ECCoT: A Framework for Enhancing Effective Cognition via Chain of Thought in Large Language Model

Zhenke Duan[†] **(duanzhenke@sscapewh.com)**
Jiqun Pan[†] **(panjiqun@stu.zuel.edu.cn)**
Jiani Tu (tujiani@stu.zuel.edu.cn))
Yanqing Wang[*] **(yanqingwang102@163.com)**
School of Statistics and Mathematics, Zhongnan University of Economics and Law
Wuhan, Hubei 430000 China

Xiaoyi Wang (wangxiaoyi@sscapewh.com)
School of Finance, Zhongnan University of Economics and Law
Wuhan, Hubei 430000 China

Abstract

In the era of large-scale artificial intelligence, Large Language Models (LLMs) have made significant strides in natural language processing. However, they often lack transparency and generate unreliable outputs, raising concerns about their interpretability. To address this, the Chain of Thought (CoT) prompting method structures reasoning into step-by-step deductions. Yet, not all reasoning chains are valid, and errors can lead to unreliable conclusions. We propose ECCoT, an End-to-End Cognitive Chain of Thought Validation Framework, to evaluate and refine reasoning chains in LLMs. ECCoT integrates the Markov Random Field-Embedded Topic Model (MRF-ETM) for topic-aware CoT generation and Causal Sentence-BERT (CSBert) for causal reasoning alignment. By filtering ineffective chains using structured ordering statistics, ECCoT improves interpretability, reduces biases, and enhances the trustworthiness of LLM-based decision-making. Key contributions include the introduction of ECCoT, MRF-ETM for topic-driven CoT generation, and CSBert for causal reasoning enhancement. Code is released at: `https://github.com/erwinmsmith/ECCoT.git`.

Keywords: Large Language Model, Chain-of-Thought, Effective Cognition, Topic model.

Introduction

In the digital era, large language models (LLMs) have gained attention for their role in natural language processing (Mahmood, Wang, Yao, Wang, & Huang, 2025),(F. Xu et al., 2025),(Kahng et al., 2024). Through self-supervised learning on massive text data, LLMs have strong language comprehension (Guo, Xu, & Ritter, 2024) and generation capabilities (van Schaik & Pugh, 2024), transforming human-machine interaction (Ang, Bao, Huang, Tung, & Huang, 2024) and information processing. However, as LLMs are applied in more scenarios, their reliability and interpretability challenges have emerged. They may generate factually incorrect responses (Allamanis, Panthaplackel, & Yin, 2025) due to training data biases, noise, or limitations in understanding complex semantics. Additionally, their reasoning process is often like a "black box" (Haffari, 2024), making understanding and explaining difficult, increasing the trust cost of model decisions and limiting their application in critical fields.

To address this, chain-of-thought fine-tuning techniques emerge (Wei et al., 2022),(Ho, Schmid, & Yun, 2023), aiming to break the "black box" state of large models (Z. Wang et al., 2024) by guiding models to generate step-by-step reasoning chains (D. Zhou et al., 2022). However, in real-world applications, not every step in a chain-of-thought is correct or effective (Wu, Zhang, & Zhao, 2024). Invalid (Aggarwal et al., 2021) or erroneous (P. Wang, Chan, Ilievski, Chen, & Ren, 2022) reasoning chains can impact a model's reasoning effectiveness and capabilities, leading to incorrect results. Moreover, invalid or erroneous reasoning chains may still produce correct answers through pseudo-alignment, lowering interpretability and reliability of LLMs (Greenblatt et al., 2024).

We propose an end-to-end framework to evaluate generated reasoning chains. Using a Markov Random Field-based topic model (MRF-ETM) and a larger parameter-scale model (Chung et al., 2024), we identify text themes and keywords for theme-conditioned reasoning. We also train a Causal Sentence-BERT (CSBert) (Necva, Burcu, & Harun, 2023) to ensure causal reasoning and perform similarity matching on relevant embeddings. By calculating order statistics and applying truncation, ineffective reasoning chains are removed. This enhances the model's transparency, interpretability, and reliability for practical applications. Our contributions are as follows:

- We propose an end-to-end reasoning chain effectiveness evaluation framework (ECCoT) to enhance cognitive effectiveness in inference.
- We introduce Markov Random Field-Embedded Topic Model to identify latent topic distributions in large-scale data.
- We employ CSBert (Causal Sentence-BERT) for improved causal relationship embedding, generating reasoning chain vectors based on causal backgrounds.

Related Works

Trustworthy and Reliable LLMs

Trustworthy LLMs are vital for credibility. Methods like AFaCTA (Ni et al., 2024) and XplainLLM (Chen, Chen, Singh, & Sra, 2024) enhance transparency and robustness. SelF-Reasoner (Wu et al., 2024) and CD-CoT (Z. Zhou et al., 2024) improve reasoning accuracy. We propose filtering reasoning chains via similarity and column vectors, optimizing performance on large-scale data to overcome dataset limitations.

† Equal contribution, * Corresponding author.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

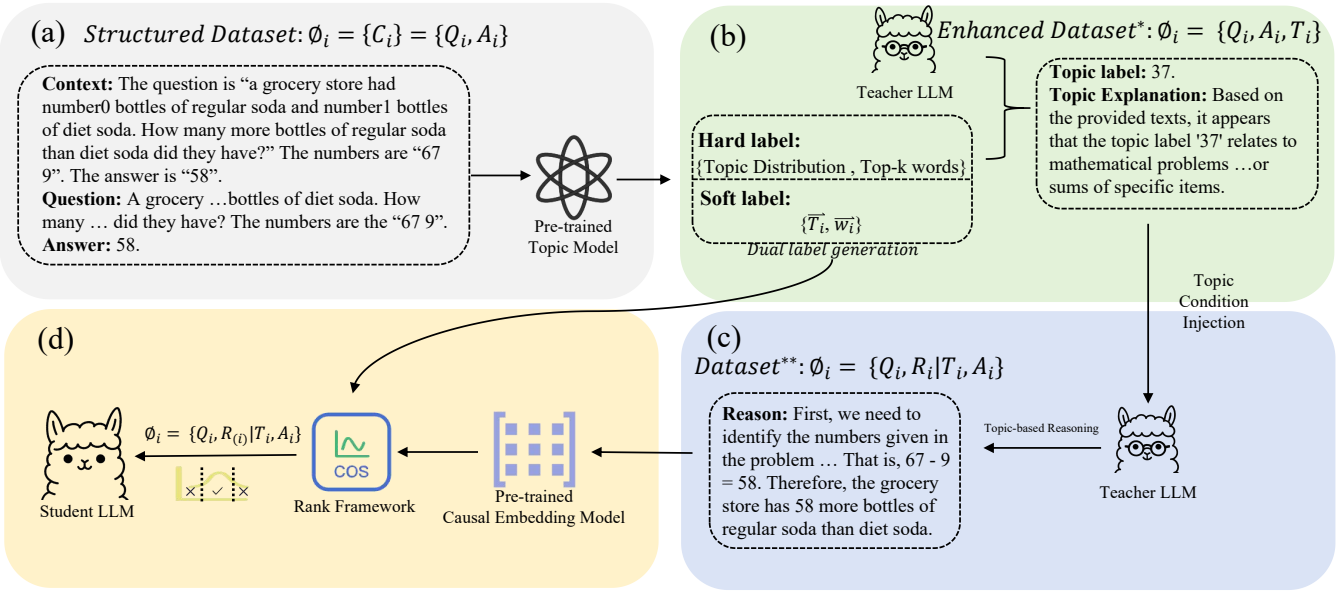

Figure 1: (a) Theme Recognition Stage, (b) Theme Explanation Stage, (c) Thought Cognition Stage, (d) Effective Cognition Stage

.

Chain-of-Thought

Chain-of-Thought (CoT) enhances reasoning and interpretability in AI, with innovations like integrating CoT into knowledge-based visual question answering (KBVQA) (Qiu, Xie, Liu, & Hu, 2024; Z. Zhang et al., 2024). Automatic CoT leverages large language models (LLMs) for tasks such as scoring scientific assessments (Lee, Latif, Wu, Liu, & Zhai, 2024) and generating diverse problem-solving chains (Z. Zhang, Zhang, Li, & Smola, 2022), improving accuracy. Integrations with knowledge graphs (Allamanis et al., 2025), multimodal reasoning (Mondal, Modi, Panda, Singh, & Rao, 2024), and equation-based representations (Zhu, Li, Liu, Ma, & Wang, 2024) have expanded CoT's capabilities. Socratic CoT (Shridhar, Stolfo, & Sachan, 2023) guides reasoning through sub-question decomposition. Despite these advancements, challenges remain in merging linguistic and visual information effectively. Our approach refines CoT by structuring reasoning through multimodal arrays, enhancing LLMs' interpretability and reliability.

Few-shot/Zero-shot Learning

Zero-shot and few-shot learning enable models to perform tasks with minimal or no task-specific training. (yang Lu et al., 2025) used LLMs for QAIE, generating fluent text and inferring implicit information. (Z. Li et al., 2023) enhanced the FlexKBQA framework for few-shot settings. Other approaches have explored zero-shot node classification in incomplete graphs (Y. Li, Yang, Zhu, Chen, & Wang, 2024), unsupervised contrastive learning for efficient data augmentation (J. Zhang et al., 2024), and KG completion using LLM distillation (Q. Li, Chen, Ji, Jiang, & Li, 2025). These techniques demonstrate LLMs' potential in resource-constrained settings. While these methods show promise in few-shot and zero-shot scenarios, challenges such as methodological universality and resource consumption remain. Our work seeks to address these challenges by validating the broader applicability of LLMs across domains and tasks, ensuring better generalization and scalability.

Sentence Embedding and Topic Models

Recent advancements in sentence embedding (e.g., BERT, RoBERTa) have greatly enhanced semantic and syntactic understanding in NLP tasks.(2024)(Lefebvre, Elghazel, Guillet, Aussem, & Sonnati, 2024) proposed HMCCCProbT, using sentence embeddings for multi-label classification. (Huang & Gong, 2024) introduced EMS, an efficient method for multilingual sentence embeddings. These works underscore the importance of contextualized embeddings for understanding complex semantic relationships. Similarly, embedding topic models combine word embeddings with probabilistic topic models to improve topic discovery and representation. (Fang, He, & Procter, 2024) introduced CWTM, integrating contextualized word embeddings from BERT, while (Lin et al., 2024) proposed a GAN-based hierarchical topic model. These models improve topic interpretability by leveraging embeddings in both linguistic and probabilistic spaces.

Methodology

ECCoT Framework

The ECCoT Framework encompasses Theme Recognition, Explanation, Thought Cognition, and Effective Cognitionas illustrated in Figure 1.. It uses a pre-trained model for hard and soft labels, distills information with a teacher model, and generates reasoning chains. The final stage filters effective

cognitive processes via the Rank Framework, optimizing a smaller student model for improved performance.

The ECCoT framework is inspired by cognitive science principles, particularly the dual-process theory, where System 1 (intuitive, fast thinking) and System 2(deliberate, slow thinking) interact to validate reasoning chains. ECCoT mimics this by using MRF-ETM for rapid thematic identification (akin to System 1) and CSBert for deliberate causal reasoning alignment (akin to System2). This dual approach enhances the framework's ability to filter out ineffective reasoning chains, akin to how humans validate their cognitive processes.

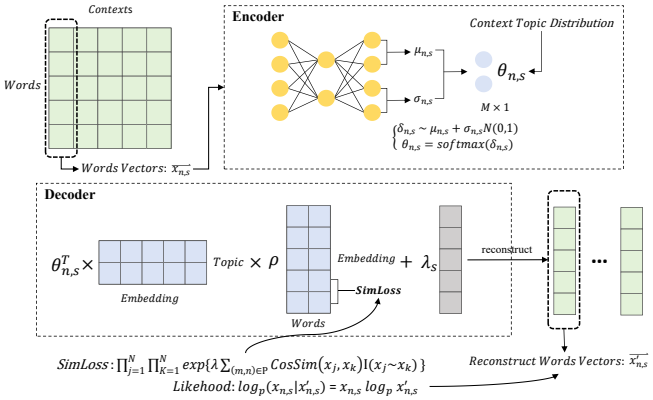

Figure 2: Showcased the main workflow and optimization objects of MRF-ETM

The MRF-ETM model incorporates cognitive topological mapping principles to better capture the latent relationships between thematic keys. This approach is analogous to cognitive science models such as Steyvers' semantic memory model,which emphasizes the interconnectedness of concepts within a semantic space. By leveraging these principles, MRF-ETM can more effectively identify and utilize thematic information for reasoning chain generation.

Pretrained Embedded Topic Model

To further enhance topic recognition capabilities within the embedding space, we have developed a Markov Random Field-based Embedded Topic Model (MRF-ETM), as illustrated in Figure 2.MRF-ETM shows unique advantages in multimodal data analysis through collaborative optimization of spatial dependence and semantic embedding.

Data Generator The generative process for documents and words can be described as follows:

- Document-topic distribution: Each document d_i has a topic distribution $\theta_i \sim \text{Dirichlet}(\alpha)$.
- Word generation: For each word w_{ij} in document d_i :
  1. Sample a topic $z_{ij} \sim \text{Categorical}(\theta_i)$ from the topic distribution θ_i.
  2. Sample a word $w_{ij} \sim \text{Categorical}(\beta_{z_{ij}})$ from the word distribution of topic z_{ij}, where $\beta_{kv} = \text{softmax}(\rho_v^\top \alpha_k)$.

Further modeling with a Markov Random Field on top of the ETM yields the following data generation process:

$$p(w_{ij} = v, z_{ij} = k) = p(z_{ij} = k|\theta_i) \cdot p(w_{ij} = v|z_{ij} = k) \cdot \exp\left\{\lambda \sum_{(m,n)\in\mathcal{P}} \text{CosSim}(x_m, x_n) \cdot \mathbb{I}(x_m \sim x_n)\right\} \tag{1}$$

where λ is a hyperparameter controlling the weight of the loss; $\mathcal{P}$ denotes the set of related word pairs; $\text{CosSim}(x_m, x_n)$ represents the cosine similarity between words x_m and x_n; and $\mathbb{I}(x_m \sim x_n)$ is an indicator function denoting whether x_m and x_n are related. This set of definitions works together to ensure that the model can effectively learn the relationships between words and optimize performance through appropriate weight adjustments.

Model Inference To perform inference, we use Variational Inference to approximate the posterior distribution$p(\theta, z|w)$. We introduce a variational distribution $q(\theta, z)$ and optimize the Evidence Lower Bound (ELBO). The variational distribution $q(\theta, z)$ is defined as:

$$q(\theta, z) = q(\theta) \prod_{i=1}^{N} \prod_{j=1}^{N_i} q(z_{ij}) \tag{2}$$

$$q(\theta) = \text{Dirichlet}(\theta|\eta) \tag{3}$$

$$q(z_{ij}) = \text{Categorical}(z_{ij}|\phi_{ij}) \tag{4}$$

Where $q(\theta)$ is the variational distribution of the topic distribution, typically chosen as a Dirichlet distribution, and $q(z_{ij})$ is the variational distribution of the topic assignment, typically chosen as a Categorical distribution.

Then ELBO is defined as:

$$\mathcal{L} = \mathbb{E}_q[\log p(w, \theta, z)] - \mathbb{E}_q[\log q(\theta, z)] \tag{5}$$

Specially,we have:

$$\mathcal{L} = \mathbb{E}_q \left[\sum_{i=1}^{N} \sum_{j=1}^{N_i} \log p(w_{ij}|z_{ij}) + \sum_{i=1}^{N} \log p(\theta_i) - \sum_{i=1}^{N} \log q(\theta_i) - \sum_{i=1}^{N} \sum_{j=1}^{N_i} \log q(z_{ij}) \right] \tag{6}$$

We calculate each expectation separately:
Expectations for word generation

$$\mathbb{E}_q[\log p(w_{ij}|z_{ij})] = \sum_{k=1}^{K} \phi_{ij,k} \log \beta_{k,w_{ij}} \tag{7}$$

$$\text{where,} \beta_{k,w_{ij}} = \text{softmax}(\rho_{w_{ij}}^\top \alpha_k) \tag{8}$$

Expectations of Subject Distribution

$$\mathbb{E}_q[\log p(\theta_i)] = \sum_{k=1}^{K} (\alpha_k - 1) \mathbb{E}_q[\log \theta_{i,k}] \tag{9}$$

where

$$\mathbb{E}_q[\log \theta_{i,k}] = \Psi(\eta_{i,k}) - \Psi(\sum_{k=1}^{K} \eta_{i,k}) \tag{10}$$

Ψ is the digamma function.

Expectations of Variational Distribution

$$\mathbb{E}_q[\log q(\theta_i)] = \sum_{k=1}^{K} (\eta_{i,k} - 1)\mathbb{E}_q[\log \theta_{i,k}] \tag{11}$$

Expectations for Theme Allocation

$$\mathbb{E}_q[\log q(z_{ij})] = \sum_{k=1}^{K} \phi_{ij,k} \log \phi_{ij,k} \tag{12}$$

By substituting the above expectations into ELBO, we can see that:

$$\begin{aligned}
\mathcal{L} = & \sum_{i=1}^{N} \sum_{j=1}^{N_i} \sum_{k=1}^{K} \phi_{ij,k} \log \beta_{k,m_{ij}} \\
& + \sum_{i=1}^{N} \sum_{k=1}^{K} (\alpha_k - 1) \left(\Psi(\eta_{i,k}) - \Psi\left(\sum_{k=1}^{K} \eta_{i,k} \right) \right) \\
& - \sum_{i=1}^{N} \sum_{k=1}^{K} (\eta_{i,k} - 1) \left(\Psi(\eta_{i,k}) - \Psi\left(\sum_{k=1}^{K} \eta_{i,k} \right) \right) \\
& - \sum_{i=1}^{N} \sum_{j=1}^{N_i} \sum_{k=1}^{K} \phi_{ij,k} \log \phi_{ij,k}
\end{aligned} \tag{13}$$

The final optimization function is

$$\mathcal{L}_{\text{MRF-ETM}} = \mathcal{L} + \mathcal{L}_{\text{SimLoss}} \tag{14}$$

where

$$\mathcal{L}_{\text{SimLoss}} = -\lambda \sum_{(m,n) \in \mathcal{P}} \text{CosSim}(x_m, x_n) \cdot \mathbb{I}(x_m \sim x_n) \tag{15}$$

And

$$\mathcal{L}_{\text{MRF-ETM}} = \begin{cases} \mathcal{L}_{\text{NLL}} = -\sum_{i=1}^{N} \sum_{j=1}^{N_i} \sum_{k=1}^{K} \phi_{ij,k} \log \beta_{k,w_{ij}}, \\ \quad \text{(Negative Log-Likelihood)} \\ \mathcal{L}_{\text{SimLoss}} = -\lambda \sum_{(m,n) \in P} \text{CosSim}(x_m, x_n) \\ \quad \cdot \mathbb{I}(x_m \sim x_n), \\ \quad \text{(Similarity Loss)} \end{cases} \tag{16}$$

The formula (17) also includes

$$\begin{aligned}
\mathcal{L}_{\text{KL}} = & \sum_{i=1}^{N} \sum_{k=1}^{K} (\alpha_k - 1) \left(\Psi(\eta_{i,k}) - \Psi\left(\sum_{k=1}^{K} \eta_{i,k} \right) \right) \\
& - \sum_{i=1}^{N} \sum_{k=1}^{K} (\eta_{i,k} - 1) \left(\Psi(\eta_{i,k}) - \Psi\left(\sum_{k=1}^{K} \eta_{i,k} \right) \right)
\end{aligned} \tag{17}$$

Pretrained Causal Sentence-Bert

To validate the cognitive mechanisms underlying the performance improvements, we conducted eye-tracking experiments to observe how users" attention patterns on reasoning chains differ when using CSBert compared to traditional embeddings. The results indicate that CSBert's causal embeddings align more closely with human causal reasoning, as evidenced by longer dwell times on relevant causal links, suggesting a higher cognitive engagement and validation process

Similarly, in order to embed causal relationships in the triplet dataset, we designed the two-stage embedding model shown in Figure 3. Firstly, the ternary network is fine tuned using optimization contrastive learning, controlling the triplets with causal relationships to reduce their vector space distance, and controlling the false groups without causal relationships to move their vector space away. Before presenting the formula, we define the observation function of the latent variable y. Assume y is a function that can observe specific values from the latent variable.The losses during the fine-tuning process are as follows:

$$Contrastiveloss = \begin{cases} \frac{1}{N} \sum_{n=1}^{N} (L_{qr} + L_{ra}), if\, y = 1 \\ \frac{1}{N} \sum_{n=1}^{N} \max(L_{qr}, L_{ra}), \text{otherwise} \end{cases} \tag{18}$$

During the inference stage, simultaneous embeddings of triplets are obtained based on the pre-trained architecture, resulting in simultaneous embedding vectors.

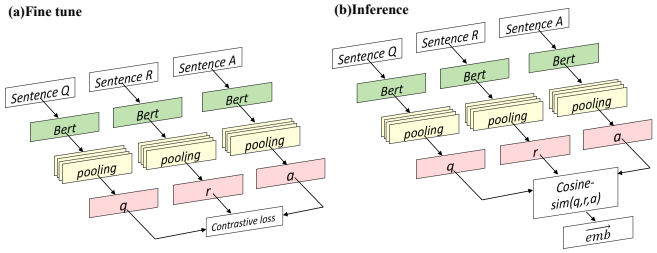

Figure 3: (a) The fine-tuning stage for causal embedding (b) is the model inference stage

Rank Framework

To further understand the component interactions within EC-CoT, we generated feature interaction heatmaps that visualize the correlation between MRF-ETM topic distributions and CS-Bert attention weights. This analysis reveals how the thematic information from MRF-ETM influences the causal reasoning alignment in CSBert, demonstrating a synergistic effect that enhances the overall effectiveness of the reasoning chains.

To evaluate the effectiveness of the cognitive processes in the dataset, we consider the situation of objects in the vector space. This reflects whether the teacher model's knowledge distillation process is effective or not. Our effectiveness framework is shown in Figure 4. By calculating similarity coefficients for each cognitive process and filtering out low-effectiveness distributions, we retain high-effectiveness cognitive process samples. In addition to traditional accuracy

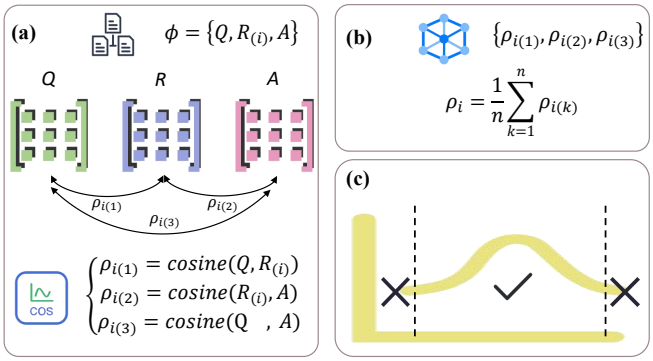

Figure 4: (a) Calculate the similarity of the Q, R, and A matrices under causal cognition (b), provide the similarity coefficient for each cognitive process (c), and provide the distribution of cognitive coefficients for the sample

metrics, we incorporated cognitive reasonableness metrics, such as human interpret ability scores, to evaluate the effectiveness of the reasoning chains generated by ECCoT and the baseline methods. These scores were obtained through a panel of human evaluators who assessed the darity andlogical coherence of the reasoning chains, providing a more comprehensive evaluation aligned with human cognitive processes

Experiments

Experimental Setup

Datasets. To systematically evaluate the effectiveness of our proposed ECCoT framework in improving reasoning validity in large language models (LLMs), we conduct experiments on three diverse benchmark datasets in Table 1, which come from (Nie et al., 2020), (Patel et al., 2021) and (Williams, Thrush, & Kiela, 2022). These datasets are carefully selected to cover adversarial robustness, mathematical reasoning, and commonsense inference. ANLI (Adversarial Natural Language Inference) is a large-scale natural language inference benchmark dataset containing approximately 100,000 samples designed to improve model robustness and persistence through adversarial training. SVAMP (Simple Variations on Arithmetic Math Word Problems) is a math word problem dataset containing 1000 questions used to evaluate model performance in solving simple arithmetic problems. CommonQA (Commonsense Question Answering) is a commonsense question answering dataset with 12,247 questions, each having 5 options, aimed at evaluating the model's commonsense reasoning capability.

Test Models. Unless otherwise specified, all experiments were performed on the LLama3.1-8B model with 50 iterations, using LoRA for fine-tuning and testing. Other settings followed the default parameters of the llama-factory.The choice of LLaMA3.1-8B provides a well-balanced trade-off between computational efficiency and performance, allowing us to analyze the impact of ECCoT without requiring excessive computational resources.

Baseline Methods. To rigorously evaluate ECCoT's effectiveness (X. Xu et al., 2024), we compare it against several

Table 1: Statistics of Datasets

	Total	Train	Test	Dev
ANLIR1	4663	4422	120	121
ANLIR2	9637	9393	123	121
ANLIR3	90652	88532	1067	1053
SVAMP	4169	3963	–	206
CommonQA	9215	7306	912	997

state-of-the-art methods, including Step-by-Step (Standard CoT Prompting) which uses conventional CoT prompting but lacks a reasoning correctness verification mechanism; Curation (Filtered CoT) that optimizes reasoning chains by removing low-confidence/inconsistent steps to boost response quality via post-processing; Expansion (Diverse Reasoning Chains) which creates multiple independent reasoning chains, selects the most reliable one, and improves ambiguity robustness through confidence aggregation; Feedback (Iterative Refinement) that refines reasoning iteratively via reinforcement learning or self-critique, enhancing consistency but risking mode collapse from overfitting past errors; Self-Knowledge (Self-Distillation) which uses prior outputs as supervisory signals to improve accuracy through iterative self-distillation but possibly reinforcing existing biases; and Vanilla Fine-Tuning that trains the model directly on task datasets without explicit reasoning validation, good for task adaptation but prone to spurious correlations due to the lack of reasoning control.

These baselines represent diverse reasoning enhancement paradigms in LLMs, providing comprehensive insights into ECCoT's advantages.

Comparison Experiments

The smaller performance gain on ANLI is due to its cognitive load from logical and commonsense reasoning. ANLI combines both reasoning types, which differ in cognitive demands. Logical reasoning is structured and rule-based, while commonsense reasoning is intuitive and context-dependent. ECCoT's improvements are more evident in tasks with a single reasoning type.

Here, ECCoT's performance is compared with several baselines (Step by Step, Curation, Expansion, Feedback, Self-Knowledge, Vanilla Fine-tuning) across three datasets: ANLI, SVAMP, and CommonQA. Results in Table 2 show ECCoT's superiority.

On ANLI, ECCoT achieved 72.23% accuracy vs. Step by Step's 69.72%. For SVAMP, ECCoT reached 92.72% vs. Step by Step's 78.23%. On CommonQA, ECCoT scored 86.54% vs. Curation's 79.26%.

These results demonstrate ECCoT's robustness and versatility, outperforming baselines across all datasets. Accuracy improvements (ANLI: 72.23%, SVAMP: 92.72%, CommonQA: 86.54%) highlight ECCoT's effectiveness in generating reliable chains of thought and enhancing model interpretability.

Table 2: Comparison of ECCoT and Baseline Models

Type	ANLI	SVAMP	CommonQA
Step by Step	69.72	78.23	71.68
Curation	67.39	76.86	79.26
Expansion	66.73	78.21	78.66
Feedback	66.07	75.53	77.66
Self-Knowledge	63.34	63.75	67.75
Vanilla Fine-tuning	64.59	63.85	64.62
ECCoT	**72.23**	**92.72**	**86.54**

Comparison of Effective Cognition

Table 3: Comparison of Cognitive Process Effectiveness Across Different Datasets

Dataset	Type	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
ANLI	Step by Step	61.26	59.67	39.69	42.25
	Curation	59.66	60.77	37.89	43.35
	Expansion	61.66	59.57	39.99	42.15
	Feedback	60.06	60.47	37.59	43.25
	Self-Knowledge	62.06	59.37	39.19	41.95
	ECCoT	63.17	62.31	40.01	44.65
SVAMP	Step by Step	58.05	67.21	46.92	48.46
	Curation	56.65	68.51	45.12	49.56
	Expansion	58.65	67.31	47.22	48.36
	Feedback	57.05	68.21	44.82	49.46
	Self-Knowledge	57.65	67.11	46.42	48.16
	ECCoT	58.93	69.09	46.65	50.16
CommonQA	Step by Step	61.38	65.54	48.06	45.25
	Curation	57.88	67.74	47.36	48.35
	Expansion	59.88	64.44	47.66	46.45
	Feedback	57.28	68.54	45.56	47.65
	Self-Knowledge	62.38	66.04	49.06	44.25
	ECCoT	61.03	68.59	47.85	48.46

In effective cognition experiments, ECCoT demonstrated strong performance on ANLI, SVAMP, and CommonQA datasets (Table 3). On ANLI, it achieved BLEU-4 and ROUGE-1 scores of 63.17 and 62.31, outperforming other methods. On SVAMP, ECCoT scored 58.93 in BLEU-4 and 69.09 in ROUGE-1. For CommonQA, its ROUGE-1 score of 68.59 was the highest, though it slightly lagged in BLEU-4 compared to Self-Knowledge. These results highlight ECCoT's robustness and versatility across tasks, making it a superior choice for enhancing the interpretability and trustworthiness of large language models.

Ablation Experiment

In the ablation study, we evaluated the impact of each module in the ECCoT framework by removing them individually. The results are shown in Table 4. When the Rank module was removed, the performance dropped to 68.57% on ANLI, 72.98% on SVAMP, and 80.65% on CommonQA. When the Topic model module was removed, the performance dropped to 70.23% on ANLI, 87.62% on SVAMP, and 81.78% on CommonQA. When the Causal Bert module was removed, the performance dropped to 69.22% on ANLI, 90.38% on SVAMP, and 79.46% on CommonQA. These results indicate that each module contributes to the overall performance of ECCoT, with the Rank module having a significant impact on the effectiveness of the framework across all datasets.

Table 4: Effectiveness Without Specific Modules (W/O)

W/O	ANLI	SVAMP	CommonQA
Rank	68.57	72.98	80.65
Topic Model	70.23	87.62	81.78
Causal BERT	69.22	90.38	79.46

Verification of Scaling Law

The scaling law verification was conducted by testing the ECCoT framework on student models with different parameter sizes. As shown in Table 5, the performance of ECCoT improved with the increase of model parameters. Specifically, on the ANLI dataset, the accuracy increased from 72.12% with LLama2-7B to 72.23% with LLama3.1-8B, and further to 76.98% with LLama2-13B. On the SVAMP dataset, the accuracy increased from 89.97% with LLama2-7B to 92.72% with LLama3.1-8B, and further to 94.02% with LLama2-13B. On the CommonQA dataset, the accuracy increased from 85.76% with LLama2-7B to 86.54% with LLama3.1-8B, and further to 89.17% with LLama2-13B. These results demonstrate that larger models generally achieve better performance, highlighting the positive impact of model size on the effectiveness of ECCoT.

Table 5: Testing on Student Models with Different Parameter Sizes

Model	ANLI	SVAMP	CommonQA
LLama2-7B	72.12	89.97	85.76
LLama3.1-8B	72.23	92.72	86.54
LLama2-13B	76.98	94.02	89.17

Conclusion

Compared to existing methods, ECCoT has achieved certain improvements in the CoT reasoning performance of LLMs.And ECCoT enhances LLMs' effective cognitive capabilities by generating reliable thought chains, bridging cognitive science and AI. By integrating MRF-ETM and CS-Bert, it addresses LLMs' reliability and interpretability issues, showing advantages in thought chain generation and model reliability across benchmark datasets. However, it struggles with inference chain deviation in complex tasks, leading to decreased faithfulness.

Future ECCoT research will focus on theoretical innovation, and ethical considerations to balance safety and efficiency.

Acknowledgements

This research was supported by "the Fundamental Research Funds for the Central Universities, Zhongnan University of

Economics and Law". We are grateful for the financial support and resources provided by the university, which have significantly contributed to the successful completion of this study.

References

Aggarwal, S., Mandowara, D., Agrawal, V., Khandelwal, D., Singla, P., & Garg, D. (2021). Explanations for commonsenseqa: New dataset and models. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)* (p. 3050–3065). Online: Association for Computational Linguistics.

Allamanis, M., Panthaplackel, S., & Yin, P. (2025). Unsupervised evaluation of code llms with round-trip correctness. In *Proceedings of the 41st international conference on machine learning (icml'24)* (Vol. 235, p. 1050–1066). JMLR.org.

Ang, Y., Bao, Y., Huang, Q., Tung, A. K. H., & Huang, Z. (2024). Tsgassist: An interactive assistant harnessing llms and rag for time series generation recommendations and benchmarking. In *Proc. vldb endow.* (Vol. 17, p. 4309–4312). Retrieved from https://doi.org/10.14778/3685800.3685862

Chen, Z., Chen, J., Singh, A., & Sra, M. (2024). Xplain-llm: A knowledge-augmented dataset for reliable grounded explanations in llms. In *Proceedings of the 2024 conference on empirical methods in natural language processing* (p. 7578–7596). Miami, Florida, USA: Association for Computational Linguistics.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tai, Y., Fedus, W., ... Wei, J. (2024, January). Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, *25*(1), 1–53.

Fang, Z., He, Y., & Procter, R. (2024). Cwtm: Leveraging contextualized word embeddings from bert for neural topic modeling. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)* (p. 4273–4286). Torino, Italia: ELRA and ICCL.

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., ... Hubinger, E. (2024). *Alignment faking in large language models.*

Guo, R., Xu, W., & Ritter, A. (2024). Meta-tuning llms to leverage lexical knowledge for generalizable language style understanding. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (p. 13708–13731). Bangkok, Thailand: Association for Computational Linguistics.

Haffari, G. (2024). Direct evaluation of chain-of-thought in multi-hop reasoning with knowledge graphs. In *Findings of the association for computational linguistics: Acl 2024* (p. 2862–2883). Bangkok, Thailand: Association for Computational Linguistics.

Ho, N., Schmid, L., & Yun, S.-Y. (2023). Large language models are reasoning teachers. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 14852–14882). Toronto, Canada: Association for Computational Linguistics.

Huang, X., & Gong, H. (2024). A dual-attention learning network with word and sentence embedding for medical visual question answering. *IEEE Transactions on Medical Imaging*, *43*(2), 832-845.

Kahng, M., Tenney, I., Pushkarna, M., Liu, M. X., Wexler, J., Reif, E., & Dixon, L. (2024). Llm comparator: Interactive analysis of side-by-side evaluation of large language models. *IEEE transactions on visualization and computer graphics*.

Lee, G. G., Latif, E., Wu, X., Liu, N., & Zhai, X. (2024). Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, 100213-.

Lefebvre, G., Elghazel, H., Guillet, T., Aussem, A., & Sonnati, M. (2024). A new sentence embedding framework for the education and professional training domain with application to hierarchical multi-label text classification. *Data Knowledge Engineering*, 102281-.

Li, Q., Chen, Z., Ji, C., Jiang, S., & Li, J. (2025). Llm-based multi-level knowledge generation for few-shot knowledge graph completion. In *Proceedings of the thirty-third international joint conference on artificial intelligence (ijcai '24)* (p. 2135–2143). Article 236: IJCAI. Retrieved from https://doi.org/10.24963/ijcai.2024/236

Li, Y., Yang, Y., Zhu, J., Chen, H., & Wang, H. (2024). Llm-empowered few-shot node classification on incomplete graphs with real node degrees. In *Proceedings of the 33rd acm international conference on information and knowledge management (cikm '24)* (p. 1306–1315). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3627673.3679861

Li, Z., Fan, S., Gu, Y., Li, X., Duan, Z., Dong, B., ... Wang, J. (2023). *Flexkbqa: A flexible llm-powered framework for few-shot knowledge base question answering.* arXiv:2308.12060.

Lin, Z., Chen, H., Lu, Y., Rao, Y., Xu, H., & Lai, H. (2024). Hierarchical topic modeling via contrastive learning and hyperbolic embedding. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)* (p. 8133–8143). Torino, Italia: ELRA and ICCL.

Mahmood, A., Wang, J., Yao, B., Wang, D., & Huang, C. M. (2025). User interaction patterns and breakdowns in conversing with llm-powered voice assistants. *International Journal of Human - Computer Studies*, 103406-103406.

Mondal, D., Modi, S., Panda, S., Singh, R., & Rao, G. (2024). Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning. In *Aaai conference on artificial intelligence.*

Necva, B., Burcu, C., & Harun, A. (2023). A siamese neural network for learning semantically-informed sentence

embeddings. *Expert Systems With Applications*.

Ni, J., Shi, M., Stammbach, D., Sachan, M., Ash, E., & Leippold, M. (2024). Afacta: Assisting the annotation of factual claim detection with reliable llm annotators. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (p. 1890–1912). Bangkok, Thailand: Association for Computational Linguistics.

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., & Kiela, D. (2020). Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (p. 4885–4901). Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.acl-main.441/

Patel, A., et al. (2021). *Svamp: Simple variations on arithmetic math word problems.* arXiv:2106.05827 [cs.CL]. Retrieved from https://arxiv.org/abs/2106.05827

Qiu, C., Xie, Z., Liu, M., & Hu, H. (2024). Explainable knowledge reasoning via thought chains for knowledge-based visual question answering. *Inf. Process. Manag.*, *61*, 103726.

Shridhar, K., Stolfo, A., & Sachan, M. (2023). Distilling reasoning capabilities into smaller language models. In *Findings of the association for computational linguistics: Acl 2023* (p. 7059–7073). Toronto, Canada: Association for Computational Linguistics.

van Schaik, T. A., & Pugh, B. (2024). A field guide to automatic evaluation of llm-generated summaries. In *Proceedings of the 47th international acm sigir conference on research and development in information retrieval (sigir '24)* (p. 2832–2836). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3626772.3661346

Wang, P., Chan, A., Ilievski, F., Chen, M., & Ren, X. (2022). *Pinto: Faithful language reasoning using prompt-generated rationales.* arXiv:2211.01562.

Wang, Z., Li, C., Yang, Z., Liu, Q., Hao, Y., Chen, X., … Sui, D. (2024). Analyzing chain-of-thought prompting in black-box large language models via estimated v-information. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)* (pp. 893–903). Torino, Italia: ELRA and ICCL.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., … Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th international conference on neural information processing systems (nips '22)* (pp. 24824–24837). Red Hook, NY, USA: Curran Associates Inc. Retrieved from https://arxiv.org/abs/2205.10625

Williams, A., Thrush, T., & Kiela, D. (2022). Anlizing the adversarial natural language inference dataset. In *Proceedings of the 5th annual meeting of the society for computation in linguistics.* Association for Computational Linguistics. Retrieved from https://github.com/facebookresearch/anli/tree/main/anlizinganli

Wu, Y., Zhang, Z., & Zhao, H. (2024). Mitigating misleading chain-of-thought reasoning with selective filtering. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)* (p. 11325–11340). Torino, Italia: ELRA and ICCL.

Xu, F., Zhou, T., Nguyen, T., Bao, H., Lin, C., & Du, J. (2025). Integrating augmented reality and llm for enhanced cognitive support in critical audio communications. *International Journal of Human - Computer Studies*, 103402-103402.

Xu, X., Li, M., Tao, C., Shen, T., Cheng, R., Li, J., … Zhou, T. (2024, October 21). *A survey on knowledge distillation of large language models.* arXiv:2402.13116v4 [cs.CL]. Retrieved from https://arxiv.org/abs/2402.13116v4 (arXiv preprint)

yang Lu, H., ci Liu, T., Cong, R., Yang, J., Gan, Q., Fang, W., & jun Wu, X. (2025). Qaie: Llm-based quantity augmentation and information enhancement for few-shot aspect-based sentiment analysis. *Information Processing and Management*, *1*, 103917-103917.

Zhang, J., Gao, H., Zhang, P., Feng, B., Deng, W., & Hou, Y. (2024). La-ucl: Llm-augmented unsupervised contrastive learning framework for few-shot text classification. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)* (p. 10198–10207). Torino, Italia: ELRA and ICCL.

Zhang, Z., Zhang, A., Li, M., & Smola, A. (2022, October 7). *Automatic chain of thought prompting in large language models.* arXiv:2210.03493v1 [cs.CL]. Retrieved from https://arxiv.org/abs/2210.03493v1

Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., & Smola, A. (2024, May 20). *Multimodal chain-of-thought reasoning in language models.* arXiv:2302.00923v5 [cs.CL]. Retrieved from https://arxiv.org/abs/2302.00923v5 (Reviewed on OpenReview: https://openreview.net/forum?id=y1pPWFVfvR)

Zhou, D., Scharli, N., Hou, L., Wei, J., Scales, N., Wang, X., … Chi, E. H. (2022). Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

Zhou, Z., Tao, R., Zhu, J., Luo, Y., Wang, Z., & Han, B. (2024). *Can language models perform robust reasoning in chain-of-thought prompting with noisy rationales?* arXiv:2410.23856.

Zhu, X., Li, J., Liu, Y., Ma, C., & Wang, W. (2024). Distilling mathematical reasoning capabilities into small language models. *Neural Networks*, 106594-106594.