CHAIN-OF-THOUGHT DEGRADES ABSTENTION IN LARGE LANGUAGE MODELS, UNLESS INVERTED

Anonymous authors
Paper under double-blind review

ABSTRACT

For Large Language Models (LLMs) to be reliably deployed, models must effectively know when not to answer: *abstain*. Chain-of-Thought (CoT) prompting has been gained popularity for improving model performance by ensuring structured outputs that follow a logical sequence. In this paper, we first investigate how current abstention methods perform with CoT outputs, finding that direct use of reasoning traces can degrade performance of existing abstention methods by more than 5%. As a result, we introduce a new framework for thinking about hallucinations in LLMs not as answering a question incorrectly but instead as LLMs answering the *wrong* question. Based on this framework, we develop a new class of state-of-the-art abstention methods called **Trace Inversion**. First, we generate the reasoning trace of a model. Based on only the trace, we then reconstruct the most likely query that the model responded to. Finally, we compare the initial query with the reconstructed query. Low similarity score between the initial query and reconstructed query suggests that the model likely answered the question incorrectly and is flagged to abstain. We perform extensive experiments to find impressive performance gains with our Trace Inversion methods. The code is publicly available at: `https://anonymous.4open.science/r/trace-inversion-9EE0/`.

1 INTRODUCTION

Large Language Models (LLMs) have demonstrated impressive performance across question answering (Tan et al., 2023; Li et al., 2024; Yang et al., 2024b), text-generation (Mo et al., 2024; Kurihara et al., 2025; Wu, 2024; Mahapatra & Garain, 2024), and complex problem solving tasks (Ge et al., 2023; Gao et al., 2024; Pei et al., 2025; Renze & Guven, 2024; Singhi et al., 2025). However, LLMs also have a tendency to "hallucinate" information (Zhang et al., 2025b; Yao et al., 2023; Tonmoy et al., 2024; Huang et al., 2025), generate overly certain responses (Xiong et al., 2023; Tao et al., 2024; Yang et al., 2024a), answer with conflicting or incomplete information (Xu et al., 2024a; Lee et al., 2024; Tan et al., 2024; Xu et al., 2023), and perpetuate social biases (Wan et al., 2023; Kong et al., 2024; Taubenfeld et al., 2024). For LLMs to be reliably deployed, models must be able to ***abstain*** from answering questions they do not know the answers to. Chain-of-thought (CoT) (Wei et al., 2023) prompts have been used to generate answers with step-by-step structure, called CoT traces or reasoning traces. In doing so, users require the model's output to have more structure and logical processing, inherently beneficial for domains like mathematical problem solving (Fung et al., 2023; Yang et al., 2024c). It has been empirically shown that language models have improved performance if they output reasoning trace tokens first (Nye et al., 2021; Zhang et al., 2022; Hsieh et al., 2023), resulting in an interest to fine-tune models with these traces. Reasoning fine-tuning LLMs has provided performance gains on various benchmarks (Vaillancourt & Thompson, 2024; Zhang et al., 2025a; Sprague et al., 2024; Zelikman et al., 2022; Luo et al., 2025). However, reasoning fine-tuning has been shown to further degrade abstention ability (Kirichenko et al., 2025). We thus pose the question: *can we use reasoning traces to improve model abstention?*

Previous approaches have posed abstention as a function of uncertainty, where a model should abstain from generating low-confidence outputs. These abstention methods have employed techniques to estimate the model's confidence and then ensure the model abstains if the confidence score for a response falls below some threshold (Feng et al., 2024). Model confidence has been calculated

using token probabilities (Radford et al., 2019; Gupta et al., 2024) or even verbalized confidence from the model itself (Lin et al., 2022; Tian et al., 2023). Self-consistency of generated reasoning traces has also been used as a metric of model certainty (Wang et al., 2022; Besta et al., 2024) where more inconsistent or contradictory traces signify that the model should abstain. While these methods have the potential to build upon a rich landscape of uncertainty quantification research, model certainty may not be the best signal for model correctness (Xiao et al., 2025; von Clarmann et al., 2021), as seen by high-certainty hallucinations (Simhi et al., 2025) where models confidently answer questions incorrectly. Instead, we position model abstention as a decision based on the model's knowledge gap corresponding to the user's question. But how can we detect such knowledge gaps? Prompting approaches and multi-LLM systems review model responses in an attempt to identify gaps in model knowledge (Wen et al., 2025; Feng et al., 2024). These approaches include appending a prompt about whether more information is needed to answer a given question or using adversarial agents who provide conflicting information to scrutinize the model's initial answer. However, several works have explored how LLM errors are may be correlated with one another (Laurito et al., 2024; Kim et al., 2025), potentially causing issues with prompting and multi-LLM hallucination detection.

In this work, we first investigate how current confidence estimation and answer reviewing methods for abstention perform with CoT outputs. In addition to exploring how current methods perform with CoT outputs, we propose a new class of methods with reasoning traces called **Trace Inversion**. We introduce a new framework for thinking about abstention in LLMs as query-based knowledge gap detection. In our framework, an abstention decision, or potential hallucination, is a consequence of the model answering the *wrong* question rather than the model answering a question incorrectly. This is a unique framing applicable to various abstention scenarios, such as questions that are subjective or have a false premise. First, we generate the reasoning trace of a model. Based on only the trace, we then reconstruct the most likely query that the model responded to. Finally, we compare the initial query with the reconstructed query. Low similarity score between the initial query and reconstructed query suggests that the model likely answered the question incorrectly and is flagged to abstain (see Figure 1). We perform extensive experiments on eight datasets across domains with five diverse models.

The main contributions of this work are as follows:

1. We demonstrate how direct use of reasoning traces can degrade performance of existing abstention methods by an average 3.47%, reaching >5% for reading comprehension and bias benchmarking datasets.
2. We introduce a new framework to think about hallucinations in LLMs as models answering a different question than the one posed by the user.
3. We provide a new set of state-of-the-art method in abstention by *inverting* reasoning traces, resulting in performance gains up to 19.8%.

2 Related Work

Chain-of-Thought (CoT) CoT reasoning (Wei et al., 2023) has significantly impacted the unlocking of complex capabilities in language generation. By explicitly eliciting a series of intermediate reasoning steps, in the form of a scratchpad (Nye et al., 2021) or interpretable window, CoT has become a powerful tool in enhancing the performance of LLMs on tasks that require structured and logical processing (Lightman et al., 2023; Lee et al., 2025). Hu et al. (2024) studies this through a theoretical lens by showing CoT as a statistical estimation process, where a model using CoT operates as a Bayesian estimator. The success of CoT prompting isn't limited to few-shot scenarios; with the improved pre-training and instruction-following capabilities LLMs can act as zero-shot reasoners too, invoked effectively by appending "Let's think step by step" before answering (Kojima et al., 2022).

Limitations of Chain-of-Thought While the "interpretable window" of human-like step-by-step reasoning appears to offer an understanding into the internal thinking of LLMs, recent studies (Chen et al., 2025; Arcuschin et al., 2025; Turpin et al., 2023) have revealed this interpretability to be superficial. The perceived effectiveness of this interpretability might not align with the model's true

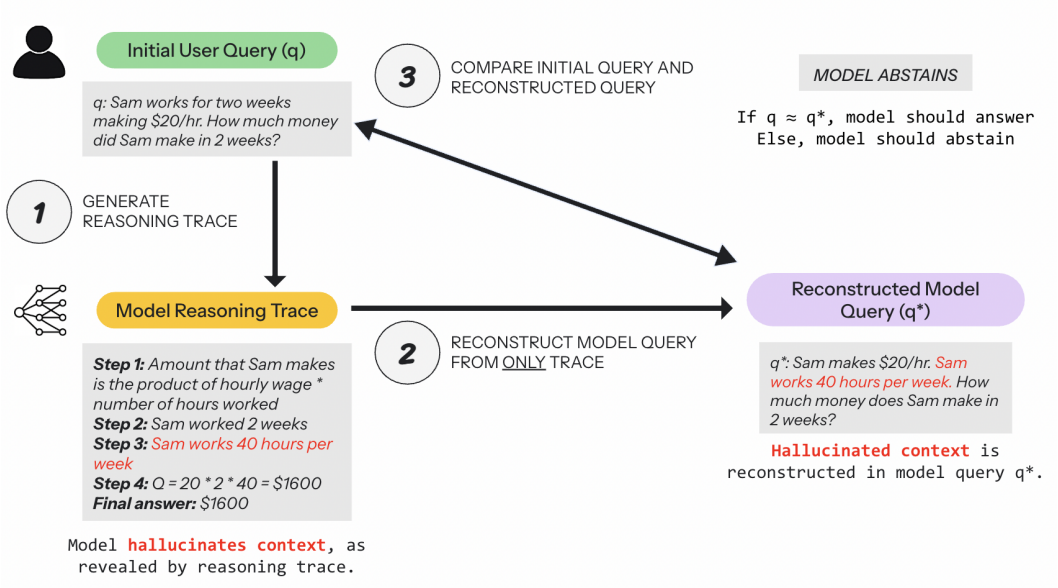

Figure 1: Overview of our three step trace inversion approach. We provide an example of how our method particularly detects subtle hallucinations in a reasoning trace by comparing the user query q with the model-interpreted query q^*. In this provided example, the reconstructed query q^* based on the trace includes the hallucinated context of how many hours Sam works per week. We interpret this as a different question, since the ambiguity from the initial user query q (which is unanswerable without more information) is not present in the model query, hence the model is answering the *wrong question*. Because the user query and model query are meaningfully different, the model abstains.

internal workings (Bhambri et al., 2025; Korbak et al., 2025). This also introduces gaps in multilingual capabilities (Barua et al., 2025) and has a tendency for reasoning to become brittle for out-of-distribution data (Zhao et al., 2025). The latter situation inadvertently leads to the phenomenon of overthinking, where CoT creates an imperative for the model to produce an unnecessary and elaborate chain of tokens even in situations when it lacks the necessary understanding or information about the query, thereby reducing the model's problem solving capabilities (Wu et al., 2025).

Model Abstention As use of LLMs has exploded in various user-facing applications while the interpretability of such models remains limited, the greater community has steered into enforcing reliability mechanisms that address 'abstention' (Wen et al., 2025; Kirichenko et al., 2025), a meta-capability enabling a model to decline providing a definitive answer for uncertain, unanswerable, or potentially harmful prompts. Tomani et al. (2024) have investigated the model's ability to detect its own knowledge gaps and to signal uncertainty as a safeguard against overconfidence or hallucinated generations. Even with a model's statistical uncertainty (via token probabilities), semantic uncertainty, or verbalized uncertainty (Xiong et al., 2023; Xu et al., 2024b; Lin et al., 2022), they often fail to correlate faithfully with actual correctness (Madhusudhan et al., 2025; Yadkori et al., 2024). Feng et al. (2024) overcomes this limitation by exploring multi-LLM collaboration rather than relying on a single monolithic model. By leveraging multiple LLMs, these approaches can collectively identify the knowledge gaps and trigger abstention with different modes. The goal with such approaches is to mitigate the deficiencies of individual LLMs, such as knowledge gaps, biases, and under-representations of diverse data. However, multi-LLM approaches may suffer from error correlation (Kim et al., 2025; Laurito et al., 2024), self-bias (Xu et al., 2024c; Panickssery et al., 2024), and other documented LLM-as-judge limitations (Wang et al., 2024; Szymanski et al., 2025).

3 Reasoning Traces Can Degrade Model Abstention

Despite the proliferation of reasoning fine-tuned models and interest in using reasoning traces for performance gains, the use of CoT outputs in the abstention setting has remained limited. We

investigate how the use of CoT outputs affects current abstention methods. Specifically, we use five baseline methods from two representative groups: confidence estimation and answer reviewing. Confidence estimation methods estimate the model's certainty and then ensure the model abstains if the certainty score for a response falls below some threshold. These methods rely on good calibration between the notions of model certainty and correctness. Answer reviewing methods use LLMs to evaluate outputs in order to identify gaps in knowledge. We compare the abstention performance of baselines when relying solely on the model's final answer versus when incorporating CoT-prompted outputs.

3.1 BASELINES

For confidence estimation methods, we use a held-out development set $\mathcal{H} = \{(q_i, \bar{a}_i)\}_{i=1}^{N}$. For each question q_i, the LLM produces an answer $a_i = \text{LLM}(q_i)$ and calculate a confidence score $p_i \in [0, 1]$. We define correctness labels as

$$y_i = \begin{cases} 1 & \text{if } a_i = \bar{a}_i, \\ 0 & \text{if } a_i \neq \bar{a}_i. \end{cases}$$

Candidate thresholds are taken from a discretized grid $\mathcal{T} = \{0.01, 0.02, \ldots, 0.99\}$. For each threshold $t \in \mathcal{T}$, we apply the abstention rule and compute the abstain error

$$\hat{a}_i(t) = \begin{cases} \text{abstain}, & p_i < t, \\ a_i, & p_i \geq t, \end{cases} \qquad E(t) = \sum_{i=1}^{N} \mathbf{1}\big(p_i < t \ \wedge \ y_i = 1\big) \ + \ \mathbf{1}\big(p_i \geq t \ \wedge \ y_i = 0\big).$$

The first term in $E(t)$ penalizes unnecessary abstentions on correct answers, while the second penalizes failures to abstain on incorrect answers. The abstention threshold is then chosen as $p^* = \arg\min_{t \in \mathcal{T}} E(t)$. At inference time, the model answers if $p_i \geq p^*$ and abstains otherwise (Feng et al., 2024). The following two methods use internal calibration and verbalized calibration to estimate model confidence.

Token probability (`TOKENPROB`) We compute the confidence score p_i for a question using the top-k token probabilities over the entire answer span where P is the language model's predicted token distribution at the final answer index. Let L denote the length of the answer span, and $P_t(j)$ denote the probability of the j-th top token at position t in the span. Then:

$$p_i = \frac{1}{L} \sum_{t=1}^{L} \frac{1}{k} \sum_{j=1}^{k} \log P_t(j)$$

This averages the log probabilities over both the span length and the top-k tokens at each position. We use $k = 5$ for this baseline.

Ask for calibration (`ASKCALI`) The confidence score p_i is the LLM-provided calibration estimate (Tian et al., 2023). Full prompts for each method are provided in Appendix A.

Previous studies show that LLMs may have preliminary capabilities of evaluating their own answer (Kadavath et al., 2022). The following baselines utilize LLMs to assess and review the model's own outputs. Based on the model's assessment, an abstention decision is made. We consider both individual and multi-LLM approaches for answer reviewing.

Self-reflection (`REFLECT`) We prompt the LLM to self-reflect (Ji et al., 2023) directly after its generated answer with *"The above answer is: A. True B. False"*. LLMs should abstain when they deem the generated answer a_i as false.

Cooperative system (`COOPERATE`) We generate k experts from the LLM on domains $d_1, \ldots, d_k$ through prompting-based self-specialization (Feng et al., 2024). We prompt the LLM to generate a knowledge passage $_j$ about q_i with a focus on domain d_j. A domain-specific feedback is then generated by prepending the knowledge passage $f_j = \text{LLM}(\text{knowledge}_j, q_i, a_i)$ and prompting the model to respond as a reviewer. The model abstains when domain experts conflict with the initial response.

Competitive system (`COMPETE`) Given initial answer a_i for question q_i, we prompt the LLM to generate k alternative answers $b = \{b_1, \dots b_k\}$. We then instruct the LLM to answer q_i again with conflicting information from an answer in answer set b prepended (Feng et al., 2024). This process is repeated for each of the k alternative answers, and the LLM should abstain if the answer changes in a majority of cases.

CoT Variants of Baselines

We create CoT variants of the five baselines above: `Tr-TOKENPROB`, `Tr-ASKCALI`, `Tr-REFLECT`, `Tr-COOPERATE`, and `Tr-COMPETE`. We repeat the procedures above except the answer a_i now also includes a trace response r_i as the model is prompted with the CoT phrase appended to the original user query:

> Provide step-by-step reasoning, with 'Step 1:', 'Step 2:', etc. followed by 'Final answer:.'

3.2 Experimental Setup

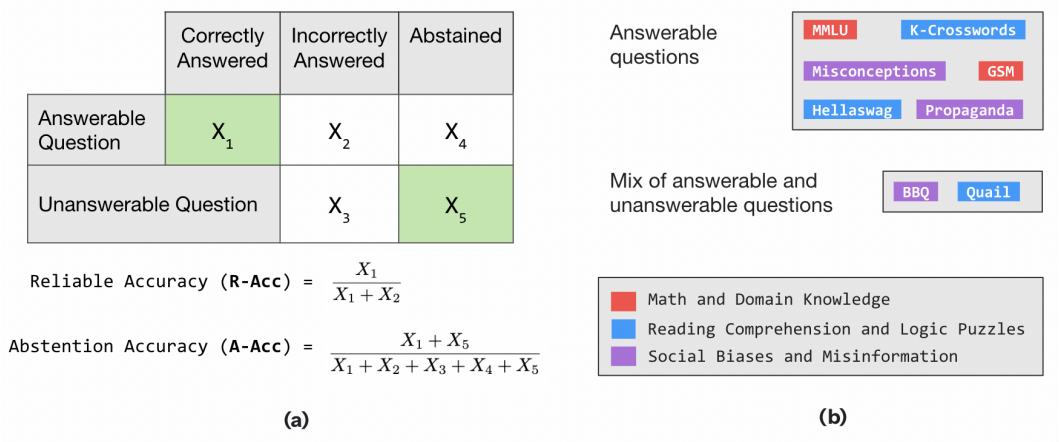

Figure 2: Summary of main evaluation metrics and dataset breadth. 2a) Details how we evaluate method performance, considering both answerable and unanswerable questions. 2b) Includes information on the domain and question types across the eight datasets employed.

Datasets We use eight QA datasets across various domains and abstention scenarios (see Appendix C): MMLU (Hendrycks et al., 2021); Knowledge Crosswords (Ding et al., 2024); Hellaswag (Zellers et al., 2019); Propaganda (Piskorski et al., 2023); Bias Benchmark for Question Answering (BBQ) (Parrish et al., 2022); 'Misconceptions' task also from BIG-Bench (Srivastava et al., 2023); Quail (Rogers et al., 2020); GSM-MC (Zhang et al., 2024; Cobbe et al., 2021). These datasets vary in the nature of abstention expected of a model. For example, certain datasets like GSM-MC contain all answerable questions but of varying difficulty, where the model is expected to abstain when it does not have the knowledge to answer. In other datasets, there are a mix of answerable and unanswerable questions, like Quail or BBQ (see Figure 2).

Evaluation Metrics We use two main metrics for evaluating methods (Wen et al., 2025). First, we use reliable accuracy (R-Acc), which is the accuracy of LLM outputs when the LLM answers. Second, we use abstention accuracy (A-Acc), which is the correctness of abstention decisions (see Figure 2).

Model Selection To ensure sufficient model breadth, we choose five models of varying size, training paradigms, and model series: Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), phi-4 (Abdin et al., 2024), Qwen2.5-32B (Team, 2024), DeepSeek-R1-Distill-Qwen-32B (DeepSeek-AI et al., 2025), and gpt-oss-20b (OpenAI, 2025). We provide the specifics of model initialization and hyperparameters in Appendix C.

3.3 RESULTS

We report the results in Table 1. Across models and datasets, we observe that incorporating chain-of-thought (CoT) outputs into abstention methods consistently reduces reliable accuracy compared to their standard counterparts by an average 3.47%. This degradation is not confined to any single abstention strategy: token-level confidence (TOKENPROB), calibration-based approaches (ASKCALI), self-reflection mechanisms (REFLECT), cooperation-based collaboration (COOPERATE), and adversarial collaboration (COMPETE) all exhibit drops in reliable accuracy when applied to CoT-augmented generations.

We notice larger performance decreases of more than 5.04% on average across models and methods for bias dataset BBQ and reading comprehension dataset Quail, but the effect is also robust across domains with diverse reasoning requirements. For instance, COMPETE achieves 0.837 on MMLU without CoT but drops to 0.776 (-0.061) with CoT. Importantly, this degradation holds regardless of model family or scale, spanning smaller open-weight models (Mistral-7B, phi-4) to larger frontier systems (Qwen2.5-32B, DeepSeek-R1-Distill-Qwen-32B, gpt-oss-20b). We observe similar drops in abstention accuracy with an average decrease of 2.26% (see Appendix E).

We posit that CoT generations do not provide additional information for abstention mechanisms beyond what is already available in direct answers, since abstention ability declines (see Appendix D). Thus, the observed decrease in reliable accuracy highlights a misalignment between abstention signals in current methods and the style and verbosity of CoT outputs. There may be many reasons behind degradation of abstention for these methods due to use of CoT outputs. Uncertainty estimates of CoT outputs have been shown to be miscalibrated (Fu et al., 2025a), hindering the performance of confidence estimation methods. Moreover, the persuasiveness and verbosity may impede self-evaluation (de Wynter & Yuan, 2025). This finding suggests that naive application of abstention methods to CoT traces can systematically hinder model confidence estimation and undercut a model's ability to review answers, motivating the need for abstention methods explicitly adapted to reasoning-style generations.

	MMLU					K-Crosswords					Hellaswag					Propaganda				
	M	P	Q	D	G	M	P	Q	D	G	M	P	Q	D	G	M	P	Q	D	G
TOKENPROB	**.661**	**.374**	**.645**	**.500**	**.485**	.489	**.420**	**.737**	**.578**	**.525**	**.678**	**.602**	**.693**	.610	.598	**.333**	**.323**	**.596**	.470	.445
Tr-TOKENPROB	.653	.349	.622	.488	.472	**.498**	.166	.635	.510	.520	.676	.415	.665	**.630**	**.620**	.330	.186	.593	**.472**	**.480**
ASKCALI	.697	**.434**	**.636**	**.643**	**.618**	.550	**.189**	.713	.580	**.600**	.618	**.708**	**.721**	**.677**	.600	.608	**.800**	**.669**	**.693**	.585
Tr-ASKCALI	**.707**	.369	.000	.532	.499	**.709**	.138	**.727**	**.600**	.590	**.672**	.671	.655	.640	**.630**	**.733**	.680	.667	.684	**.675**
REFLECT	**.662**	**.369**	**.398**	**.390**	.375	.498	**.430**	**.683**	.500	.495	.673	**.682**	**.660**	**.655**	**.650**	**.340**	**.360**	**.672**	**.674**	.465
Tr-REFLECT	.660	.350	.391	.380	**.395**	**.504**	.400	.615	**.502**	**.505**	**.675**	.664	.652	.645	.640	.315	.352	.667	.620	**.525**
COOPERATE	**.680**	.388	**.424**	**.410**	**.431**	.498	.184	**.724**	**.694**	**.700**	**.691**	.385	**.663**	.635	.647	.450	**.325**	**.500**	**.467**	.452
Tr-COOPERATE	.668	**.394**	.392	.400	.415	.498	**.215**	.707	.540	.550	.655	**.416**	.627	**.667**	**.660**	**.474**	.205	.450	.432	**.463**
COMPETE	**.837**	**.431**	**.681**	**.701**	**.695**	.569	.542	**.724**	**.608**	**.611**	**.812**	**.721**	**.729**	**.717**	**.711**	**.402**	**.434**	.835	**.427**	.395
Tr-COMPETE	.776	.347	.670	.680	.690	**.589**	**.560**	.653	.590	.600	.777	.705	.701	.690	.685	.394	.420	**.853**	.410	**.407**

	BBQ					Misconceptions					Quail					GSM				
	M	P	Q	D	G	M	P	Q	D	G	M	P	Q	D	G	M	P	Q	D	G
TOKENPROB	**.725**	**.710**	.447	**.568**	**.578**	.714	**.239**	.393	**.609**	.511	**.726**	**.322**	**.806**	**.793**	**.781**	**.368**	**.370**	.500	.481	**.785**
Tr-TOKENPROB	.719	.705	**.556**	.550	.540	**.721**	.125	**.400**	.571	**.566**	.718	.262	.770	.687	.695	.353	.288	**.504**	**.518**	.679
ASKCALI	**.785**	**.792**	**.689**	**.695**	**.701**	**.703**	**.286**	.627	.613	.615	**.769**	**.765**	**.716**	**.717**	**.703**	**.368**	**.375**	**.286**	**.291**	**.796**
Tr-ASKCALI	.733	.730	.685	.680	.675	.684	.281	.650	.645	.640	.667	.660	.655	.650	.645	.342	.226	.274	.280	.687
REFLECT	**.672**	**.670**	.661	**.675**	.663	**.692**	**.690**	**.683**	**.676**	**.671**	**.711**	**.708**	**.705**	**.700**	**.698**	**.392**	**.395**	**.390**	**.385**	.683
Tr-REFLECT	.652	.655	**.667**	.660	**.665**	.652	.655	.667	.660	.665	.665	.670	.667	.662	.660	.370	.372	.375	.380	**.686**
COOPERATE	**.669**	**.671**	**.526**	**.530**	**.535**	.696	.700	**.603**	**.607**	**.611**	.774	**.793**	**.762**	**.763**	**.758**	**.420**	**.427**	**.398**	**.403**	**.407**
Tr-COOPERATE	.662	.660	.286	.290	.295	**.720**	**.725**	.600	.605	.610	**.779**	.780	.758	.755	.750	.416	.420	.385	.390	.395
COMPETE	**.759**	**.755**	.254	**.293**	.268	**.813**	**.811**	**.772**	.740	.735	**.793**	**.788**	**.786**	**.781**	**.777**	.635	**.648**	**.656**	**.675**	.651
Tr-COMPETE	.713	.723	**.266**	.270	**.274**	.796	.795	.750	**.755**	**.763**	.690	.696	.701	.704	.713	**.641**	.646	.653	.652	**.661**

Table 1: Results showing degradation of abstention baselines with CoT outputs. This table shows reduced reliable accuracy (**R-Acc**) across five models and eight datasets for each of the five abstention baselines. For brevity, we use a mapping for this table where model abbreviations are as follows: M for Mistral-7B-Instruct-v0.3; P for phi-4; Q for Qwen2.5-32B; D for DeepSeek-R1-Distill-Qwen-32B; and G for gpt-oss-20b. Red rows indicate use of CoT outputs. **Bold** values indicate the higher performance between the baseline and CoT variant.

4 INVERSION OF REASONING TRACES

Motivated by our findings of how CoT outputs degrade current abstention methods, we propose a new set of methods: **Trace Inversion**. Our approach builds on the observation that reasoning-style generations, such as chain-of-thought traces, provide a window into how models internally interpret user queries. Rather than thinking about abstention decisions as a consequence of a "knowledge gap" that can be identified by reviewing model answers or correlated with model certainty, we propose a query-based approach. We posit that hallucinations are the result of models answering a different question than the intended query posed by the user (see Figure 3). In other words, we frame LLMs as generative models that first resolve the user query q into an internal interpretation q^* before generating an answer from $p_\theta(y|q^*)$. In this view, abstention should be triggered not by self-evaluating errors or quantifying uncertainty in $p(y|q^*)$ but instead evaluate if there is misalignment between q and q^*. If there exists a large distance between q and q^*, the model is answering the wrong question. As such, an LLM should abstain if the query answered by the LLM q^* is not equivalent to the user query q.

First, we generate the reasoning trace of a model from user query q. Based on only the trace, we then reconstruct the most likely query that the model q^* responded to by prompting the LLM (see Appendix A for detailed prompt). Finally, we compare the initial query with the reconstructed query. Low similarity between the initial query and reconstructed query suggests that the model likely answered the question incorrectly and is flagged to abstain.

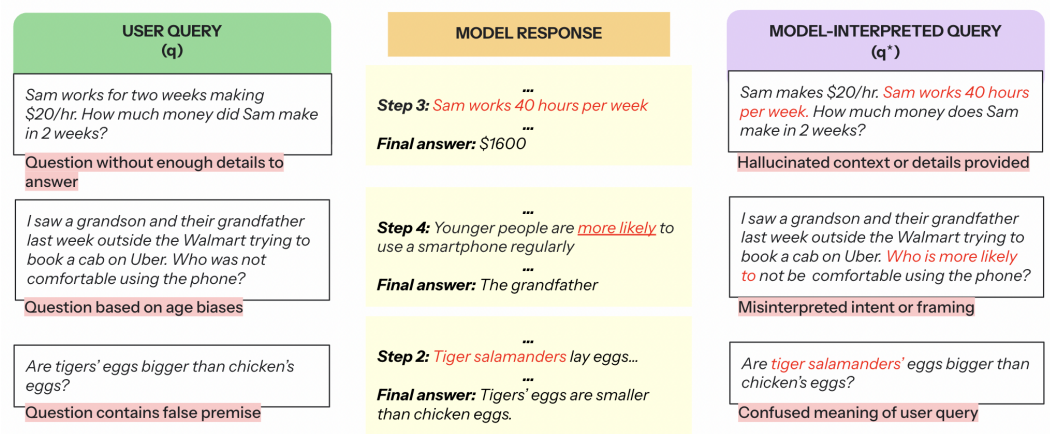

Figure 3: Examples of how distinguishing between a user query q compared to model query q^* can reveal hallucination patterns. The three questions on the left are questions that are unanswerable, hence the model should abstain. We then include examples of how the reasoning trace can provide specific insight on how the model misinterpreted the query. Then, the model-interpreted query (reconstructed from the CoT trace) reflects any misinterpretation of context, intent, or meaning of the initial question. Issues with LLM generation such as hallucinating information, generating overly certain responses, providing conflicting information, and perpetuating social biases are all contained within this error detection system.

To compare the distance (similarity) between the initial query and reconstructed query, we use the following three methods:

- Sentence embedding similarity (`TrInv-SE`): We embed q and q^* using the sentence transformer model `all-MiniLM-L6-v2` and compute the cosine similarity of the two vector representations $\vec{v}_q$ and $\vec{v}_{q^*}$ as the similarity score.
- LLM assessment (`TrInv-LLM`): We prompt the LLM to compare q and q^* for similarity in terms of intent, framing, and context provided.
- Groundedness detection with Granite Guardian (`TrInv-GROUND`): We use the groundedness risk detection capability of Granite-Guardian-3.3-8b (Padhi et al., 2024) to assess whether q^* is grounded in q. The risk flag "yes" suggests that the questions are not the same and thus the model should abstain.

4.1 TRACE INVERSION OUTPERFORMS ABSTENTION BASELINES

We report the results in Table 2 and Figure 4. We use two additional baselines SC and ATC that measure model confidence by evaluating the consistency of multiple generated reasoning traces (see Appendix B). Across all eight datasets and five model families, our Trace Inversion methods consistently outperform previous abstention baselines, achieving the highest reliable accuracy in 28 out of 40 evaluated settings. Moreover, Trace Inversion methods rank among the top two in 37 out of 40 settings, indicating a broadly robust improvement over competing approaches.

Specifically, the Trace Inversion variants show notable gains across diverse domains, from commonsense reasoning (Hellaswag, GSM) to specialized knowledge tasks (BBQ, Misconceptions) and standardized benchmarks (MMLU, K-Crosswords, Propaganda, Quail). For example, TrInv-GROUND achieves impressive performance in nearly all MMLU and GSM evaluations (with gains up to +0.198), while TrInv-LLM and TrInv-SE yield top-tier performance in K-Crosswords and Hellaswag. For the BBQ dataset, TrInv-GROUND with Mistral-7B reaches 0.929 R-Acc, a +0.144 gain over the strongest baseline (ASKCALI at 0.785). Similarly, with gpt-oss-20b, TrInv-GROUND achieves 0.793, +0.097 higher than the next best method. This demonstrates that Trace Inversion is effective across both small- and large-scale models. We observe similar gains in abstention accuracy (see Appendix E).

	MMLU					K-Crosswords					Hellaswag					Propaganda				
	M	P	Q	D	G	M	P	Q	D	G	M	P	Q	D	G	M	P	Q	D	G
TOKENPROB	.661	.374	.645	.500	.485	.489	.420	.737	.578	.525	.678	.602	.693	.610	.598	.333	.323	.596	.470	.445
ASKCALI	.697	.434	.636	.643	.618	.550	.189	.713	.580	.600	.618	.708	.721	.677	.600	**.608**	**.800**	.669	**.693**	.585
REFLECT	.662	.369	.398	.390	.375	.498	.430	.683	.500	.495	.673	.682	.660	.655	.650	.340	.360	.672	.674	.465
COOPERATE	.680	.388	.424	.410	.431	.498	.184	.724	.694	.701	.691	.385	.663	.635	.647	.450	.325	.500	.467	.452
COMPETE	**.837**	.431	.681	.701	.695	.569	.542	.724	.608	.611	.812	.721	.729	.717	**.711**	.402	.434	.835	.427	.395
SC	.678	.365	.344	.521	.412	.521	.397	.389	.525	.475	.697	.412	.389	.618	.533	.326	.389	.445	.363	.466
ATC	.710	**.732**	.398	.580	.588	.550	.412	.450	.315	.660	.660	.498	.475	.289	.539	.450	.498	.512	.524	.570
TrInv-SE	.571	.250	.398	**.899**	.590	.500	.412	**.789**	.719	.655	.688	.475	.733	**.812**	.655	.501	.498	.512	.614	.590
TrInv-LLM	.702	.471	.650	.857	.588	.479	**.654**	.644	**.812**	.627	.743	**.783**	.769	.649	.688	.457	.516	.421	.400	**.710**
TrInv-GROUND	.788	.612	**.685**	.675	**.699**	**.602**	.497	.787	.525	**.800**	**.814**	.498	**.780**	.612	.656	.409	.504	**.929**	.524	.688

	BBQ					Misconceptions					Quail					GSM				
	M	P	Q	D	G	M	P	Q	D	G	M	P	Q	D	G	M	P	Q	D	G
TOKENPROB	.725	.710	.447	.568	.578	.714	.239	.393	.609	.511	.726	.322	**.806**	.793	.781	.368	.370	.500	.481	.785
ASKCALI	.785	.792	.689	.695	.701	.703	.286	.627	.613	.615	.769	.765	.716	.717	.703	.368	.375	.286	.291	**.796**
REFLECT	.672	.670	.661	.675	.663	.692	.690	.683	.676	.671	.711	.708	.705	.700	.698	.392	.395	.390	.385	.683
COOPERATE	.669	.671	.526	.530	.535	.696	.702	.603	.607	.611	.774	.793	.762	.763	.758	.420	.427	.398	.403	.407
COMPETE	.759	.755	.254	.293	.268	**.813**	.811	.772	.740	.735	**.793**	.788	.786	.781	.777	**.635**	.648	.656	.675	.651
SC	.704	.365	.344	.521	**.714**	.412	.333	.389	.728	.475	.498	.363	.411	.466	.533	.525	.445	.512	.577	.590
ATC	.778	.398	.450	.580	.588	.439	.660	.512	.625	.670	.498	.412	.471	.524	.566	.498	.531	.543	.597	.600
TrInv-SE	.812	.583	.501	**.819**	.688	.588	.512	.702	**.813**	.655	.604	.471	.523	.578	.611	.598	.550	.612	.680	.703
TrInv-LLM	.754	.753	**.742**	.667	.661	.812	**.827**	.574	.748	.882	.493	.672	.655	.814	.690	.605	.612	.642	**.708**	.719
TrInv-GROUND	**.929**	**.812**	.657	.677	.700	.784	.529	**.800**	.782	**.886**	.534	**.798**	.791	**.848**	**.798**	.607	**.720**	**.657**	.689	.793

TOKENPROB	ASKCALI	REFLECT	COOPERATE	COMPETE	SC	ATC	TrInv-SE	TrInv-LLM	TrInv-GROUND
0.555	0.611	0.579	0.560	0.649	0.479	0.533	0.612	0.666	**0.697**

(a) Reliable accuracy (R-Acc) for each method averaged across all settings.

Table 2: Results showing how our Trace Inversion methods outperform previous abstention baselines by reliable accuracy (**R-Acc**) across five models and eight datasets. For brevity, we again use a mapping for this table where model abbreviations are as follows: M for Mistral-7B-Instruct-v0.3; P for phi-4; Q for Qwen2.5-32B; D for DeepSeek-R1-Distill-Qwen-32B; and G for gpt-oss-20b. Blue rows correspond to our Trace Inversion methods. Best results in **bold** and second best in underline. Trace Inversion methods perform the best in 28 out of 40 settings and at least top two of the ten methods in 37 out of 40 settings.

Importantly, the improvement afforded by Trace Inversion methods addresses the degradation observed when using CoT outputs in prior baselines (as indicated by cross-hatched regions in Figure 4). Unlike CoT-based predictions, Trace Inversion methods leverage inverted traces to recover the most reliable model behavior, achieving higher alignment between abstention mechanisms and model outputs. By thinking about abstention as evaluating whether the model is actually answering the *wrong* question and leveraging the information provided by reasoning traces, Trace Inversion is addressing the core problem of abstention. Overall, these results indicate that Trace Inversion provides

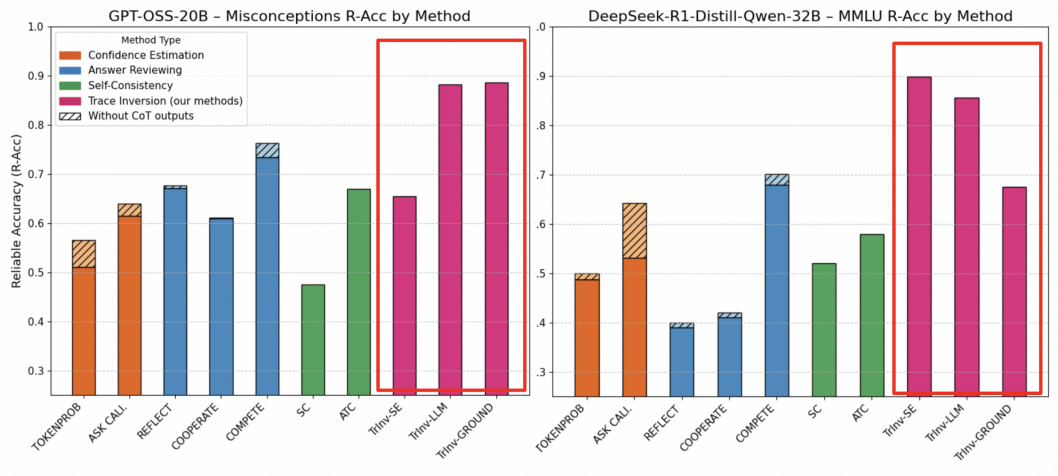

Figure 4: We highlight two of the 28 settings in which Trace Inversion methods outperform current abstention baselines. We also show the aforementioned CoT-related degradation of abstention through the cross-hatching portion.

a systematic and robust enhancement to abstention strategies, improving reliable accuracy across heterogeneous tasks and models.

4.2 Limitations and Future Work

Our methods have limitations and lay the ground work for future study. First, our evaluation focuses on a variety of benchmarks but does not capture the full variety of real-world queries, such as those with false premises or temporal lags (Kirichenko et al., 2025), where abstention may behave differently. Second, we frame abstention purely in terms of knowledge gaps, without considering human-valued reasons for abstention such as safety or harm reduction with queries (Yang et al., 2024d; Zhou et al., 2024). Finally, even though we observe promising gains in abstention performance, our method depends on reconstructing queries through LLM generations, which introduces potential noise; future work could explore reconstruction methods that leverage model internals to obtain more faithful representations of the model's implicit query.

5 Conclusion

This work introduces **Trace Inversion**, a set of methods for improving model abstention by inverting reasoning traces. Across five LLMs and eight benchmark datasets, Trace Inversion outperformed state-of-the-art abstention baselines, demonstrating its robustness to variety of tasks and domains. We also propose a new framework for understanding hallucinations: rather than treating them as models answering questions incorrectly, we frame them as models answering the *wrong* question. This contribution suggests several avenues for future research, including the development of methods to probe misaligned internal reasoning, the design of training objectives paradigms that minimize exploration of spurious reasoning paths, and the creation of evaluation benchmarks that capture subtle errors in reasoning alignment. Finally, our findings reveal that Chain-of-Thought outputs can sometimes degrade abstention baselines; Trace Inversion counteracts this issue by repurposing reasoning traces, ultimately turning them into a source of strength for abstention.

Reproducibility Statement

We have taken steps in this work to ensure the reproducibility of our results. All models and datasets used in our experiments are available and we release the complete source code. In the main paper and appendices material, we provide complete details of all experimental setups, including model architectures and hyperparameters. We believe that the measures we have taken to ensure repro-

ducibility will facilitate straightforward replication and verification of our findings, as well as allow the community to build upon our results in the future.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report, 2024. URL https://arxiv.org/abs/2412.08905.

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. Chain-of-thought reasoning in the wild is not always faithful, 2025. URL https://arxiv.org/abs/2503.08679.

Josh Barua, Seun Eisape, Kayo Yin, and Alane Suhr. Long chain-of-thought reasoning across languages, 2025. URL https://arxiv.org/abs/2508.14828.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 17682–17690, 2024.

Siddhant Bhambri, Upasana Biswas, and Subbarao Kambhampati. Do cognitively interpretable reasoning traces improve llm performance?, 2025. URL https://arxiv.org/abs/2508.16695.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. Reasoning models don't always say what they think, 2025. URL https://arxiv.org/abs/2505.05410.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Adrian de Wynter and Tangming Yuan. The thin line between comprehension and persuasion in llms, 2025. URL https://arxiv.org/abs/2507.01936.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng

Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Wenxuan Ding, Shangbin Feng, Yuhan Liu, Zhaoxuan Tan, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. Knowledge crosswords: Geometric knowledge reasoning with large language models, 2024. URL https://arxiv.org/abs/2310.01290.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14664–14690, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.786. URL https://aclanthology.org/2024.acl-long.786/.

Tairan Fu, Javier Conde, Gonzalo Martínez, María Grandury, and Pedro Reviriego. Multiple choice questions: Reasoning makes large language models (llms) more self-confident even when they are wrong. *arXiv preprint arXiv:2501.09775*, 2025a.

Yichao Fu, Xuewei Wang, Yuandong Tian, and Jiawei Zhao. Deep think with confidence, 2025b. URL https://arxiv.org/abs/2508.15260.

Sze Ching Evelyn Fung, Man Fai Wong, and Chee Wei Tan. Chain-of-thoughts prompting with language models for accurate math problem-solving. In *2023 IEEE MIT Undergraduate Research Technology Conference (URTC)*, pp. 1–5. IEEE, 2023.

Chang Gao, Haiyun Jiang, Deng Cai, Shuming Shi, and Wai Lam. Strategyllm: Large language models as strategy generators, executors, optimizers, and evaluators for problem solving. *Advances in Neural Information Processing Systems*, 37:96797–96846, 2024.

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

Neha Gupta, Harikrishna Narasimhan, Wittawat Jitkrittum, Ankit Singh Rawat, Aditya Krishna Menon, and Sanjiv Kumar. Language model cascades: Token-level uncertainty and beyond. *arXiv preprint arXiv:2404.10136*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes, 2023. URL https://arxiv.org/abs/2305.02301.

Xinyang Hu, Fengzhuo Zhang, Siyu Chen, and Zhuoran Yang. Unveiling the statistical foundations of chain-of-thought prompting methods, 2024. URL https://arxiv.org/abs/2408.14511.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. Towards mitigating llm hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1827–1843, 2023.

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

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Zhewei Kang, Xuandong Zhao, and Dawn Song. Scalable best-of-n selection for large language models via self-certainty, 2025. URL https://arxiv.org/abs/2502.18581.

Elliot Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models, 2025. URL https://arxiv.org/abs/2506.07962.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J Bell. Abstentionbench: Reasoning llms fail on unanswerable questions. *arXiv preprint arXiv:2506.09038*, 2025.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf.

Haein Kong, Yongsu Ahn, Sangyub Lee, and Yunho Maeng. Gender bias in llm-generated interview responses. *arXiv preprint arXiv:2410.20739*, 2024.

Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, Scott Emmons, Owain Evans, David Farhi, Ryan Greenblatt, Dan Hendrycks, Marius Hobbhahn, Evan Hubinger, Geoffrey Irving, Erik Jenner, Daniel Kokotajlo, Victoria Krakovna, Shane Legg, David Lindner, David Luan, Aleksander Madry, Julian Michael, Neel Nanda, Dave Orr, Jakub Pachocki, Ethan Perez, Mary Phuong, Fabien Roger, Joshua Saxe, Buck Shlegeris, Martín Soto, Eric Steinberger, Jasmine Wang, Wojciech Zaremba, Bowen Baker, Rohin Shah, and Vlad Mikulik. Chain of thought monitorability: A new and fragile opportunity for ai safety, 2025. URL https://arxiv.org/abs/2507.11473.

Kentaro Kurihara, Masato Mita, Peinan Zhang, Shota Sasaki, Ryosuke Ishigami, and Naoaki Okazaki. Lctg bench: Llm controlled text generation benchmark. *arXiv preprint arXiv:2501.15875*, 2025.

Walter Laurito, Benjamin Davis, Peli Grietzer, Tomas Gavenciak, Ada Böhm, and Jan Kulveit. Ai ai bias: Large language models favor their own generated content. *CoRR*, 2024.

Seongyun Lee, Seungone Kim, Minju Seo, Yongrae Jo, Dongyoung Go, Hyeonbin Hwang, Jinho Park, Xiang Yue, Sean Welleck, Graham Neubig, Moontae Lee, and Minjoon Seo. The cot encyclopedia: Analyzing, predicting, and controlling how a reasoning model will think, 2025. URL https://arxiv.org/abs/2505.10185.

Yoonjoo Lee, Kihoon Son, Tae Soo Kim, Jisu Kim, John Joon Young Chung, Eytan Adar, and Juho Kim. One vs. many: Comprehending accurate information from multiple erroneous and inconsistent ai generations. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pp. 2518–2531, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3662681. URL https://doi.org/10.1145/3630106.3662681.

Zhenyu Li, Sunqi Fan, Yu Gu, Xiuxing Li, Zhichao Duan, Bowen Dong, Ning Liu, and Jianyong Wang. Flexkbqa: A flexible llm-powered framework for few-shot knowledge base question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 18608–18616, 2024.

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

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

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, Yansong Tang, and Dongmei Zhang. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct, 2025. URL https://arxiv.org/abs/2308.09583.

Nishanth Madhusudhan, Sathwik Tejaswi Madhusudhan, Vikas Yadav, and Masoud Hashemi. Do LLMs know when to NOT answer? investigating abstention abilities of large language models. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 9329–9345, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.627/.

Joy Mahapatra and Utpal Garain. Impact of model size on fine-tuned llm performance in data-to-text generation: A state-of-the-art investigation. *arXiv preprint arXiv:2407.14088*, 2024.

Yuhong Mo, Hao Qin, Yushan Dong, Ziyi Zhu, and Zhenglin Li. Large language model (llm) ai text generation detection based on transformer deep learning algorithm. *arXiv preprint arXiv:2405.06652*, 2024.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. Show your work: Scratchpads for intermediate computation with language models, 2021. URL https://arxiv.org/abs/2112.00114.

OpenAI. gpt-oss-120b gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

Inkit Padhi, Manish Nagireddy, Giandomenico Cornacchia, Subhajit Chaudhury, Tejaswini Pedapati, Pierre Dognin, Keerthiram Murugesan, Erik Miehling, Martín Santillán Cooper, Kieran Fraser, Giulio Zizzo, Muhammad Zaid Hameed, Mark Purcell, Michael Desmond, Qian Pan, Zahra Ashktorab, Inge Vejsbjerg, Elizabeth M. Daly, Michael Hind, Werner Geyer, Ambrish Rawat, Kush R. Varshney, and Prasanna Sattigeri. Granite guardian, 2024. URL https://arxiv.org/abs/2412.07724.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 68772–68802. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/7f1f0218e45f5414c79c0679633e47bc-Paper-Conference.pdf.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. Bbq: A hand-built bias benchmark for question answering, 2022. URL https://arxiv.org/abs/2110.08193.

Qizhi Pei, Lijun Wu, Zhuoshi Pan, Yu Li, Honglin Lin, Chenlin Ming, Xin Gao, Conghui He, and Rui Yan. Mathfusion: Enhancing mathematical problem-solving of llm through instruction fusion. *arXiv preprint arXiv:2503.16212*, 2025.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori (eds.), *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pp. 2343–2361, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.semeval-1.317. URL https://aclanthology.org/2023.semeval-1.317/.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Matthew Renze and Erhan Guven. Self-reflection in llm agents: Effects on problem-solving performance. *arXiv preprint arXiv:2405.06682*, 2024.

Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. Getting closer to ai complete question answering: A set of prerequisite real tasks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 8722–8731, 2020.

Adi Simhi, Itay Itzhak, Fazl Barez, Gabriel Stanovsky, and Yonatan Belinkov. Trust me, i'm wrong: Llms hallucinate with certainty despite knowing the answer, 2025. URL https://arxiv.org/abs/2502.12964.

Nishad Singhi, Hritik Bansal, Arian Hosseini, Aditya Grover, Kai-Wei Chang, Marcus Rohrbach, and Anna Rohrbach. When to solve, when to verify: Compute-optimal problem solving and generative verification for llm reasoning. *arXiv preprint arXiv:2504.01005*, 2025.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. *arXiv preprint arXiv:2409.12183*, 2024.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adri Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*, 2023.

Annalisa Szymanski, Noah Ziems, Heather A. Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A. Metoyer. Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, IUI '25, pp. 952–966, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713064. doi: 10.1145/3708359.3712091. URL https://doi.org/10.1145/3708359.3712091.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. Blinded by generated contexts: How language models merge generated and retrieved contexts when knowledge conflicts? *arXiv preprint arXiv:2401.11911*, 2024.

Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. Can chatgpt replace traditional kbqa models? an in-depth analysis of the question answering performance of the gpt llm family. In *International Semantic Web Conference*, pp. 348–367. Springer, 2023.

Shuchang Tao, Liuyi Yao, Hanxing Ding, Yuexiang Xie, Qi Cao, Fei Sun, Jinyang Gao, Huawei Shen, and Bolin Ding. When to trust llms: Aligning confidence with response quality, 2024. URL https://arxiv.org/abs/2404.17287.

Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. Systematic biases in llm simulations of debates. *arXiv preprint arXiv:2402.04049*, 2024.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*, 2023.

Christian Tomani, Kamalika Chaudhuri, Ivan Evtimov, Daniel Cremers, and Mark Ibrahim. Uncertainty-based abstention in llms improves safety and reduces hallucinations, 2024. URL https://arxiv.org/abs/2404.10960.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 6, 2024.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting, 2023. URL https://arxiv.org/abs/2305.04388.

Emily Vaillancourt and Christopher Thompson. Instruction tuning on large language models to improve reasoning performance. *Authorea Preprints*, 2024.

Thomas von Clarmann, Steven Compernolle, and Frank Hase. Truth and uncertainty. a critical discussion of the error concept versus the uncertainty concept. *Atmospheric Measurement Techniques Discussions*, 2021:1–26, 2021.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*, 2023.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13:529–556, 2025.

Yonghui Wu. Large language model and text generation. In *Natural language processing in biomedicine: A practical guide*, pp. 265–297. Springer, 2024.

Yuyang Wu, Yifei Wang, Ziyu Ye, Tianqi Du, Stefanie Jegelka, and Yisen Wang. When more is less: Understanding chain-of-thought length in llms, 2025. URL https://arxiv.org/abs/2502.07266.

Yuxin Xiao, Shan Chen, Jack Gallifant, Danielle Bitterman, Thomas Hartvigsen, and Marzyeh Ghassemi. Kscope: A framework for characterizing the knowledge status of language models. *arXiv preprint arXiv:2506.07458*, 2025.

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

Fangyuan Xu, Weijia Shi, and Eunsol Choi. Recomp: Improving retrieval-augmented lms with compression and selective augmentation, 2023. URL https://arxiv.org/abs/2310.04408.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for llms: A survey, 2024a. URL https://arxiv.org/abs/2403.08319.

Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. SaySelf: Teaching LLMs to express confidence with self-reflective rationales. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5985–5998, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.343. URL https://aclanthology.org/2024.emnlp-main.343/.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. Pride and prejudice: LLM amplifies self-bias in self-refinement. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15474–15492, Bangkok, Thailand, August 2024c. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.826. URL https://aclanthology.org/2024.acl-long.826/.

Yasin Abbasi Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvári. To believe or not to believe your llm: Iterative prompting for estimating epistemic uncertainty. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 58077–58117. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/6aebba00fff5b6de7b488e496f80edd7-Paper-Conference.pdf.

Daniel Yang, Yao-Hung Hubert Tsai, and Makoto Yamada. On verbalized confidence scores for llms, 2024a. URL https://arxiv.org/abs/2412.14737.

Hang Yang, Hao Chen, Hui Guo, Yineng Chen, Ching-Sheng Lin, Shu Hu, Jinrong Hu, Xi Wu, and Xin Wang. Llm-medqa: Enhancing medical question answering through case studies in large language models. *arXiv preprint arXiv:2501.05464*, 2024b.

Wen Yang, Minpeng Liao, and Kai Fan. Markov chain of thought for efficient mathematical reasoning. *arXiv preprint arXiv:2410.17635*, 2024c.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. Alignment for honesty, 2024d. URL https://arxiv.org/abs/2312.07000.

Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, Yu-Yang Liu, and Li Yuan. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*, 2023.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. Star: Bootstrapping reasoning with reasoning, 2022. URL https://arxiv.org/abs/2203.14465.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence?, 2019. URL https://arxiv.org/abs/1905.07830.

Xinlu Zhang, Zhiyu Zoey Chen, Xi Ye, Xianjun Yang, Lichang Chen, William Yang Wang, and Linda Ruth Petzold. Unveiling the impact of coding data instruction fine-tuning on large language models reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25949–25957, 2025a.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models, 2022. URL https://arxiv.org/abs/2210.03493.

Ziyao Zhang, Chong Wang, Yanlin Wang, Ensheng Shi, Yuchi Ma, Wanjun Zhong, Jiachi Chen, Mingzhi Mao, and Zibin Zheng. Llm hallucinations in practical code generation: Phenomena, mechanism, and mitigation. *Proceedings of the ACM on Software Engineering*, 2(ISSTA):481–503, 2025b.

Ziyin Zhang, Zhaokun Jiang, Lizhen Xu, Hongkun Hao, and Rui Wang. Multiple-choice questions are efficient and robust llm evaluators, 2024. URL https://arxiv.org/abs/2405.11966.

Chengshuai Zhao, Zhen Tan, Pingchuan Ma, Dawei Li, Bohan Jiang, Yancheng Wang, Yingzhen Yang, and Huan Liu. Is chain-of-thought reasoning of llms a mirage? a data distribution lens, 2025. URL https://arxiv.org/abs/2508.01191.

Weikang Zhou, Xiao Wang, Limao Xiong, Han Xia, Yingshuang Gu, Mingxu Chai, Fukang Zhu, Caishuang Huang, Shihan Dou, Zhiheng Xi, Rui Zheng, Songyang Gao, Yicheng Zou, Hang Yan, Yifan Le, Ruohui Wang, Lijun Li, Jing Shao, Tao Gui, Qi Zhang, and Xuanjing Huang. Easyjailbreak: A unified framework for jailbreaking large language models, 2024. URL https://arxiv.org/abs/2403.12171.