

Deep Hidden Cognition Facilitates Reliable Chain-of-Thought Reasoning

Zijun Chen, Wenbo Hu*, Richang Hong

Hefei University of Technology
 chenzijun@mail.hfut.edu.cn, {wenbohu,hongrc}@hfut.edu.cn

Abstract

Chain of Thought (CoT) reasoning has demonstrated remarkable deep reasoning capabilities in both large language models (LLMs) and multimodal large language models (MLLMs). However, its reliability is often undermined by the accumulation of errors in intermediate steps. This paper proposes a novel approach to calibrating CoT reasoning accuracy by leveraging the model’s internal cognition of truthfulness. Our findings suggest that the model implicitly tracks the evolving veracity of intermediate steps throughout the dynamic, progressive reasoning process. We train a confidence predictor to quantify the model’s internal cognition of truthfulness at each reasoning step, enabling dynamic selection of the most plausible reasoning path through beam search. Experimental results demonstrate that our method significantly outperforms the state-of-the-art baselines (e.g., Self-Consistency, and PRM Guided Search) across the mathematical, symbolic, and commonsense reasoning tasks, exhibiting superior accuracy and reliability in both unimodal and multimodal settings. This study proposes a novel path toward improving the reliability of CoT reasoning, demonstrating strong potential for wide-ranging applications.

Code — <https://github.com/hfutml/cog-cot>

Introduction

Chain of Thought (CoT) reasoning has emerged as a powerful paradigm for unlocking deep reasoning capabilities in large language models (LLMs, (Wei et al. 2022; Lyu et al. 2023)) and multimodal large language models (MLLMs, (Zhang et al. 2023; Zheng et al. 2023)). CoT not only enhances the model’s performance but also ensures that the model has greater adaptability and controllability in complex, dynamic real-world tasks.

Yet, the CoT approach is not without its Achilles’ heel. Evidence suggests that CoT does not consistently outperform direct question-answering methods (Sprague et al. 2024), a limitation rooted in the progressive accumulation of errors. When an intermediate step contains an error, it can lead to the final result being incorrect. A key issue in mitigating this phenomenon is identifying potential errors generated by

*Corresponding Author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

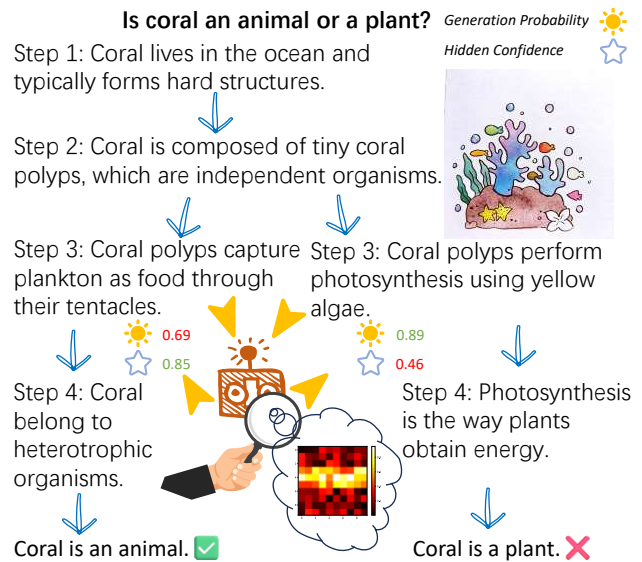


Figure 1: Demonstrate of the dissociation between surface-level generation probabilities and latent cognition in reasoning chains. The internal activations contain hidden representations of true information, and these sensitive activations are primarily concentrated in the intermediate layers.

the model during the generation process and then correcting them. Some researchers have explored this problem: Xie et al. (2023) utilized large models for self-evaluation, taking the probability of the model assigning items to the True category as an indicator of the correctness of intermediate steps. Yin et al. (2024) identified errors by detecting abrupt changes in the negative log probability values of tokens during the model’s generation process. Self-consistency (Wang et al. 2023) selects candidates with high voting rates or generation probabilities from multiple samplings as decoding results. Although such methods have achieved certain effects, they all rely solely on the surface probabilities of tokens generated by the model. However, the overconfidence of large models has been widely confirmed (Kadavath et al. 2022; Tian et al. 2023; Chen et al. 2025), and their generation probabilities do not match the actual truth, which inevitably raises concerns about the reliability of similar methods.

A profound paradox of human nature lies in the frequent disparity between what is spoken and what is thought. Intriguingly, this duality mirrors latent mechanisms in LLMs: even when generating erroneous outputs, their pretrained latent spaces retain structured factual knowledge, as evidenced by neural activation patterns that encode verifiable information (Burns et al. 2022; Moschella et al. 2022; Rai and Yao 2024; Skean et al. 2025).

Building on these studies, we consider whether the internal states of the model during CoT reasoning can generate exploitable authenticity signals. Such signals do not rely on surface-level generation probabilities and can assign appropriate probabilities even to incorrect outputs. These authenticity signals can then guide more reliable CoT decoding. Liu, Fatahi Bayat, and Wang (2024) used the hidden states of the model’s final layer to train a confidence predictor, achieving higher calibration than generation probabilities. However, the hidden states of the model’s final layer have been condensed into the probability prediction of the next token, and the authenticity signals encoded therein may not be the most dense. Studies have demonstrated that large models handle information differently across layers: lower layers focus on basic semantics, while higher layers process more abstract knowledge (Tenney, Das, and Pavlick 2019; Dai et al. 2021; Meng et al. 2022). Research by Li et al. (2023) shows that the encoding of such authenticity is dispersed and primarily concentrated in the middle layers. Therefore, extracting these dispersed authenticity signals may be more beneficial for quantifying authenticity probabilities compared to using only the final layer.

In this paper, we propose a novel method: using probing techniques to identify attention heads most sensitive to truthfulness, extracting their activations to construct a multi-layered feature representation, and training a high-precision confidence predictor. This predictor can accurately assess the correctness of each reasoning step. Furthermore, we integrate this predictor into a step-by-step decomposed CoT framework, combining it with beam search to dynamically select high-confidence reasoning steps. This strategy effectively suppresses error accumulation and enhances the overall reliability of CoT. This work advances the field through three key contributions:

1. we identified that a model’s internal activations during CoT reasoning encode truthfulness information, enabling us to train a confidence predictor with exceptional calibration performance;
2. we developed a binary-labeled CoT dataset, which supports training the predictor and fosters research into CoT reliability; this dataset will be made publicly available;
3. we introduced a novel method that leverages the model’s internal cognitive verification to enhance the reliability of CoT generation, with its effectiveness demonstrated across diverse unimodal and multimodal tasks and various models, including the large reasoning model.

Related Work

CoT Decoding Strategy

Chain-of-Thought (CoT) prompting (Wei et al. 2022) has revolutionized multi-step reasoning in LLMs by interleaving intermediate rationales into prompts. Early work demonstrated its efficacy across arithmetic, commonsense, and symbolic tasks, leveraging in-context learning to decompose problems into explicit reasoning chains. Subsequent advancements extended CoT to zero-shot settings (Kojima et al. 2022) using generic instructions like “Let’s think step by step,” revealing LLMs’ emergent reasoning capabilities. For structured tasks, few-shot CoT (Wei et al. 2022; Zhang et al. 2022) further improved performance through task-specific demonstrations, highlighting the importance of prompt complexity and relevance. In multimodal contexts, adapted CoT frameworks (Wang et al. 2024; Zhang et al. 2023) often involve extracting image information as intermediate steps. For example, scene graph generation (Mitra et al. 2024) and subtask decomposition (Gao et al. 2024) enhance MLLM reasoning by incorporating visual context into sequential rationales. To address error accumulation in long chains, UAG (Yin et al. 2024) identifies uncertain steps using token entropy and refines chains via certified clues. Techniques like ITI (Li et al. 2023) and DOLA (Chuang et al. 2023) further optimize reasoning by manipulating attention head activations or layer-specific logits. Xie et al. (2024) introduced “internal consistency” to calibrate LLM reasoning by measuring agreement between intermediate and final layer predictions.

Uncertainty Quantification

Uncertainty quantification, a cornerstone of machine learning, measures a model’s confidence in its predictions to ensure reliability. Typical approaches for deep learning include MC dropout (Srivastava et al. 2014), Laplace approximation (Ritter, Botev, and Barber 2018), and deep ensembles (Lakshminarayanan, Pritzel, and Blundell 2017). In the era of LLMs, novel methods have emerged to address uncertainty in sequential generation. For example, predictive entropy (Kadavath et al. 2022) quantifies uncertainty via logit distributions, while semantic entropy (Kuhn, Gal, and Farquhar 2023; Cui, Hu, and Zhu 2020) captures semantic variability across multiple model responses. Recent work like ACT-CAB (Liu, Fatahi Bayat, and Wang 2024) further refines this by training a linear layer on LM’s last-layer activations to get more precise confidence estimation. Researchers have leveraged these techniques to detect hallucinations in LLMs (Farquhar et al. 2024; Zhang et al. 2024), enhancing their trustworthiness. Notably, integrating uncertainty into CoT prompting has emerged as a promising direction. Methods like UAG (Yin et al. 2024) dynamically identify uncertain reasoning steps using token-level entropy. Similarly, Mo and Xin (2024) incorporates uncertainty-guided tree of thought to explore multiple reasoning paths.

Confidence from Truth-Relevant Attention Head Activations

The research by Li et al. (2023) made an important discovery: the internal activation patterns of LLMs implicitly encode the

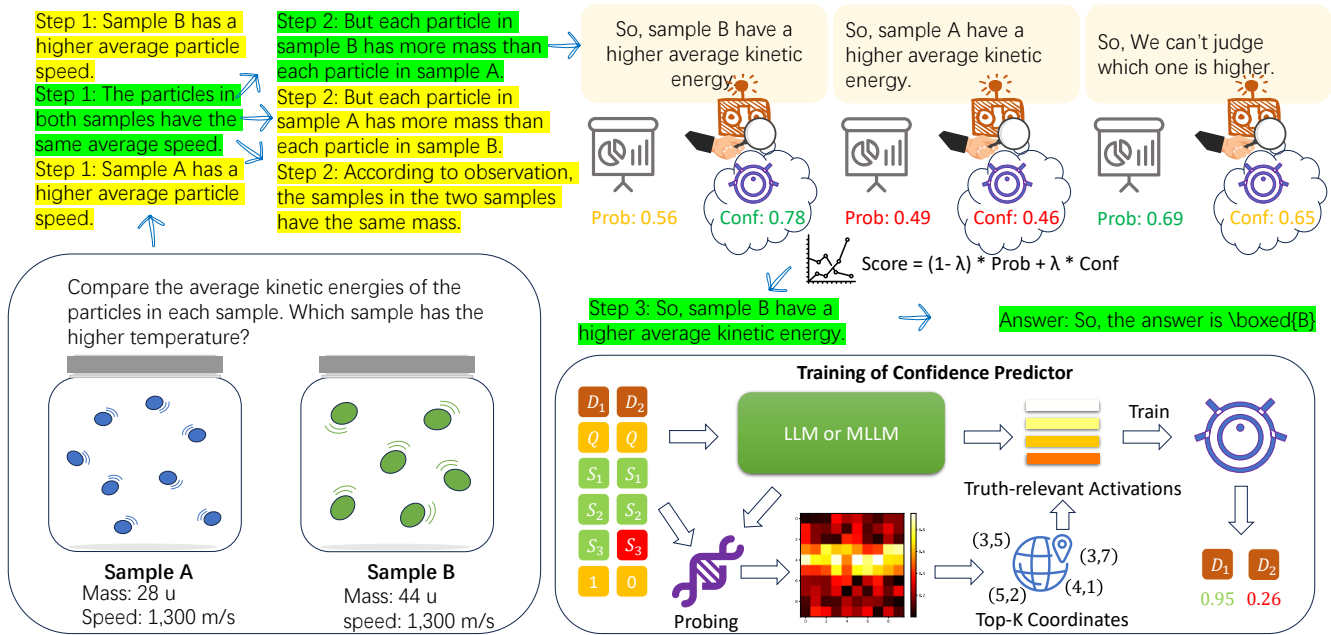


Figure 2: An overview of our method. Using the binary-labeled data to train a confidence predictor, and then introducing this predictor into the CoT reasoning process to select multiple generated candidates with the highest confidence.

veracity of answers. Specifically, they used a labeled dataset containing both correct and incorrect answers, concatenated “question-answer pairs” as model inputs, and trained binary classification probes to analyze the sensitivity of each attention head to truthfulness. Their results revealed that attention heads in the middle layers of the Transformer exhibit the highest sensitivity to veracity signals. This finding provided the first empirical evidence of an implicit veracity judgment mechanism at the neural level, laying a critical foundation for understanding how LLMs “reason” about truth internally.

However, this framework is limited to static “question + final answer” paradigms and fails to capture the dynamic, step-by-step nature of CoT reasoning. In CoT processes, each intermediate reasoning step not only relies on its own logic but also depends on the reasoning context established by preceding steps. This dynamic characteristic raises a more complex question: Given a question and the generated preceding reasoning steps $(S_{1,\dots,n-1})$, do the model’s internal activations contain implicit signals about the veracity of the “next reasoning step” (S_n) ?

To address this question, we designed a binary classification dataset specifically tailored for CoT scenarios. Unlike the static question-answer data used by Li et al. (2023), each sample in our dataset captures the reasoning context of CoT: formatted as $(Q, S_{1,\dots,n-1}, S_n^{\text{true or false}})$ where Q is the original question, $S_{1,\dots,n-1}$ represents the generated preceding steps, S_n is the next reasoning step to be evaluated (either true S_n^{true} or false S_n^{false}), Appendix A for detailed dataset construction.

Using this dataset, we trained a targeted binary classification probe to test a core hypothesis: Can the model’s internal activations effectively distinguish between truthful and false

branches in CoT reasoning? By measuring the classification accuracy of this probe on a validation set, we were able to quantify the sensitivity of activations in different layers and attention heads to the “truthfulness of the next step”—this approach extends the probe analysis methodology of Li et al. (2023) and for the first time explores implicit veracity signals in dynamic reasoning scenarios, providing an actionable framework for studying CoT reasoning. To provide a better explanation, we introduce the following concepts:

Multi-head attention architecture. To probe the internal activations of the model, we describe how to select the activation locations we need to probe within the Transformer architecture. A single transformer layer contains two key modules: a MHA mechanism and a standard multilayer perceptron (MLP) layer. During inference, the tokens are initially embedded into a high-dimensional space $x_0 \in \mathbb{R}^{D_{\text{high}}}$, which serves as the starting point for the residual stream. This vector initiates the sequence x_0, x_1, \dots, x_n of vectors in the stream. Each transformer layer processes the current vector x_i , performs computations, and adds the result to generate the next vector x_{i+1} . The final vector is then decoded into a prediction for the next-token distribution. In each layer, the MHA consists of H separate linear transformations. Specifically, the MHA can be written as:

$$x_{l+1} = x_l + \sum_{h=1}^H Q_h^l \text{Att}^l(P_h^l x_l), \quad (1)$$

where $P_h^l \in \mathbb{R}^{D_{\text{head}} \times D_{\text{high}}}$ maps the activations into a head space, and $Q_h^l \in \mathbb{R}^{D_{\text{high}} \times D_{\text{head}}}$ maps them back. The operator Att facilitates communication between input tokens. Following Li et al. (2023), the activations used for probe de-

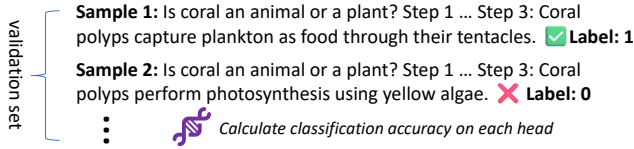


Figure 3: This figure shows the binary dataset of the constructed CoT.

tection and training the confidence predictor occur after Att and before Q_h^l , with activations denoted by $x_h^l \in \mathbb{R}^{D_{head}}$. Specifically, these activations take the form:

$$\mathbf{x}_h^l = Att(P_h^l \mathbf{x}_l). \quad (2)$$

Probing. A classification probe (Alain and Bengio 2018) is a lightweight classifier used to examine whether a model’s internal representations encode specific discrete features. Its core idea is to model the mapping between internal activations and target labels in a formalized way, thereby quantifying the model’s implicit ability to encode features.

Typically, classification is performed using a linear transformation followed by a softmax function: $f(h) = \text{softmax}(W^\top h + b)$. Given an input sample set $X = \{x_1, x_2, \dots, x_N\}$, we extract the activation a from a specific head of the model: $H = \{h_1, h_2, \dots, h_N\}$, where $h_i = \text{Model}(x_i; \theta)$ ^(a) Here, θ denotes the original model parameters, and a is the target activation. Using H as input and the label set: $Y = \{y_1, y_2, \dots, y_N\}$ as supervision, we train the probe parameters $\{W, b\}$. In our experiment, we selected the hidden state of the last token in the sequence X .

After training, the classification accuracy of the probe on an independent validation set is used to quantify the encoding ability of the activations: $\text{Acc} = \frac{1}{M} \sum_{j=1}^M \mathbf{1}(f(h_j^*) = y_j^*)$ where M is the number of validation samples, and (h_j^*, y_j^*) are the activation-label pairs from the validation set. If the probe achieves accuracy significantly higher than random chance, it suggests that the corresponding layer’s activations encode substantial information about the target feature.

Interpretability case studies. We conducted probing experiments on both unimodal and multimodal models. As illustrated in Figure 2, we used heatmaps to visualize the veracity classification accuracy of the trained probes on the validation set. It is evident that certain attention heads in the Transformer can distinguish between positive and negative CoT branch samples with remarkably high accuracy, reaching up to 85%. This indicates that both unimodal and multimodal models possess an internal latent cognition capable of discerning the veracity of dynamically progressive CoT reasoning. Consistent with the findings of Li et al. (2023), the activations sensitive to the veracity of CoT patterns are predominantly distributed in the middle layers. Notably, the distribution patterns of the multimodal LLaVA model and the unimodal LLaMA model exhibit a high degree of similarity. Given that LLaVA is trained on the LLaMA base model, this similarity suggests that such latent cognition may originate from pre-training rather than task-specific fine-tuning.

Training of Confidence Predictor

Merely identifying attention heads sensitive to truthfulness is insufficient to solve the problem. To transform these latent signals into operational tools, a rigorous quantification mechanism is required to convert distributed neural activations into interpretable confidence metrics. Specifically, we propose training a “truthfulness-aware confidence predictor”—an auxiliary module that learns to map the model’s internal states to a scalar value, which characterizes the prediction of the truthfulness of the next reasoning step. This module learns to associate the activation dynamics of sensitive heads with human-judged truthfulness.

Let x denote the input query, with y representing the model’s output sequence of length $|y|$. For the final token $y_{|y|}$, we retrieve its hidden state \mathbf{h}^l from the l -th layer of the language model, which consists of H attention heads. Specifically, according to Equation 2, this state is decomposed into a multi-head representation: $\mathbf{h}^l = [\mathbf{h}_1^l, \mathbf{h}_2^l, \dots, \mathbf{h}_H^l] \in \mathbb{R}^{H \times D_{hidden}}$, where $\mathbf{h}_h^l \in \mathbb{R}^{D_{hidden}}$ denotes the attention output for head h . Unlike previous methods that aggregate activations across the entire sequence, our method focuses on the activations corresponding to the final token. This choice is motivated by the idea that the final token’s representation encapsulates the model’s judgment of the overall correctness of the generated sequence (Li et al. 2023).

To identify which attention heads provide the most useful information for assessing truthfulness, we use a probing technique to analyze the correlation between each head’s activation and the ground-truth label. This allows us to rank the attention heads by their relevance to truthfulness detection. We then select the top- K heads ($K = H$, we choose to select the same number of attention heads as each layer) with the highest correlation, ensuring that we capture the most informative heads from the model’s intermediate layers: $\mathcal{S} = \{(l_1, h_1), (l_2, h_2), \dots, (l_K, h_K)\}$. Here, $l_i \in \{1, 2, \dots, L\}$ and $h_i \in \{1, 2, \dots, H\}$ denote the layer and head indices of the selected heads.

Next, the activations of the selected heads are concatenated into a unified representation vector:

$$\mathbf{v} = \text{Concat}(\mathbf{h}_{h_1}^{l_1}, \mathbf{h}_{h_2}^{l_2}, \dots, \mathbf{h}_{h_K}^{l_K}) \in \mathbb{R}^{K \cdot D_{hidden}}. \quad (3)$$

This concatenation captures the spatial dependencies encoded by each head while increasing the representational capacity of the feature vector.

To predict the confidence of the generated sequence y , we use a simple classifier: $p_\theta(y|\mathbf{x}) = \sigma(\mathbf{W} \cdot \mathbf{v} + b)$, where $\mathbf{W} \in \mathbb{R}^{1 \times (K \cdot D_{hidden})}$ and $b \in \mathbb{R}$ are learnable parameters. The sigmoid function $\sigma(\cdot)$ normalizes the output to a confidence score in $[0, 1]$. This classifier explicitly models the relationship between the attention head activations and the model’s confidence, enabling a more reliable estimation of the output’s correctness.

ECE Loss. To train the confidence predictor, the standard mean squared error (MSE) loss function is commonly applied in calibration tasks. However, it inherits a fundamental limitation from binary labels, This sharp 0/1 target often leads to overconfidence or underconfidence. (Guo et al. 2017).

To overcome this limitation, we adopt the ‘‘ECE loss’’ proposed by Liu, Fatahi Bayat, and Wang (2024) to improve calibration. They demonstrated its effectiveness through ablation experiments.

The ECE loss replaces binary correctness labels with soft targets derived from cross-validation accuracy distributions. Specifically, K -fold cross-validation is used to generate confidence scores for each fold, and the data is partitioned into B equal-interval bins. For each bin B_i , the empirical accuracy is computed as: $\text{acc}(B_i) = \frac{1}{|B_i|} \sum_{(x,y) \in B_i} \mathbf{1}(y \text{ is correct})$, where $\text{acc}(B_i)$ represents the proportion of correct predictions in the bin B_i . These accuracies provide a soft label that better reflects the true confidence-accuracy relationship. The final ECE loss is then computed as:

$$\mathcal{L}_{\text{ECE}} = \frac{1}{|\mathcal{D}|} \sum_{(x,y) \in \mathcal{D}} (\text{acc}(B_i) - p_\theta(y|x))^2. \quad (4)$$

This loss function directly aligns the predicted confidence with the empirical accuracy of the model’s predictions, leading to more accurate calibration of the confidence scores.

Confidence Predictor Evaluation

Given the limited persuasiveness of comparing the performance of the trained predictor with other baseline methods when evaluating it on intermediate steps of CoT reasoning, we ultimately opted to use a non-CoT dataset to assess its calibration for predicting the truthfulness of final answers. The construction of the dataset required for predictor training is detailed in Appendix A.

Model. We tested our approach using LLaMA2-7B-Chat, LLaMA2-13B-Chat, and LLaMA-3-8B-Instruct models.

Metrics. We employ three complementary metrics to evaluate calibration and prediction quality: Expected Calibration Error (ECE), Brier Score and Area Under the ROC Curve (AUC). See Appendix B for specific definitions.

A model may achieve high AUC yet exhibit poor calibration (high ECE/Brier Score), underscoring the necessity of evaluating all three metrics to fully characterize predictive reliability.

Baseline. We compared four methods for quantifying confidence to demonstrate the superiority of our approach.

- **Sequence Likelihood:** The average probability value assigned by the model to the answer tokens.
- **‘‘Is True’’ probability:** The probability assigned by the language model to ‘‘True’’ when prompted to verify the correctness of a question-answer pair (Tian et al. 2023).
- **Verbalization:** Method where the language model is prompted to verbally articulate its confidence of answers (Tian et al. 2023).
- **ACTCAB:** Method that exclusively uses the final layer of the model to train a confidence predictor (Liu, Fatahi Bayat, and Wang 2024).

Results. To evaluate the effectiveness of our method, we compared it against four confidence quantification approaches on the test sets of four non-CoT datasets: CommonsenseQA (Talmor et al. 2019), SciQ (Johannes Welbl

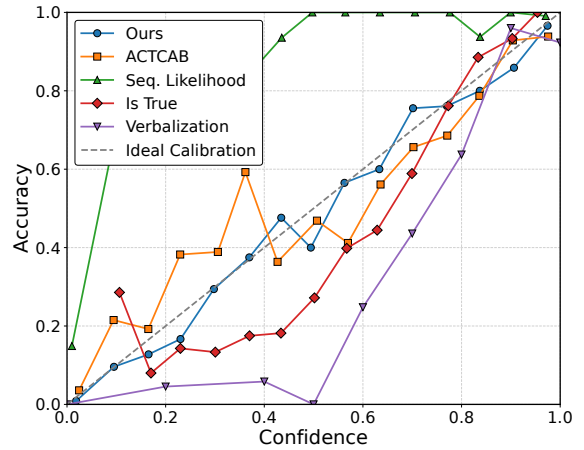


Figure 4: This figure shows the calibration curves of LLaMA2-13B-Chat in SciQ, where closer proximity to the Ideal Calibration curve indicates better calibration.

2017), WikiQA (Yang, Yih, and Meek 2015), and TruthfulQA (Lin, Hilton, and Evans 2021). Table 1 presents the results (The remaining results are in Appendix B). Notably, ACTCAB and our proposed method (both trained approaches) demonstrated substantial performance gaps over untrained baselines. Specifically, our approach, which incorporates Truth-Relevant activations identified via probing, outperformed ACTCAB’s exclusive reliance on the model’s final layer for confidence prediction. Additionally, the use of ECE loss optimized both ACTCAB and our method for improved calibration, reflected in their strong ECE scores. While Brier Score and AUC provided broader assessments beyond calibration-specific optimization, our method consistently outperformed competitors across datasets, models, and metrics. Figure 4 displays a visual comparison of calibration curves.

Dataset	Metric	Ours	ACTCAB	Seq. Likelihood	Is True	Verbalization
WikiQA	ECE↓	0.037	0.058	0.271	0.146	0.254
	Brier↓	0.102	0.149	0.302	0.231	0.291
	AUC↑	0.934	0.868	0.678	0.747	0.640
TruthfulQA	ECE↓	0.019	0.038	0.284	0.155	0.178
	Brier↓	0.122	0.140	0.344	0.286	0.296
	AUC↑	0.907	0.881	0.480	0.510	0.570
SciQ	ECE↓	0.018	0.052	0.223	0.132	0.194
	Brier↓	0.095	0.132	0.183	0.178	0.221
	AUC↑	0.937	0.897	0.949	0.848	0.784
CommonQA	ECE↓	0.024	0.029	0.167	0.482	0.224
	Brier↓	0.133	0.137	0.184	0.481	0.290
	AUC↑	0.891	0.885	0.873	0.422	0.604

Table 1: Evaluation of our confidence predictor against state-of-the-art baselines on LLaMA2-13B-Chat.

More Reliable Reasoning Chain Generation

A single reasoning path is often constrained by local optima or implicit biases, making it difficult to fully unleash the model’s internal knowledge reserves. Therefore, exploring multiple potential reasoning branches through multi-path

sampling has emerged as a key approach to excavate the model’s deep-seated knowledge and mitigate the risks of a single path. To this end, we integrate the well-trained confidence predictor into the CoT generation process: it quantitatively evaluates the next reasoning step of each candidate path, ultimately selecting the step with the highest confidence as the cornerstone for subsequent reasoning.

Confidence Predictor Guided Beam Search

We propose an improved sequence generation method that optimizes CoT reasoning through dynamic fusion of confidence predictions and generation probabilities. The core workflow can be summarized as follows: at each decoding step, beam search generates multiple candidate reasoning steps, then a pretrained confidence predictor evaluates their quality, and a weighted scoring function ultimately selects the optimal path. Specifically, given the current generated chain-of-thought sequence $S_t = \{s_1, s_2, \dots, s_t\}$, where each s_i represents a reasoning step, the model extends it by generating M candidate subsequent steps $\{C_{t+1}^1, C_{t+1}^2, \dots, C_{t+1}^M\}$ through beam search. Each candidate step C_{t+1}^m consists of a sequence of tokens $(x_1^m, x_2^m, \dots, x_{L_m}^m)$, generated with probability distribution $P(C_{t+1}^m | S_t)$. The pretrained confidence predictor $\mathcal{F}_{\text{conf}}$ computes a confidence score $\beta(C_{t+1}^m)$ for each candidate. These scores are combined with the normalized token-level generation probability to construct a global scoring function:

$$\bar{P}(C_{t+1}^m) = \exp\left(\frac{1}{L_m} \sum_{k=1}^{L_m} \log P(x_k^m | S_t, x_1^m, \dots, x_{k-1}^m)\right),$$

$$\text{Score}(C_{t+1}^m) = \lambda \cdot \beta(C_{t+1}^m) + (1 - \lambda) \cdot \bar{P}(C_{t+1}^m), \quad (5)$$

where $\lambda \in [0, 1]$ is a balancing weight (we performed grid search on several datasets and ultimately selected $\lambda = 0.5$.) and L_m denotes the number of tokens in the candidate step C_{t+1}^m . The algorithm appends the highest scoring candidate C_{t+1}^* to the current chain S_t , forming $S_{t+1} = S_t \oplus C_{t+1}^*$, and iterates until the final answer is generated. By guiding the search direction with the confidence predictor, our method mitigates the degradation of the reasoning path caused by the accumulation of errors. Detail settings are in Appendix C. Figure 2 shows the overall process of our method.

Experiments

We systematically evaluate the performance of our proposed method by leveraging state-of-the-art LLM and MLLM backbones across a diverse range of reasoning benchmarks.

Benchmarks. The evaluation framework comprises two primary categories: *Mathematical/Symbolic Reasoning* and *Commonsense Reasoning*, each spanning both unimodal and multimodal scenarios.

- **Mathematical & Symbolic Reasoning** The unimodal benchmarks include *GSM8K* (Cobbe et al. 2021), *Boolean Expressions* (Suzgun et al. 2023) and *SVAMP* (Patel, Bhattamishra, and Goyal 2021). Multimodal evaluation is represented by *CLEVR-Math* (Lindström and Abraham 2022).

- **Commonsense Reasoning** Unimodal evaluation includes *StrategyQA* (Geva et al. 2021) and *BoolQ* (Clark et al. 2019). Multimodal benchmarks include *ScienceQA* (Lu et al. 2022), *RealWorldQA* (Rea 2023), and *MMStar* (Chen et al. 2024).

Models. To comprehensively evaluate our method, we conducted experiments using state-of-the-art LLMs and MLLMs spanning diverse parameter scales and architectures. For unimodal reasoning, we employed LLaMA-2 in 7B, 13B, and 70B (Touvron et al. 2023) configurations to examine scalability effects. In the multi-modal domain, we utilized LLaVA v1.6 (7B/13B, (Liu et al. 2023)) and Qwen2.5-VL-Instruct (7B, (Team 2025)) to assess cross-modal capabilities.

Baseline. To conduct a comprehensive comparison of our method, we selected three representative baselines: (1) Few-Shot CoT (Wei et al. 2022), this approach provides a small number of reasoning examples to prompt the model for step-by-step reasoning; (2) Self-Consistency (Wang et al. 2023), it generates multiple reasoning paths through sampling and selects the optimal chain via majority voting; (3) Self-Evaluation Guided Beam Search (Xie et al. 2023), for fair comparison, it shares identical hyperparameters and prompt templates with our proposed approach, except replacing our scoring function (Equation 5) with their metric. (4) Process Reward Models (Zhao et al. 2025) aim to score each step of the reasoning process and guide model learning through fine-grained supervision. PRMs replace the models with a classification head to provide reward scores.

Main Results

Table 2 present the performance of our proposed method in unimodal and multimodal tasks respectively. The results clearly show that our method achieves excellent performance under each setting. When compared with the few - shot CoT under the same settings, our method demonstrates significant improvements in most tests. For instance, in the unimodal task, for the SVAMP dataset, our method has a 5% improvement over the few shot CoT (48.3 vs. 43.3). In the multimodal task, for the realworldqa dataset, our method shows a 10.7% improvement. Overall, our method outperforms the baseline few - shot CoT in most cases, whether it is in Mathematical & Symbolic Reasoning, Commonsense Reasoning tasks, or unimodal and multimodal tasks. This fully demonstrates that the confidence extracted from the internal states of the model can effectively guide the generation of more reliable reasoning chains. Some examples shown in Appendix C.

The Impact of Confidence Selection

It is important to rely on confidence predictors for selection among candidates. To verify the effect of the confidence predictor we introduced on the results, we conducted an ablation study. Instead of selecting candidates with higher confidence based on Equation 5, we randomly selected from these candidates. With all other parameter settings unchanged, we tested on multiple datasets and compared few-shot CoT, our method, and the random selection strategy, as shown in Figure 5. The results indicate that (the remaining results are in Appendix D), compared to when the confidence predictor is used, the

Model	Method	Datasets				AVG	
		GSM8K	SVAMP	StrategyQA	BoolQ		Boolean
LLaMA2-7B	CoT-few	24.4	43.3	63.0	54.8	62.4	49.6
	SC	24.3	47.7	63.0	58.0	70.4	52.7
	SE	24.9	40.0	63.0	58.2	72.0	51.6
	PRM	24.4	44.0	63.8	56.6	67.2	51.2
	Ours	25.2	48.3	64.5	58.6	70.0	53.3
LLaMA2-13B	CoT-few	39.9	53.7	66.0	57.6	68.8	57.2
	SC	39.3	54.0	65.9	56.6	70.0	57.2
	SE	38.4	51.7	66.7	56.4	69.6	56.6
	PRM	39.2	55.0	65.8	53.8	68	56.4
	Ours	42.8	55.7	66.8	59.2	68.8	58.7
LLaMA2-70B	CoT-few	52.3	70.3	72.0	67.0	79.6	68.2
	SC	59.0	73.0	74.2	67.4	79.2	70.6
	SE	52.6	66.3	74.3	68.6	79.6	68.3
	PRM	54.6	70.3	74.4	68.2	81	69.9
	Ours	58.3	74.7	73.7	69.8	81.2	71.5

Table 2: Comparison of Chain-of-Thought Decoding Processes in Unimodal (left) and Multimodal (right) tasks. SC: Self-Consistency, SE: Self-Evaluation Guided Beam Search.

Model	Method	Datasets				AVG
		ScienceQA	realworldqa	clevr-math	MMStar	
LLaVA-7B	CoT-few	58.3	23.0	18.0	40.3	34.9
	SC	55.7	18.8	15.7	41.3	32.9
	SE	61.6	28.2	18.0	44.0	38.0
	PRM	65.0	33.1	17.0	44.0	39.7
	Ours	62.4	33.7	18.0	45.7	40.0
LLaVA-13B	CoT-few	61.9	31.9	10.7	41.0	36.4
	SC	64.0	25.6	13.0	39.3	35.5
	SE	65.4	30.3	11.7	41.7	37.3
	PRM	61.3	33.9	9.7	41.3	36.6
	Ours	69.2	31.9	14.0	42.0	39.3
Qwen2.5-VL-7B	CoT-few	86.0	52.2	64.7	58.0	65.2
	SC	84.0	52.7	67.7	57.3	65.4
	SE	76.3	53.7	68.5	57.7	64.1
	PRM	81.0	53.0	70.0	58.1	65.5
	Ours	82.0	51.2	71.0	58.3	65.6

Model	Method	Datasets				AVG	
		GSM8K	SVAMP	StrategyQA	BoolQ		Boolean
DS-R1-7B	CoT-few	85.3	87.3	56.9	56.6	90.4	75.3
	SC	84.2	87.5	60.2	56.6	95.2	76.7
	SE	82.3	85.7	59.8	59.8	96	76.7
	PRM	86.2	88.2	60.7	57.4	95.2	77.5
	Ours	84.0	88.3	59.7	59.4	96.8	77.6

Table 3: This table demonstrates the performance of our approach on the DeepSeek R1 distilled reasoning model.

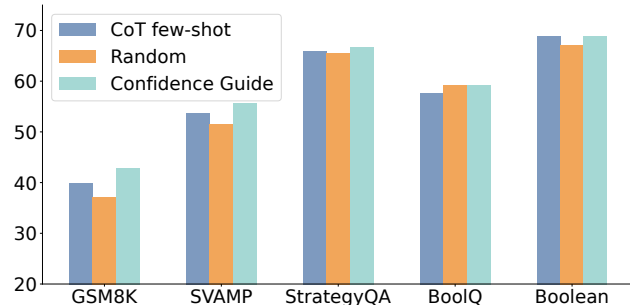


Figure 5: This figure demonstrates the effect (LLaMA2-13B) of confidence-guided reasoning.

random strategy shows a significant drop in performance, and in many cases, its performance is even lower than the baseline few-shot CoT. This ablation study strongly demonstrates the importance of the confidence predictor.

Effectiveness of Self-Correcting Error Steps

Overall effective, indicating the compatibility of our method. In selecting the highest confidence candidate, low confidence levels across all candidates can lead to error accumulation in the CoT process. An intuitive solution is to set a threshold: if the confidence falls below it, the model self-corrects the potentially erroneous step. Some research suggests that unbiased prompts—those not explicitly instructing the model to correct errors—can encourage limited self-correction (Liu

Method	GSM8K	SVAMP	StrategyQA	BoolQ	Boolean	AVG
Ours	25.2	48.3	64.5	58.6	70.0	53.3
Revise	26.0	49.0	64.5	59.6	69.6	53.7

Table 4: The effect of model (LLaMA2-7B) self-correction on erroneous steps in CoT reasoning.

et al. 2024), while overly directive prompts may lead to blind adherence to instructions rather than reasoning. To test this, we applied the unbiased prompting method from Liu et al. (2024) (see Appendix D) and set a threshold of 0.5. The results show an overall improvement (see Table 4). These findings indicate that self-correction during CoT generation is effective, and our method is well compatible with some CoT error correction techniques.

Performance in LRM

Still effective, indicating the generalization capability of our method. We aim to investigate the applicability of our method to reasoning models like DeepSeek R1, which have undergone specialized CoT training, including reinforcement learning on long-chain reasoning datasets. We tested language models distilled from DeepSeek R1, such as DS-Distill-Qwen-7/14B.

Our results, shown in Table 3 (full details in Appendix D), the overall performance still favors our method.

Conclusion

This paper presents a method leveraging the internal truthfulness sensitivity of LLMs and MLLMs to significantly enhance the reliability of CoT reasoning. By analyzing attention head activations, we train a confidence predictor to evaluate the correctness of each reasoning step and design a confidence-driven reasoning generation strategy that effectively reduces error accumulation. Experimental results demonstrate superior performance across diverse tasks and model scales, outperforming existing approaches. Further analysis reveals the model’s self-correction mechanism in CoT reasoning.

Acknowledgments

This paper is supported by the National Science Foundation of China Project (No. 62306098), Fundamental Research Funds for the Central Universities (No. JZ2024HGTTB0256) and the Open Project of Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University (No. MMC202412). The computation is completed on the HPC Platform of Hefei University of Technology.

References

2023. RealWorldQA. <https://huggingface.co/datasets/lmms-lab/RealWorldQA>.
- Alain, G.; and Bengio, Y. 2018. Understanding intermediate layers using linear classifier probes, 2017. In *URL* <https://openreview.net/forum>.
- Burns, C.; Ye, H.; Klein, D.; and Steinhardt, J. 2022. Discovering latent knowledge in language models without supervision. *International Conference on Learning Representations*.
- Chen, L.; Li, J.; Dong, X.; Zhang, P.; Zang, Y.; Chen, Z.; Duan, H.; Wang, J.; Qiao, Y.; Lin, D.; et al. 2024. Are We on the Right Way for Evaluating Large Vision-Language Models? *NeurIPS*.
- Chen, Z.; Hu, W.; He, G.; Deng, Z.; ZHANG, Z.; and Hong, R. 2025. Unveiling Uncertainty: A Deep Dive into Calibration and Performance of Multimodal Large Language Models. In Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B. D.; and Schockaert, S., eds., *Proceedings of the 31st International Conference on Computational Linguistics*, 3095–3109. Abu Dhabi, UAE: Association for Computational Linguistics.
- Chuang, Y.-S.; Xie, Y.; Luo, H.; Kim, Y.; Glass, J.; and He, P. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *International Conference on Learning Representations*.
- Clark, C.; Lee, K.; Chang, M.-W.; Kwiatkowski, T.; Collins, M.; and Toutanova, K. 2019. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. In *NAACL*.
- Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; Hesse, C.; and Schulman, J. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*.
- Cui, P.; Hu, W.; and Zhu, J. 2020. Calibrated reliable regression using maximum mean discrepancy. *Advances in Neural Information Processing Systems*, 33: 17164–17175.
- Dai, D.; Dong, L.; Hao, Y.; Sui, Z.; Chang, B.; and Wei, F. 2021. Knowledge neurons in pretrained transformers. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.
- Farquhar, S.; Kossen, J.; Kuhn, L.; and Gal, Y. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017): 625–630.
- Gao, T.; Chen, P.; Zhang, M.; Fu, C.; Shen, Y.; Zhang, Y.; Zhang, S.; Zheng, X.; Sun, X.; Cao, L.; et al. 2024. Cantor: Inspiring Multimodal Chain-of-Thought of MLLM. *Proceedings of the 32nd ACM International Conference on Multimedia*.
- Geva, M.; Khashabi, D.; Segal, E.; Khot, T.; Roth, D.; and Berant, J. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9: 346–361.
- Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International conference on machine learning*, 1321–1330. PMLR.
- Johannes Welbl, M. G., Nelson F. Liu. 2017. Crowdsourcing Multiple Choice Science Questions.
- Kadavath, S.; Conerly, T.; Askell, A.; Henighan, T.; Drain, D.; Perez, E.; Schiefer, N.; Hatfield-Dodds, Z.; DasSarma, N.; Tran-Johnson, E.; et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.
- Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213.
- Kuhn, L.; Gal, Y.; and Farquhar, S. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *International Conference on Learning Representations*.
- Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.
- Li, K.; Patel, O.; Viégas, F.; Pfister, H.; and Wattenberg, M. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36: 41451–41530.
- Lin, S.; Hilton, J.; and Evans, O. 2021. TruthfulQA: Measuring How Models Mimic Human Falsehoods. *arXiv:2109.07958*.
- Lindström, A. D.; and Abraham, S. S. 2022. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *International Joint Conference on Learning and Reasoning, 16th International Workshop on Neural-Symbolic Learning and Reasoning (NeSy 2022), Windsor, UK, September 28-30, 2022*.
- Liu, D.; Nassereldine, A.; Yang, Z.; Xu, C.; Hu, Y.; Li, J.; Kumar, U.; Lee, C.; Qin, R.; Shi, Y.; et al. 2024. Large language models have intrinsic self-correction ability. *arXiv preprint arXiv:2406.15673*.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.
- Liu, X.; Fatahi Bayat, F.; and Wang, L. 2024. Enhancing Language Model Factuality via Activation-Based Confidence Calibration and Guided Decoding. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 10436–10448. Miami, Florida, USA: Association for Computational Linguistics.
- Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

- Lyu, Q.; Havaldar, S.; Stein, A.; Zhang, L.; Rao, D.; Wong, E.; Apidianaki, M.; and Callison-Burch, C. 2023. Faithful chain-of-thought reasoning. In *The 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2023)*.
- Meng, K.; Bau, D.; Andonian, A.; and Belinkov, Y. 2022. Locating and editing factual associations in gpt. *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, 35: 17359–17372.
- Mitra, C.; Huang, B.; Darrell, T.; and Herzig, R. 2024. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14420–14431.
- Mo, S.; and Xin, M. 2024. Tree of uncertain thoughts reasoning for large language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 12742–12746. IEEE.
- Moschella, L.; Maiorca, V.; Fumero, M.; Norelli, A.; Locatello, F.; and Rodolà, E. 2022. Relative representations enable zero-shot latent space communication. *The Eleventh International Conference on Learning Representations*.
- Patel, A.; Bhattamishra, S.; and Goyal, N. 2021. Are NLP models really able to solve simple math word problems? *NAACL*.
- Rai, D.; and Yao, Z. 2024. An Investigation of Neuron Activation as a Unified Lens to Explain Chain-of-Thought Eliciting Arithmetic Reasoning of LLMs. *arXiv preprint arXiv:2406.12288*.
- Ritter, H.; Botev, A.; and Barber, D. 2018. A scalable laplace approximation for neural networks. In *6th international conference on learning representations, ICLR 2018-conference track proceedings*, volume 6. International Conference on Representation Learning.
- Skean, O.; Arefin, M. R.; Zhao, D.; Patel, N.; Naghiyev, J.; LeCun, Y.; and Shwartz-Ziv, R. 2025. Layer by Layer: Uncovering Hidden Representations in Language Models. *arXiv preprint arXiv:2502.02013*.
- Sprague, Z.; Yin, F.; Rodriguez, J. D.; Jiang, D.; Wadhwa, M.; Singhal, P.; Zhao, X.; Ye, X.; Mahowald, K.; and Durrett, G. 2024. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. *arXiv preprint arXiv:2409.12183*.
- Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1): 1929–1958.
- Suzgun, M.; Scales, N.; Schärli, N.; Gehrmann, S.; Tay, Y.; Chung, H. W.; Chowdhery, A.; Le, Q. V.; Chi, E. H.; Zhou, D.; ; and Wei, J. 2023. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. *ACL*.
- Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4149–4158. Minneapolis, Minnesota: Association for Computational Linguistics.
- Team, Q. 2025. Qwen2.5-VL.
- Tenney, I.; Das, D.; and Pavlick, E. 2019. BERT rediscovers the classical NLP pipeline. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.
- Tian, K.; Mitchell, E.; Zhou, A.; Sharma, A.; Rafailov, R.; Yao, H.; Finn, C.; and Manning, C. D. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.
- Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Wang, L.; Hu, Y.; He, J.; Xu, X.; Liu, N.; Liu, H.; and Shen, H. T. 2024. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19162–19170.
- Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023. Self-consistency improves chain of thought reasoning in language models. *The Eleventh International Conference on Learning Representations*.
- Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.
- Xie, Y.; Kawaguchi, K.; Zhao, Y.; Zhao, J. X.; Kan, M.-Y.; He, J.; and Xie, M. 2023. Self-evaluation guided beam search for reasoning. *Advances in Neural Information Processing Systems*, 36: 41618–41650.
- Xie, Z.; Guo, J.; Yu, T.; and Li, S. 2024. Calibrating reasoning in language models with internal consistency. *NeurIPS 2024*.
- Yang, Y.; Yih, W.-t.; and Meek, C. 2015. WikiQA: A Challenge Dataset for Open-Domain Question Answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2013–2018. Lisbon, Portugal: Association for Computational Linguistics.
- Yin, Z.; Sun, Q.; Guo, Q.; Zeng, Z.; Li, X.; Dai, J.; Cheng, Q.; Huang, X.-J.; and Qiu, X. 2024. Reasoning in flux: Enhancing large language models reasoning through uncertainty-aware adaptive guidance. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2401–2416.
- Zhang, R.; Jiang, D.; Zhang, Y.; Lin, H.; Guo, Z.; Qiu, P.; Zhou, A.; Lu, P.; Chang, K.-W.; Gao, P.; et al. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *Computer Vision – ECCV 2024*.
- Zhang, Z.; Zhang, A.; Li, M.; and Smola, A. 2022. Automatic chain of thought prompting in large language models. *International Conference on Learning Representations*.
- Zhang, Z.; Zhang, A.; Li, M.; Zhao, H.; Karypis, G.; and Smola, A. 2023. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*.

Zhao, J.; Liu, R.; Zhang, K.; Zhou, Z.; Gao, J.; Li, D.; Lyu, J.; Qian, Z.; Qi, B.; Li, X.; et al. 2025. Genprm: Scaling test-time compute of process reward models via generative reasoning. *arXiv preprint arXiv:2504.00891*.

Zheng, G.; Yang, B.; Tang, J.; Zhou, H.-Y.; and Yang, S. 2023. Ddcot: Duty-distinct chain-of-thought prompting for multi-modal reasoning in language models. *Advances in Neural Information Processing Systems*, 36: 5168–5191.