Verb perform poorly under the same conditions, with ECEs of 28.8 and 29.4. Moreover, when faced with the more challenging AIME24 dataset, which involves long-form mathematical reasoning, most traditional optimization methods suffer catastrophic failures. Their breakdown is striking: on the Llama3.1-8B model, Verb and Multi-step report ECEs of 73.4 and 77.2, indicating a complete collapse in confidence calibration. In comparison, LePe maintains robust performance with an ECE of 18.5, highlighting its strong calibration capability under difficult settings.

	Baselines	GSM8K			AIME24			CSQA			MMLU			TriviaQA			NQ-Open		
		$r\uparrow$	$ECE\downarrow$	$AU\uparrow$	$r\uparrow$	$ECE\downarrow$	$AU\uparrow$	$r\uparrow$	$ECE\downarrow$	$AU\uparrow$	$r\uparrow$	$ECE\downarrow$	$AU\uparrow$	$r\uparrow$	$ECE\downarrow$	$AU\uparrow$	$r\uparrow$	$ECE\downarrow$	$AU\uparrow$
Llama2-13B	*P(IK)*	0.894	14.5	64.8	0.693	31.4	72.1	0.812	29.9	59.5	0.764	17.3	67.6	0.901	18.7	65.0	0.731	18.3	70.7
	First-Prob	0.767	23.4	59.7	0.629	42.0	61.2	0.763	22.4	60.1	0.687	19.4	64.3	0.741	27.6	57.1	0.636	22.1	65.1
	SuC	0.702	28.8	57.3	0.547	37.3	57.3	0.700	27.2	56.7	0.628	22.1	65.2	0.718	23.5	58.2	0.610	24.6	66.4
	Verb	0.558	29.4	56.2	0.284	82.3	14.9	0.453	21.8	58.3	0.659	32.6	61.1	0.644	27.2	53.7	0.637	29.8	62.4
	Multi-Step	0.701	27.2	58.3	0.403	76.4	23.1	0.738	24.1	59.2	0.574	33.4	62.5	0.751	26.4	61.1	0.635	28.7	63.1
	SE	0.673	18.4	68.6	0.748	32.7	65.1	0.712	16.3	65.4	0.727	20.3	69.4	0.772	19.5	63.1	0.705	24.1	70.2
	LePe(ours)	**0.932**	**8.9**	**67.3**	**0.903**	**24.8**	**78.4**	**0.982**	**16.2**	**69.3**	**0.964**	**15.0**	**72.6**	**0.927**	**15.5**	**68.4**	**0.941**	**13.9**	**74.3**
Llama3.1-8B	*P(IK)*	0.793	17.6	72.8	0.765	33.1	67.9	0.874	19.4	**68.7**	0.724	18.3	72.1	0.762	20.4	67.7	0.751	22.4	68.2
	First-Prob	0.812	26.2	66.2	0.609	40.3	65.0	0.834	23.5	66.8	0.627	21.4	68.4	0.790	24.9	65.1	0.646	29.4	66.5
	SuC	0.603	28.4	62.0	0.577	42.7	62.2	0.630	32.7	59.1	0.611	24.7	66.3	0.663	29.7	60.4	0.563	27.3	61.4
	Verb	0.652	20.4	72.9	0.175	73.4	6.1	0.731	28.0	68.4	0.608	31.2	62.7	0.701	30.1	69.1	0.657	34.0	65.2
	Multi-Step	0.694	25.9	65.4	0.354	77.2	16.3	0.680	27.4	67.2	0.671	29.6	62.0	0.713	27.8	66.1	0.603	31.9	63.1
	SE	0.703	17.6	73.5	0.715	20.9	68.5	0.705	21.3	66.7	0.773	17.2	71.2	0.753	19.4	66.4	0.721	22.3	70.4
	LePe(ours)	**0.963**	**13.5**	**76.4**	**0.892**	**18.5**	**73.1**	**0.973**	**16.0**	68.4	**0.913**	**14.3**	**76.2**	**0.983**	**15.5**	**69.8**	**0.937**	**20.9**	**73.1**
Qwen2.5-7B	*P(IK)*	0.757	17.4	68.3	0.763	27.9	66.3	0.731	16.3	68.4	0.789	16.1	69.8	0.873	21.6	67.9	0.814	20.8	72.3
	First-Prob	0.831	25.4	66.4	0.576	35.8	57.4	0.874	26.6	65.2	0.673	30.3	68.0	0.882	25.9	62.3	0.626	24.5	68.5
	SuC	0.642	29.0	57.4	0.548	38.4	60.4	0.671	28.2	63.1	0.594	27.0	62.4	0.624	32.7	58.5	0.507	24.1	63.1
	Verb	0.603	15.3	72.2	0.251	78.7	11.3	0.734	**12.4**	70.3	0.624	29.4	63.3	0.661	22.0	68.4	0.677	33.6	62.4
	Multi-Step	0.662	22.9	68.0	0.405	68.3	31.2	0.691	24.2	65.7	0.574	28.4	65.1	0.721	21.7	66.4	0.631	29.0	65.3
	SE	0.739	18.6	72.1	0.720	25.1	73.5	0.693	19.3	69.4	0.674	22.4	68.3	0.741	22.5	68.4	0.781	23.8	71.8
	LePe(ours)	**0.946**	**11.4**	**72.3**	**0.911**	**21.2**	**76.2**	**0.971**	14.7	**70.6**	**0.952**	**15.6**	**73.1**	**0.992**	**15.2**	**69.2**	**0.927**	**17.4**	**76.2**

Table 1: Results of confidence estimates for all baseline methods. AU means AUROC. The best results are **bolded**, and the second best ones are underlined.

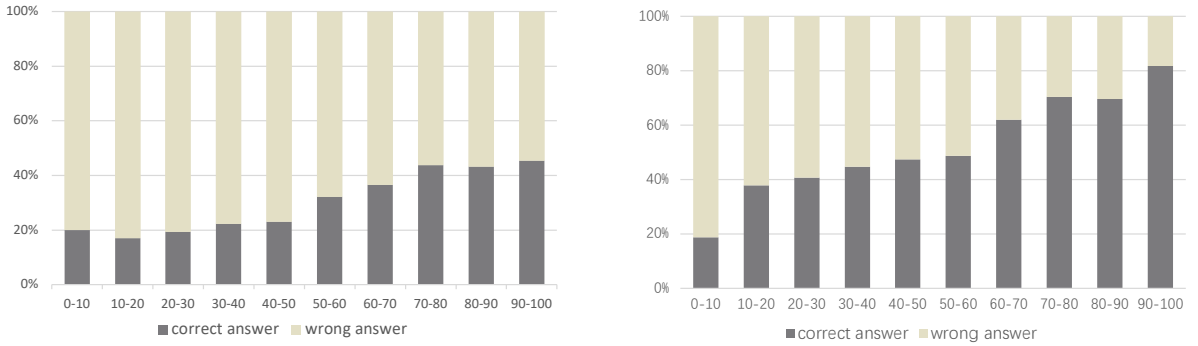

Figure 5: Testing on the out-of-domain datasets. The LLM is trained using LePe on CSQA and tested on OpenBookQA (left) and GSM8K (right).

Dataset	t	Model	ACC	ACC_t	DP
GSM8K	65	Llama2-13B	33.6	**66.5**	29.4%
		Llama3.1-8B	61.7	**77.3**	76.1%
		Qwen2.5-7B	73.4	**82.1**	73.4%
CSQA	85	Llama2-13B	65.6	**85.5**	26.5%
		Llama3.1-8B	77.4	**88.4**	62.9%
		Qwen2.5-7B	81.1	**90.3**	60.2%
TriviaQA	95	Llama2-13B	64.8	**85.1**	48.4%
		Llama3.1-8B	73.9	**87.2**	54.3%
		Qwen2.5-7B	77.3	**89.4**	57.1%

Table 2: The accuracy performance of our foundational models is evaluated across three distinct datasets at varying acceptable confidence thresholds. DP represents the proportion of data for which the confidence level exceeds the t.

As shown in Figure 3 and Figure 4, the use of the LePe method results in a strong positive correlation between the model's predicted confidence and its actual accuracy, indicating effective calibration. In contrast, traditional methods exhibit notable issues. For example, on the Llama2-13B model with the GSM8K dataset, the Verb method shows only a weak correlation between predicted confidence and true accuracy. Even when the model assigns high confidence scores (90%–100%), the actual accuracy remains around 40%. Furthermore, on the Llama3.1-8B model, the problem of overconfidence becomes more pronounced. The model rarely produces low-confidence predictions and shows similar accuracy across different confidence intervals. This undermines the reliability and usability of its confidence scores.

4.3 Confidence threshold analysis

As shown in Table 2, we analyze model confidence and identify a practical threshold for evaluating the reliability of LLMs. This threshold enables actionable reliability: predictions with confidence

Baselines	GSM8K				CSQA				TriviaQA			
	ACC	ACC_{t5}	ACC_{t3}	ACC_{t1}	ACC	ACC_{t5}	ACC_{t3}	ACC_{t1}	ACC	ACC_{t5}	ACC_{t3}	ACC_{t1}
P(IK)	30.4	45.1	50.5	55.7	66.9	74.3	78.6	80.1	66.2	72.3	74.9	78.2
First-Prob	30.4	40.2	43.7	47.1	62.5	69.2	73.1	76.1	63.1	68.8	70.5	73.3
SuC	31.0	36.6	37.8	39.2	60.1	65.3	67.7	69.4	62.8	65.3	67.2	69.9
Verb	31.0	35.2	38.8	41.3	64.3	67.1	69.5	72.6	65.1	66.7	69.3	71.4
Multi-Step	31.3	39.4	41.0	45.9	63.7	68.7	71.4	74.5	63.3	69.2	72.0	73.7
SE	32.6	40.1	45.7	49.2	64.7	71.2	75.4	78.2	65.1	71.0	73.2	78.1
LePe(ours)	33.6	**57.7**	**65.3**	**83.3**	65.6	**81.3**	**84.9**	**93.4**	64.8	**84.7**	**87.6**	**93.8**

Table 3: The ACC of results at different confidence score levels in Llama2-13B. ACC_{t5} represents the accuracy of the results whose confidence scores are within the top 50%, and so on.

Method	$r\uparrow$	$ECE\downarrow$	$AUROC\uparrow$
LePe(Ours)	0.982	16.2	69.3
w/o DisO	0.940(-0.042)	17.2(+1.0)	68.1(-1.2)
w/o RS	0.963(-0.019)	17.4(+1.2)	67.4(-1.9)
w/o COL	0.958(-0.024)	16.9(+0.7)	67.1(-2.2)
w/o ALL	0.917(-0.065)	18.2(+2.0)	65.8(-3.5)

Table 4: Ablation study of LePe method on CommonsenseQA with Llama2-13B. DisO: distractor options. RS: randomly shuffled. COL: change option labels.

scores above the threshold demonstrate statistically reliable performance, while those below indicate uncertainty and require further handling. For instance, on the Llama-13B model with the GSM8K dataset, LePe achieves an accuracy of 66.5% on predictions above the threshold—nearly twice the baseline overall accuracy of 33.6%.

As shown in Figure 3, we further filter predictions based on different confidence levels. At all levels, LePe consistently outperforms the baseline. For example, on the TriviaQA dataset, the top 10% most confident predictions achieve an accuracy of 93.8%, representing a 29% improvement over the overall accuracy.

4.4 Generalized experimental analysis

As shown in Figure 5, to assess the generalizability of our method, we evaluate Llama2-13B trained with LePe on CSQA, testing its confidence estimation on both in-domain (OpenBookQA) and out-of-domain (GSM8K) datasets.

Our analysis indicates that LePe maintains effective calibration across both in-domain and out-of-domain scenarios. Specifically, Llama2-13B trained on CSQA demonstrates reliable confidence estimation on OpenBookQA and GSM8K, with a significant positive correlation between confidence estimates and accuracy. This consistent trend suggests robust generalization, indicating LePe's ability to provide reliable confidence estimates even on tasks beyond its training domain.

4.5 Ablation experiment analysis

As shown in Table 4, we conduct ablation studies demonstrating that all mutation strategies are essential to the calibration effectiveness of LePe. Performance degrades progressively with component removal (r:0.982→0.940; ECE:+7.4%), reaching worst-case metrics without strategies, demonstrating their synergistic integration requirement.

5 Conclusion

In this paper, we present a method of learning from past experience to enhance the LLMs' capability for confidence expression, enabling LLMs to provide answers along with corresponding confidence levels. We first design a general pipeline to obtain the actual performance of the LLMs on the problem. Further, we utilize the performance records to construct the dataset for instruction fine-tuning so that the LLMs learn to express confidence in the generated answers. We conduct experiments on three open-source language LLMs to demonstrate the effectiveness of our method. The consistent experimental results across multiple tasks affirm that our method endows the LLMs with confidence expression capability.

6 Limitations

This study mainly evaluates the performance of LePe on general-purpose tasks (reasoning, common sense, openness), and does not validate its applicability in specialized domains (e.g., legal, medical). Moreover, as mentioned earlier, using the confidence expressed by the model, we can identify the model's weaknesses and further improve them in a targeted manner, allowing LLMs to continue to evolve. In this paper, we propose a general method to enhance the model's capability to express confidence, but do not discuss the impact of ability on the continuous learning of LLMs.

References

Yasin Abbasi-Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesv'ari. 2024. To believe or not to believe your llm. *ArXiv*, abs/2406.02543.

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

Gwern Branwen. 2020. Gpt-3 nonfiction- calibration. Technical report, The institution that published. Last accessed on 2022-04-24.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Bap tiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Cantón Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab A. AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriele Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guanglong Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Laurens Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Ju-Qing Jia, Kalyan Vasuden Alwala, K. Upasani, Kate Plawiak, Keqian Li, Ken-591 neth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuen ley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Babu Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melissa Hall Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri S. Chatterji, Olivier Duchenne, Onur cCelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasić, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Ro main Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Chandra Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whit ney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yiqian Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zhengxu Yan, Zhengxing Chen, Zoe Papakipos, Aaditya K. Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adi Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Ben Leonhardi, Po-Yao (Bernie) Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Shang-Wen Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzm'an, Frank J. Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory G. Sizov, Guangyi Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Han Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai

Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kaixing(Kai) Wu, U KamHou, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, A Lavender, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollár, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sung-Bae Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Andrei Poenaru, Vlad T. Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xia Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *ArXiv*, abs/2407.21783.

Tyler Giallanza and Declan Iain Campbell. 2024. Context-sensitive semantic reasoning in large language models. In *ICLR 2024 Workshop on Representational Alignment*.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Antonio Gutierrez de Blume and Gregory Schraw. 2014. Effects of strategy training and incentives on students' performance, confidence, and calibration. *The Journal of Experimental Education*, pages 1–19.

Haixia Han, Jiaqing Liang, Jie Shi, Qi He, and Yanghua Xiao. 2024. Small language model can self-correct. In *AAAI Conference on Artificial Intelligence*.

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

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ArXiv*, abs/2311.05232.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *ArXiv*, abs/1705.03551.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zachary Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, John Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom B. Brown, Jack Clark, Nicholas Joseph, Benjamin Mann, Sam McCandlish, Christopher Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know. *ArXiv*, abs/2207.05221.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024a. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36.

Moxin Li, Wenjie Wang, Fuli Feng, Fengbin Zhu, Qifan Wang, and Tat-Seng Chua. 2024b. Think twice before assure: Confidence estimation for large language models through reflection on multiple answers. *arXiv preprint arXiv:2403.09972*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.

Linyu Liu, Yu Pan, Xiaocheng Li, and Guanting Chen. 2024. Uncertainty estimation and quantification for llms: A simple supervised approach. *arXiv preprint arXiv:2404.15993*.

Xin Liu, Muhammad Khalifa, and Lu Wang. 2023. Litcab: Lightweight calibration of language models on outputs of varied lengths. *arXiv preprint arXiv:2310.19208*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *arXiv preprint arXiv:2307.11019*.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? *ArXiv*, abs/2303.17548.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

David Stoten. 2019. Metacognition, calibration, and self-regulated learning: an exploratory study of undergraduates in a business school. *Learning: Research and Practice*, 5:1–24.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Kikumi K. Tatsuoka. 1983. Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20(4):345–354.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D Manning, and Chelsea Finn. 2023a. Fine-tuning language models for factuality. *arXiv preprint arXiv:2311.08401*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023b. 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, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024. "my answer is c": First-token probabilities do not match text answers in instruction-tuned language models. *ArXiv*, abs/2402.14499.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and

Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.

Rongwu Xu, Brian Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2024. The earth is flat because...: Investigating LLMs' belief towards misinformation via persuasive conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16259–16303, Bangkok, Thailand. Association for Computational Linguistics.

Rongwu Xu, Brian S. Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2023. The earth is flat because...: Investigating llms' belief towards misinformation via persuasive conversation. *ArXiv*, abs/2312.09085.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, and Shanghaoran Quan. 2024. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

Yuchen Yang, Houqiang Li, Yanfeng Wang, and Yu Wang. 2023. Improving the reliability of large language models by leveraging uncertainty-aware in-context learning. *arXiv preprint arXiv:2310.04782*.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? *arXiv preprint arXiv:2305.18153*.

Mozhi Zhang, Mianqiu Huang, Rundong Shi, Linsen Guo, Chong Peng, Peng Yan, Yaqian Zhou, and Xipeng Qiu. 2024. Calibrating the confidence of large language models by eliciting fidelity. *arXiv preprint arXiv:2404.02655*.

Xinran Zhao, Hongming Zhang, Xiaoman Pan, Wenlin Yao, Dong Yu, Tongshuang Wu, and Jianshu Chen. 2024. Fact-and-reflection (far) improves confidence calibration of large language models. *ArXiv*, abs/2402.17124.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. *arXiv preprint arXiv:2302.13439*.