Exploring Selective Avoidance for Online User Behavior Analysis: A Forest of Thought Explanation

Xiaohua Wu[1,2], Lin Li[1*], Kaize Shi[3], Xiaohui Tao[3], Jianwei Zhang[4], Yuefeng Li[2]

[1] Wuhan University of Technology, Wuhan, China
[2] Queensland University of Technology, Brisbane, Australia
[3] University of Southern Queensland, Springfield, Australia
[4] Iwate University, Morioka, Japan
{xhwu, cathylilin}@whut.edu.cn, {kaize.shi, xiaohui.tao}@unisq.edu.au, zhang@iwate-u.ac.jp, y2.li@qut.edu.au

Abstract

The response behaviors observed in online user-generated content (UGC) frequently demonstrate non-linear characteristics, such as conditional branching and selective avoidance. These patterns present additional challenges for ensuring the trustworthiness of Large Language Model (LLMs) reasoning, particularly as their unidirectional, left-to-right inference mechanisms may not adequately capture such complex reasoning dynamics. To address this, we propose a Forest of Thought Explanation (FoTE), a novel prompting that models the selective avoidance in UGC while ensuring explanation consensus through reasoning paths across all decision sub-trees. FoTE firstly generates various reasoning paths through an adaptive CoT prompting. Each generated thought is subsequently evaluated through cooperative game theory to quantify its fair influence. The thoughts with the top-k contribution scores are preserved and randomly sampled to emulate selective avoidance for the next reasoning iteration. Through extensive evaluations across three open-source LLMs and two established social science problems (spanning four benchmark datasets), FoTE demonstrates superior success rates compared to competing prompting strategies. Notably, its performance gains increase with the strength of selective avoidance in social problems. The trustworthiness of our FoTE is enhanced by the incorporation of (1) a solid theoretical foundation and (2) a transparent reasoning path that converges toward consensus.

1 Introduction

User-generated content (UGC) refers to the myriad of content created and shared by users on social media platforms that often reflects personal concerns, supporting data-driven analysis in various social decision-making scenarios, such as healthcare service (Liang et al. 2025), brand engagement (Naeem 2020), and happiness prediction (Wu et al. 2025). Selective avoidance, also known as branching or conditional logic, is a survey design technique that automatically determines which question a respondent should answer next based on their previous response. This phenomenon can be widely observed in UGC, where users selectively respond to certain topics while ignoring others.

Recently, large language models (LLMs) have shown significant promise in reasoning, marking a new era in the transformation of some emerging topics in social analysis (Liu et al. 2025). The novel reasoning paradigms of X-of-Thoughts (XoT), such as Chain of Thought (Wei et al. 2022b), Tree of Thought (Yao et al. 2023), and Forest of Thoughts (Bi et al. 2025), significantly enhanced the reasoning ability of LLMs. Traditional prompting approaches rely on unidirectional, left-to-right inference mechanisms that generate constrained reasoning paths. This paradigm fundamentally fails to address two critical complexities: (1) the inherent branching logic embedded in question-answer structures, and (2) the oversight of respondent-specific conditions. Compounding these issues, selective avoidance behaviors further obscure the reasoning process, creating additional opacity in model explanations. Though novel model explanation methods, including model-specific and model-agnostic explanations such as SHAP (Lundberg and Lee 2017), LIME (Ribeiro, Singh, and Guestrin 2016), and DeepLIFT (Shrikumar, Greenside, and Kundaje 2017), have arisen to address the limited insight into underlying causal processes or theoretical implications, recent works (Kim et al. 2024b; Wu et al. 2025) have highlighted inconsistencies in instance-level explanations generated by different predictive agents. Taking the Fig. 1 as an example, LLMs A, B, and N are employed to reason about the same social problem and are expected to produce outcomes accompanied by intermediate reasoning steps. Nevertheless, the resulting conclusions may differ, even though they are based on the same questionnaire data. Inspired by some LLM-based multiple-agent systems fostering cooperative behavior within a group (de Curtò and de Zarzà 2025; Kim et al. 2024a), a Forest of Thought Explanation (FoTE) is proposed for the explanation consensus by integrating multiple reasoning paths with a novel forest of explanation prompting.

Our work delivers three key contributions:

1. The FoTE is designed to generate a consistent explanation across LLM reasoning paths for their outputs, especially for the problems with strong selective avoidance.
2. The adaptive Chain of Thought can capture respondents' primary considerations across multiple dimensions, such as family and work, and others.

*Corresponding Author.

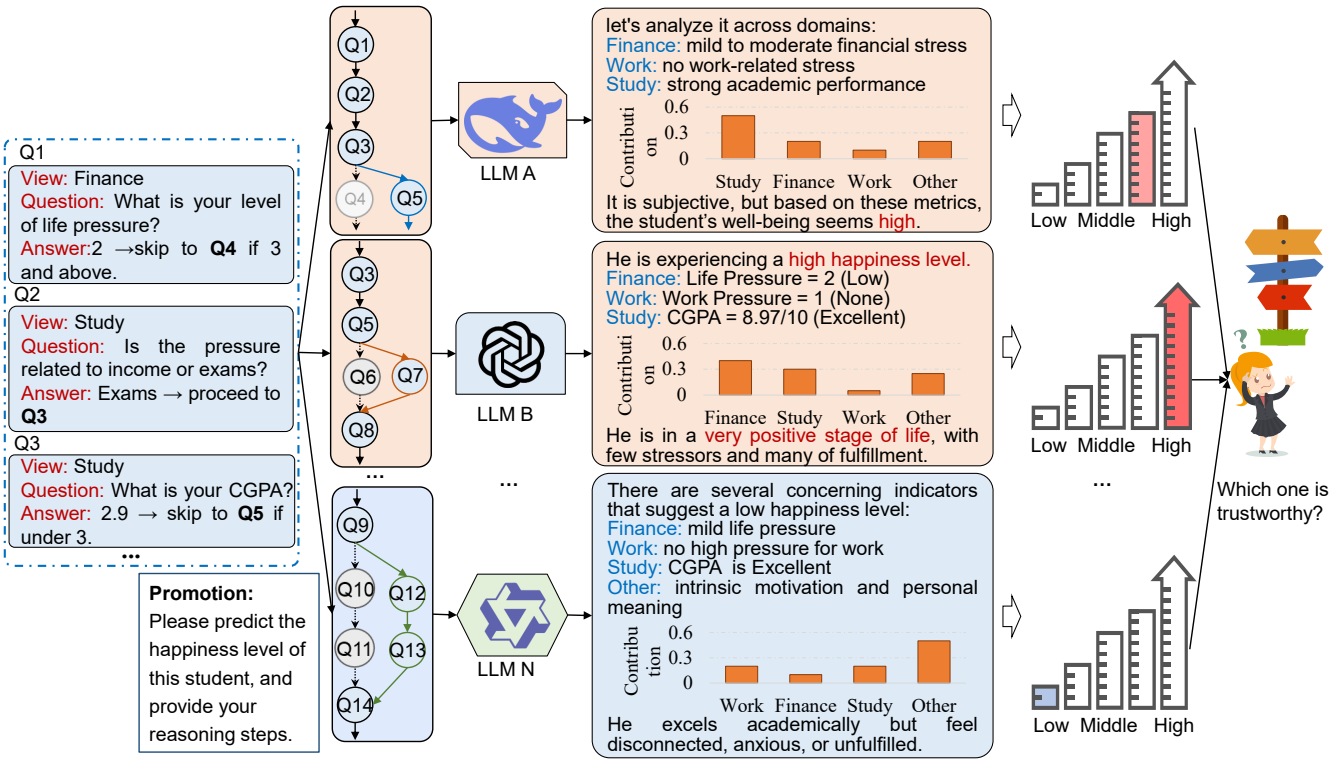

Figure 1: An example of reasoning inconsistency across different LLMs on the same survey. Based on the same questionnaire, selective avoidance can result in contradictory reasoning outputs and inconsistent explanations by LLMs, which poses challenges for reasoning trustworthiness.

3. An important scientific finding demonstrates the effectiveness of FoTE for real-world inference problems with strong selective avoidance. This discovery is confirmed by extensive experiments using three reasoning-capable LLMs and four real-world, community-recognized datasets, which cover two popular social problems and two data types. It achieved the best reasoning success rate of 0.84 and the best weighted F1 score of 0.83.

2 Related Work

2.1 LLM Prompting Reasoning

LLMs increasingly excel at a wide range of text and multimodal tasks—including classification, summarization, and complex reasoning—often achieving results that exceed human performance (Bang et al. 2023; Qin et al. 2023). Several socially oriented applications, such as misinformation identification (Choi and Ferrara 2024), sentiment classification (Xing 2025), hate speech recognition (Hong et al. 2024), and stance or humor analysis (Inácio and Oliveira 2024; Thapa et al. 2024), rely on combining diverse social cues and contextual signals. To strengthen LLM reasoning in these settings, researchers have introduced various prompting techniques, which largely fall into two groups: Input–Output (I/O) and X-of-Thoughts (XoT) prompting strategies.

Input-output (I/O) Prompting Among existing prompting methods, I/O prompting—mainly covering both zero-shot and few-shot methods (Wei et al. 2022a)—is the most widely used. By supplying natural-language task descriptions along with the desired form of the answer, these approaches help LLMs establish a clearer reasoning context, which in turn improves the quality of their inference (Hasan et al. 2024). Prior work (Brown et al. 2020; Ahuja et al. 2023) has consistently shown that few-shot examples typically yield better performance than zero-shot. These works indicate that well-designed prompting can enhance LLM reasoning performance.

X-of-Thought Prompting Chain-of-Thought (CoT) prompting (Wei et al. 2022b) was introduced to handle reasoning tasks that are difficult to formalize, especially when the mapping from a question x to an answer y is complex. By guiding models to generate intermediate steps, CoT strengthens the reasoning ability of LLMs and improves their inference on challenging tasks. This approach has been widely adopted in areas such as question answering (Rasool et al. 2023) and mathematical reasoning (Shi et al. 2023). Several extensions—including Self-Consistency CoT (SC-CoT) (Wang et al. 2023), Tree-of-Thoughts (ToT) (Yao et al. 2023), Tree of Uncertain Thoughts (TouT) (Mo and Xin 2024), and Forest of Thought (FoT) (Bi et al. 2025)—have further demonstrated strong results on tasks like mini crosswords and creative generation. These methods underscore the effectiveness of the X-of-Thoughts prompting framework.

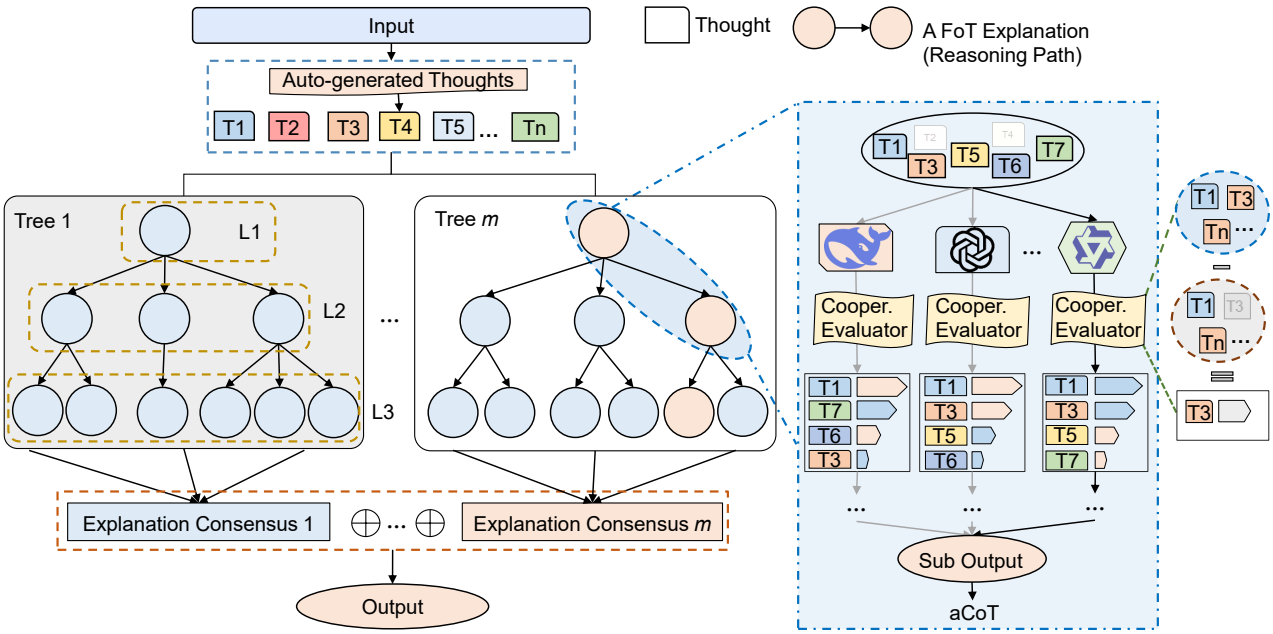

Figure 2: A FoTE framework is proposed for reaching an explanation consensus. Thought candidates are generated by LLMs and evaluated by a cooperative evaluator, then integrated into the FoTE reasoning process, which consists of m sub-trees. The pink tree nodes represent a path of the LLMs' reasoning, generating a corresponding explanation.

2.2 Explanation for LLMs

To improve the explanation of LLMs, the follow-up works proposed numerous prompting methods to generate natural language explanations for why a certain text could be hateful by Chain of Explanation (CoE) (Huang, Kwak, and An 2023), for dialogue state tracking by Chain-of-Thought-Explanatio (CoTE) (Xu et al. 2024b), and an explanation verification method named CoTEVer (Kim et al. 2023). However, these prompting methods are primarily aligned with left-to-right instruction mechanisms, which are insufficient for addressing social problems due to the absence of a selective avoidance strategy. As a result, they fail to account for individual differences among respondents.

In summary, existing research reveals two fundamental limitations in LLM reasoning: (1) the inability to effectively model selective avoidance behaviors prevalent in online UGC, primarily due to the constraints of unidirectional left-to-right inference architectures; and (2) the persistent challenge of generating transparent and reliable explanations for LLM reasoning processes.

3 Methodology

Recent studies (Mo and Xin 2024; Bi et al. 2025) indicate that uncertainty exploration can effectively navigate large combinatorial feature spaces to locate high-quality solutions. Building on this insight, a new prompting method named FoTE is introduced for selective avoidance in social reasoning tasks, as illustrated in Fig. 2. The core of this framework is the adaptive Chain of Thoughts (aCoT), designed to produce reasoning thoughts for each view by leveraging prompt-based learning to convert user-provided content into linguistic cues that guide the breakdown of intermediate reasoning steps. Distinct from prior contribution computing techniques—such as entropy measures, score-based ranking, or voting (Yao et al. 2023)—our approach evaluates thoughts through a cooperative contribution evaluator grounded in the Shapley value (Shapley 2016). FoTE then aggregates the full set of generated thoughts and reasoning paths, enabling the model to search the extensive thought space for the most effective reasoning strategy.

3.1 Problem Statement

Let a multi-turn interaction in social problems be represented as a view-annotated dialogue history

$$\mathcal{D} = \{(v_i, q_i, r_i)\}_{i=1}^{J},$$

where for the i-th turn, $v_i \in \mathcal{V}$ denotes the view (e.g., economics or family), $q_i \in \mathcal{Q}$ is the question, and $r_i \in \mathcal{R}$ is the response. To construct a sampled multi-turn dialogue for simulating the selective avoidance, we randomly select $\mathcal{D}^*$ containing m tuples from the dataset $\mathcal{D}$:

$$\mathcal{D}^* = \{(v_i, q_i, r_i)\}_{i=1}^{k}, \; where \; (v_i, q_i, r_i) \sim Uniform(\mathcal{D}).$$

An explanation generator $g : \mathcal{V} \times \mathcal{Q} \times \mathcal{R} \to \varepsilon^*$ produces a sequence of intermediate explanations $E_i = (e_1, e_2, \ldots, e_l)$ for a multi-turn interaction, following $E_i = g(v_i, q_i, r_i)$, where $e_j \in \varepsilon$ are explanation tokens. The explanation-augmented dialogue state $\mathcal{D}'$ is then constructed as:

$$\mathcal{D}' = \{(v_i, q_i, E_i, r_i)\}_{i=1}^{n} \subseteq \mathcal{V} \times \mathcal{Q} \times \varepsilon^* \times \mathcal{R},$$

where $n \leq k$. A reasoning forest $\mathcal{F}$ is a set of J tree-structured LLMs $\{T_j\}_{j=1}^{m}$. The final output is the aggregation of all leaf-node states.

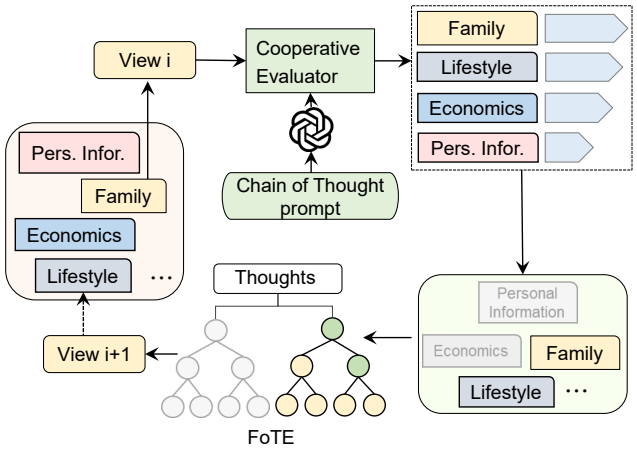

Figure 3: Hierarchical Multi-View Thought Generation Using the adaptive Chain of Thought.

3.2 Multi-View Thought Generation

Linguistic cues (Levitan, Maredia, and Hirschberg 2018; Xu et al. 2024a) provide insight into respondents on social analysis and refer to specific levels, which are beneficial to language models for the explanation of different views. The social phenomenon is influenced by various perspectives informed by social science knowledge. For example, positive emotions, engagement, relationships, meaning and purpose, and accomplishment are identified as primary factors in the mental state (Bognar 2010). Inspired by this knowledge, the aCoT is proposed to generate thoughts from different views, as shown in Fig. 3.

The layers of a language model capture different aspects of linguistic structure (van Aken et al. 2019). Therefore, we treat models at various layer depths as specialized learners representing distinct views, and their outputs are combined within a forest architecture to improve predictive accuracy. In our method, LLMs equipped with CoT prompting are employed to derive intermediate reasoning explanations from multiple perspectives associated with each view. The overall procedure can be expressed as a hierarchical, multi-level functional mapping, as defined below.

$$\mathcal{F}_{LLM} = (\mathcal{V}, \mathcal{Q}, \mathcal{R}) \rightarrow (V, K, R), \tag{1}$$

where $\mathcal{V}$ is the set of views, $\mathcal{Q}$ represents the questions, and $\mathcal{R}$ corresponds to their associated response. For a view–question–response tuple (v_i, q_i, r_i), the extraction function for generating thoughts at level L_i is by Eq. 2.

$$\begin{aligned} \mathcal{T}_{L_1} &= \mathcal{G}^{(L_1)}_{LLM}(v_i, q_i, r_i) \quad \text{(View-level thoughts)} \\ \mathcal{T}_{L_2} &= \mathcal{G}^{(L_2)}_{LLM}\mathcal{T}_{L_1} \quad \text{(Keywords-level thoughts)} \\ \mathcal{T}_{L_3} &= \mathcal{G}^{(L_3)}_{LLM}\mathcal{T}_{L_2} \quad \text{(Final reasoning outputs)} \end{aligned} \tag{2}$$

where $\mathcal{T}_{L_i}$ denotes the set of thoughts produced at level L_i by the thought generator. Specifically, $\mathcal{T}_{L_1}$, $\mathcal{T}_{L_2}$, and $\mathcal{T}_{L_3}$ correspond to the view-level outputs, emotion-keyword outputs, and final generated responses, respectively. The thoughts derived at each layer of the hierarchy collectively form the pool of candidate components.

3.3 Forest of Thought Explanation

Each thought generated by aCoT is evaluated using a cooperative thought evaluator and is constructed as a sub-tree. The sub-tree is then joined into a forest of explanations for final reasoning. As shown in Fig. 2, the FoTE is defined as a multi-reasoning path with a selection mechanism and consensus-driven explanation represented by a pink node path. The key improvements include the traceable reasoning steps and capturing selective avoidance.

Synergistic Contribution Evaluation Drawing on principles from cooperative game theory (Shapley 2016), we introduce a cooperative thought evaluator designed to identify the top-k most influential thoughts produced by aCoT, as also illustrated on the right side of Fig. 2. This evaluator ensures fair attribution of contribution within a group of thoughts and adheres to the accuracy, missingness, and consistency criteria of feature attribution (Shapley 2016). In contrast to methods such as LIME (Ribeiro, Singh, and Guestrin 2016) and DeepLIFT (Shrikumar, Greenside, and Kundaje 2017), which approximate importance through local perturbations or gradient-based decompositions, the cooperation-based evaluator directly models cooperative interactions and synergistic contributions among thoughts, providing stronger theoretical grounding and improved reliability in quantitative assessment.

The reasoning process is viewed as a cooperative effort aimed at producing the correct outcome using all available thoughts. Accordingly, the contribution of a specific thought j is denoted by ϕ_j, defined as follows:

$$\begin{aligned} \phi_j(v) = &\textstyle\sum_{S \subseteq \mathcal{T}' \setminus \{\mathcal{T}'_j\}} \frac{|S|!(|\mathcal{T}'|-|S|-1)!}{|\mathcal{T}'|!} \\ &\left[v(S \cup \{\mathcal{T}'_j\}) - v(S)\right], \end{aligned} \tag{3}$$

where $\mathcal{T}' \subseteq \mathcal{T}$, $j = \{1, \cdots, |\mathcal{T}'|\}$, and $S \subseteq \mathcal{T}' \setminus \{\mathcal{T}'_j\}$ denotes all possible subsets of thought set $\mathcal{T}'$ that excludes the j-th thought. The size of this subset is given by $|S|$. The $|S|!(|\mathcal{T}'| - |S| - 1)!/|\mathcal{T}'|!$ represents the probability of sampling subset S. $v(S \cup \mathcal{T}'_j) - v(S)$ measures the marginal contribution of thought j, where $v(\cdot) \in \mathbb{R}$ denotes the model output conditioned on the thoughts contained in the given set. Thus, the contribution score is computed as a weighted average of these marginal differences across all subsets S that exclude thought j. Because the exact computation incurs exponential complexity, we employ a Monte Carlo–based approximation (Strumbelj and Kononenko 2014) to make the evaluation tractable.

Efficient Thought Selection To improve the efficiency of the reasoning process, we retain only the top-k thoughts with the highest contribution scores $\phi_j(v)$ from different views, i.e, $\sum_{\mathcal{T}'_j \in \mathcal{T}' \subseteq \mathcal{T}} \phi_j(\mathcal{T}'_j)$, rather than utilizing the full set of generated thoughts. $\mathcal{T}$ denotes the complete set of thoughts produced by aCoT; $\mathcal{T}' \subseteq \mathcal{T}$ refers to a subset of n thoughts sampled at random; and $|\mathcal{T}^*| = k$ represents the top-k subset containing the highest-valued thoughts.

Uncertainty-Aware Reasoning Modeling uncertainty for responses remains a key challenge in reasoning tasks. Existing methods, including FoT, ToT, and SC-CoT, depend

Algorithm 1: Generating Forest of Thought Explanation

Require: Input data $\mathcal{D}$, number of sub-trees m, and thought number $|\mathcal{T}|$

1: Draw a sample x of question–answer pairs from $\mathcal{D}$;
2: **for** each sampled instance x **do**
3: Apply aCoT to obtain the complete thought pool $\mathcal{T}$;
4: Select a subset $\mathcal{T}'$ from $\mathcal{T}$ via non-replacement sampling;
5: Determine the top-k candidate thoughts $\mathcal{T}^*$ from $\mathcal{T}'$;
6: **for** each level-wise thought set $\mathcal{T}_{L_i}$ within $\mathcal{T}^*$ **do**
7: Choose a root thought according to Eq. 5;
8: Use the cooperative evaluator to identify the most informative splitting thought;
9: Expand the left branch via DFS applied to $\mathcal{T}_i^{\text{left}}$;
10: Expand the right branch via DFS applied to $\mathcal{T}_i^{\text{right}}$;
11: **end for**
12: Combine all generated sub-trees into the FoTE;
13: **end for**
14: **Return** the completed Forest of Thought Explanation

on predetermined and deterministic reasoning trajectories. By contrast, our FoTE incorporates a reasoning path fusion strategy that randomly samples from the generated thoughts to create $|\mathcal{T}^*|$ approximate, independent, and identically distributed (i.i.d.) thought sets, capturing the effect of selective avoidance across individuals (Zhang et al. 2022). As formalized in Eq. 4, this uncertainty-aware reasoning process leverages the importance weighting function $\phi_{\mathcal{T}_{L_j}}(v)$ to select more informative thoughts preferentially.

$$P(\mathcal{T}_{L_j^i} \in \mathcal{T}^*) \propto \phi_{\mathcal{T}_{L_j^i}}(v) \tag{4}$$

Using the selected thought subset $\mathcal{T}_{L_j}$, multiple trees $\{T_1, T_2, \cdots, T_m\}$ could be generated through randomized sampling. Initially, the root nodes $R = \{T_{r_1}, T_{r_2}, \cdots, T_{r_m}\}$ are chosen according to Eq. 5, where each root corresponds to the thought with the highest contribution score among its candidates, ensuring that the most informative thoughts guide the initial branching of the forest.

$$P(T_{r_j}\ is\ root) \propto \phi_j(v) \tag{5}$$

where the $T_{r_j} \in \mathcal{T}^*$ is the candidate root of each sub-tree.

Different search strategies can be integrated into the prompting method depending on the structure of the constructed RoTE. In this work, we adopt an effective tree-expansion procedure and leave the incorporation of more sophisticated algorithms for future study. Depth-First Search (DFS) is often favored over Breadth-First Search (BFS) in various computational settings due to its lower memory requirements, strong performance in deep search spaces, and natural compatibility with recursive formulations (Tarjan and Zwick 2024). The DFS-based construction process for FoTE is outlined in Algorithm 1 to provide a clear view of the computation pipeline. Step 1 samples a subset of question–answer pairs across different views to form the first level of selective avoidance. Steps 2–13 iteratively grow the random forest of thoughts. In detail, Step 4 draws a subset $\mathcal{T}'$ from $\mathcal{T}$ without replacement; Step 5 evaluates each thought and retains the top-k set $\mathcal{T}^*$. Step 7 determines the root of each tree using Eq. 5; Step 8 identifies the optimal splitting thought—the one with the highest cooperative contribution score. Steps 9–10 recursively generate the left and right subtrees. Finally, Step 12 aggregates all constructed subtrees into the complete FoTE prompting architecture.

4 Experiment

4.1 Experiment Setup

Base LLMs We employ three open-source LLMs as base models to evaluate FoTE: Qwen2.5-7B[1], Llama3-8B[2] (at Meta 2024), and the recent state-of-the-art DeepSeek-R1-8B[3] (DeepSeek-AI et al. 2025). DeepSeek-R1 is distilled from Qwen and Llama models and exhibits strong reasoning capabilities. We test all three LLMs on a variety of widely used prompting benchmarks, including standard zero-shot I/O, CoT, SC-CoT, ToT, FoT, and FoTE prompting.

Dataset The prompting method is evaluated on two representative social science problems characterized by strong selective avoidance: happiness prediction and depression detection (Yang et al. 2022). These tasks span two major categories of contemporary social science analysis—surveys and social media datasets—thus providing comprehensive coverage of mainstream application settings.

- ***Survey Data***. Two open, shared, and large-scale social online investigation datasets are employed in this work, such as 1) Social Survey containing the Chinese General Social Survey[4] and the European citizens survey dataset[5], as well as 2) Student Survey[6].
- ***Social Media Data***. Depression Post[7] is a multi-turn dialogue in online comments and can be used for depression detection. Happy Moments (Asai et al. 2018) is a large-scale collection of happy moments over three months on Amazon Mechanical Turk, containing more than 100,000 crowd-sourced happy moments in a specific area.

Metrics The success rate and weighted F1 score, as used in (Wei et al. 2022b; Yao et al. 2023), are employed to evaluate the reasoning performance of the LLMs.

Parameter Settings and Environment Following prior work (Wei et al. 2022b; Wang et al. 2023; Yao et al. 2023), each prompt is executed 100 times on randomly sampled inputs, and the mean performance is reported. FoTE uses a maximum depth of 3, a maximum thought count of $\sqrt{N}$ where $N = |\mathcal{T}'|$, and a minimum split size of 2. The number of trees is selected from $\{2^x \mid x = 1, \cdots, 7\}$. Experiments are performed in Python 3.12 and PyTorch 2.3.0 on a standard server with dual RTX 3080 (20GB) GPUs.

[1] https://ollama.com/library/qwen2.5:7b
[2] https://ollama.com/library/llama3
[3] https://ollama.com/library/deepseek-r1
[4] http://cgss.ruc.edu.cn/
[5] https://ess-search.nsd.no/
[6] https://www.kaggle.com/datasets/hopesb/
[7] https://www.kaggle.com/code/isanbel/depression-on-twitter

LLMs	Prompting	Social Survey		Student Survey		Happy Moment		Depression Post	
		Success	Weighted-F1	Success	Weighted-F1	Success	Weighted-F1	Success	Weighted-F1
Qwen2.5-7B	I/O Prompt	0.20	0.19	0.58	0.57	0.47	0.47	0.34	0.32
	CoT	0.45	0.43	0.67	0.66	0.64	0.62	0.41	0.37
	SC-CoT	0.49	0.48	0.79	0.78	0.65	0.65	0.56	0.54
	ToT	0.53	0.51	0.80	0.80	0.69	0.66	0.66	0.63
	FoT	0.56	0.55	0.82	0.81	0.71	0.70	0.69	0.68
	FoTE(Ours)	**0.60**	**0.57**	**0.84**	**0.83**	**0.73**	**0.72**	**0.73**	**0.70**
Llama3-8B	I/O Prompt	0.29	0.27	0.54	0.49	0.44	0.43	0.51	0.44
	CoT	0.51	0.47	0.61	0.59	0.59	0.60	0.75	0.74
	SC-CoT	0.62	0.61	0.65	0.62	0.70	0.69	0.75	0.75
	ToT	0.64	0.62	0.81	0.80	0.72	0.71	0.76	0.74
	FoT	0.67	0.66	0.81	0.81	0.73	0.72	0.79	0.77
	FoTE(Ours)	**0.72**	**0.71**	**0.84**	**0.83**	**0.75**	**0.74**	**0.81**	**0.79**
DeepSeek-R1-8B	I/O Prompt	0.21	0.19	0.56	0.54	0.49	0.50	0.49	0.45
	CoT	0.36	0.34	0.62	0.59	0.54	0.51	0.66	0.59
	SC-CoT	0.45	0.43	0.61	0.59	0.59	0.58	0.67	0.57
	ToT	0.50	0.49	0.64	0.61	0.61	0.59	0.77	0.72
	FoT	0.53	0.52	0.66	0.61	0.65	0.62	0.79	0.77
	FoTE(Ours)	**0.57**	**0.54**	**0.68**	**0.62**	**0.74**	**0.72**	**0.81**	**0.78**

Table 1: Reasoning results on two social problems (happiness and depression prediction) based on three open-source LLMs covering four datasets. The bold highlights the best performance, whereas the underlined indicates the second-best results.

4.2 Results and Discussion

Results on Survey Table 1 illustrates that FoTE achieves superior performance across the two survey datasets when evaluated with Qwen2.5-7B, Llama3-8B, and DeepSeek-R1-8B. Methods relying solely on direct I/O prompting, which omit intermediate reasoning, show the weakest results. This trend reinforces the well-established finding that incorporating explicit reasoning traces improves LLM reasoning quality (Wei et al. 2022b; Yao et al. 2023; Wang et al. 2023). Across all evaluated datasets and model architectures, FoTE demonstrates clear advantages by combining diverse thought processes through an information fusion procedure that explores multiple reasoning configurations. This mechanism aligns naturally with the selective avoidance phenomena commonly found in survey-based responses (see Fig. 1). Unlike traditional XoT promptings that adhere to a fixed sequential reasoning pattern, FoTE dynamically navigates the thought space, amplifying high-value reasoning components while mitigating distractions from thoughts irrelevant to mental state inference.

Results on Social-Media Referring to Table 1, FoTE demonstrates substantial gains over strong baseline methods. On the happy moments dataset with DeepSeek-R1-8B, it improves the success rate by up to 13% compared to ToT and 9% relative to FoT. These results highlight FoTE's capability to effectively tackle social science tasks characterized by complex selective avoidance patterns.

A closer examination of the social survey and social media datasets reveals that the performance gains are particularly pronounced on survey-based tasks. For instance, when using LLaMA-3-8B, FoTE achieves a peak success rate of 0.72 on the social survey dataset, exceeding FoT by 5%. This indicates that FoTE's fusion of diverse reasoning paths and selective emphasis on informative thoughts is especially beneficial for structured survey data.

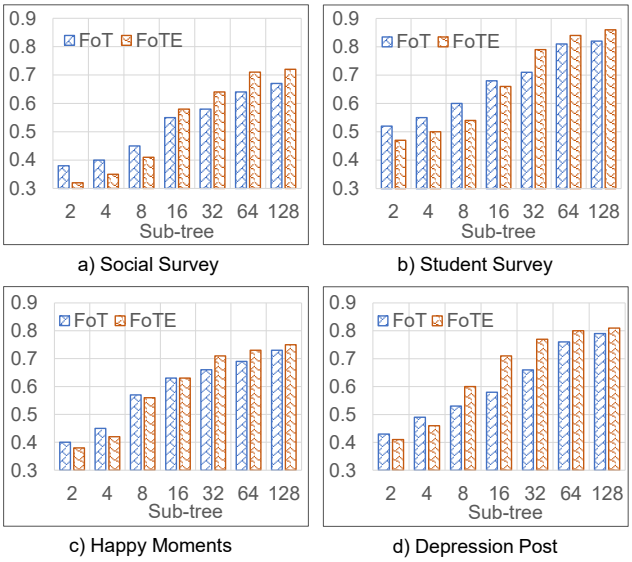

Figure 4: Sensitivity analysis for the number of sub-trees on the success rate of FoT and FoTE using LLaMA-3-8B for survey and social media tasks.

4.3 Sub-tree Sensitivity

A comprehensive parameter sensitivity analysis focusing on the number of activated reasoning trees is conducted in this section. For each parameter configuration, we calculated the mean success rate across randomized samples.

We evaluated the number of trees T within a reasonable range while holding others fixed. Figure 4 illustrates the reasoning performance on the social survey and student survey

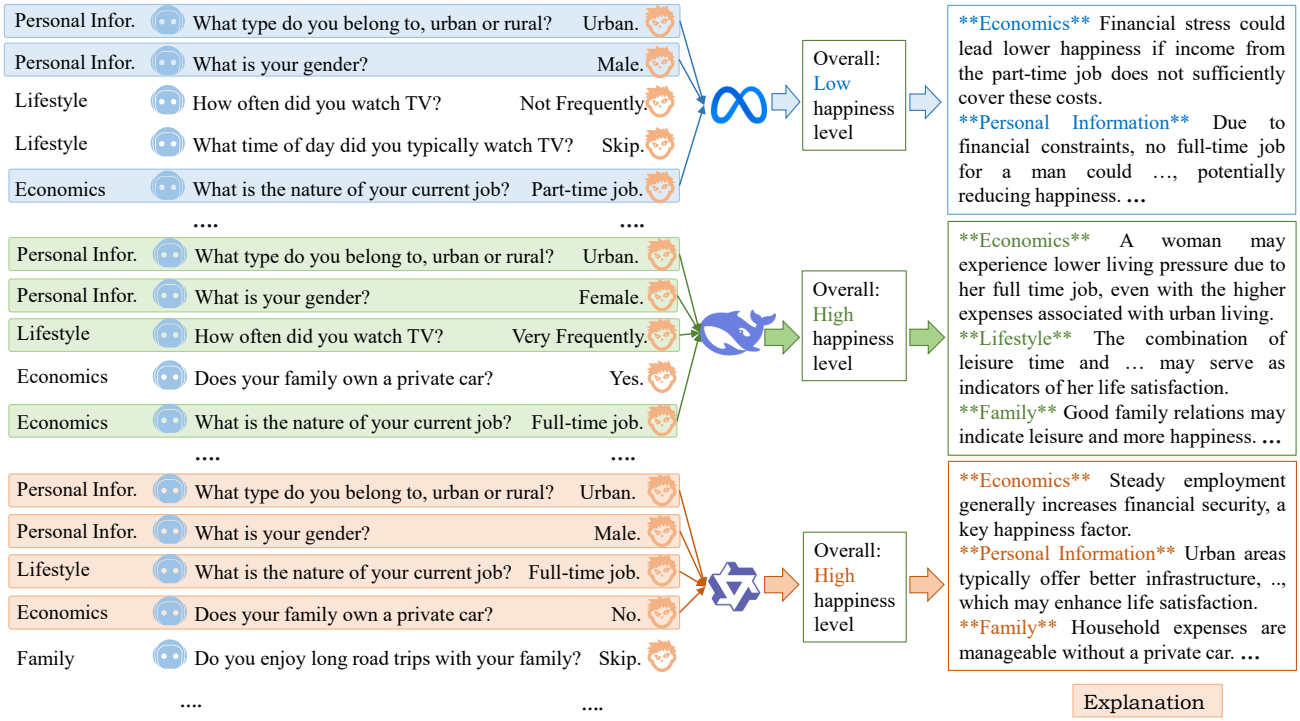

Figure 5: The case study of why the LLMs generate their outputs on the social survey data. The colored background indicates the selection of a question–answer pair, labeled by category, which is then used to generate the explanation.

datasets, revealing two notable patterns. First, employing 64 trees provides a marked improvement by broadening the exploration of the solution space, whereas increasing the number of trees beyond this point yields only marginal gains. Second, the peak success rates for social and student surveys reach 0.84, surpassing the 0.8 peak observed on depression posts, highlighting FoTE's ability to capture the selective-avoidance behavior prevalent in survey responses. Both datasets exhibit performance plateaus beyond these points, indicating an intrinsic limitation in reasoning capacity. Despite this, FoTE maintains stability across varying numbers of sub-trees. Increasing the tree count further results in minimal performance improvement while incurring higher computational cost.

Overall, these findings demonstrate that FoTE is robust to moderate hyperparameter variations, reinforcing its practical applicability in real-world scenarios.

4.4 Case Study

Explanation evaluation remains a significant challenge to date (Kim et al. 2023). To address this, prior studies have employed different forms of human feedback—such as assessments of informativeness, clarity (Huang, Kwak, and An 2023), and similarity to human-generated explanations (Chen et al. 2025)—typically through simple case studies and visualization. Building on prior works, we conducted a case study to assess the effectiveness of the proposed FoTE prompting approach in improving the explainability of LLM reasoning outputs. As illustrated in Fig. 5, it recombines question-answer pairs across different demographic categories to simulate diverse respondent profiles. The analysis revealed that LLMs generated transparent reasoning explanations when processing these structured inputs. For instance, one respondent profile - an urban-dwelling male without full-time employment - demonstrated how financial constraints negatively impact happiness levels. The FoTE framework automatically adapts to this profile by skipping irrelevant questions (e.g., television viewing habits for respondents who rarely watch TV) while maintaining coherent reasoning chains. This approach not only improved the logical consistency of LLM outputs but also provided verifiable intermediate reasoning steps. The resulting explanations significantly enhanced model transparency, suggesting broad applicability in fields requiring interpretable AI decision-making, from social science research to policy analysis.

5 Conclusion and Future Work

This work proposes FoTE that enhances reasoning performance in domain-specific applications by diversifying reasoning path exploration. It employs a cooperative thought evaluator for more reliable thought selection and produces transparent reasoning outputs, and is effective for complex, socially relevant tasks involving selective avoidance. However, FoTE relies on predefined representative samples with selective-avoidance patterns to optimize prompting. Future research will focus on developing more efficient in-context learning methods for sample contribution computing and selection, further improving its explanation.

Acknowledgments

This work is partially supported by the National Science Foundation of China (No.62276196) and the China Scholarship Council program (No.202306950097).

References

Ahuja, K.; Diddee, H.; Hada, R.; Ochieng, M.; and et al. 2023. MEGA: Multilingual Evaluation of Generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, 4232–4267.

Asai, A.; Evensen, S.; Golshan, B.; Halevy, A. Y.; and et. al. 2018. HappyDB: A Corpus of 100, 000 Crowdsourced Happy Moments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*.

at Meta, A. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date.

Bang, Y.; Cahyawijaya, S.; Lee, N.; Dai, W.; and et al. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. In *Proceedings of 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 2023 -Volume 1: Long Papers, Nusa Dua, Bali, November 1 - 4, 2023*, 675–718.

Bi, Z.; Han, K.; Liu, C.; Tang, Y.; and Wang, Y. 2025. Forest-of-Thought: Scaling Test-Time Compute for Enhancing LLM Reasoning. *arXiv preprint arXiv:2412.09078*.

Bognar, G. 2010. Authentic happiness. *Utilitas*, 22(3): 272–284.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; and et al. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 1–25.

Chen, B.; Peng, S.; Korhonen, A.; and Plank, B. 2025. A Rose by Any Other Name: LLM-Generated Explanations Are Good Proxies for Human Explanations to Collect Label Distributions on NLI. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, 10777–10802.

Choi, E.; and Ferrara, E. 2024. FACT-GPT: Fact-Checking Augmentation via Claim Matching with LLMs. In *Companion Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, Singapore, May 13-17, 2024*, 883–886.

de Curtò, J.; and de Zarzà, I. 2025. LLM-Driven Social Influence for Cooperative Behavior in Multi-Agent Systems. *IEEE Access*, 13: 44330–44342.

DeepSeek-AI; Guo, D.; Yang, D.; Zhang, H.; Song, J.; and et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *CoRR*, abs/2501.12948.

Hasan, M. A.; Das, S.; Anjum, A.; Alam, F.; Anjum, A.; Sarker, A.; and Noori, S. R. H. 2024. Zero- and Few-Shot Prompting with LLMs: A Comparative Study with Fine-tuned Models for Bangla Sentiment Analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, 17808–17818.

Hong, L.; Luo, P.; Blanco, E.; and Song, X. 2024. Outcome-Constrained Large Language Models for Countering Hate Speech. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, 4523–4536.

Huang, F.; Kwak, H.; and An, J. 2023. Chain of Explanation: New Prompting Method to Generate Quality Natural Language Explanation for Implicit Hate Speech. In *Companion Proceedings of the ACM Web Conference 2023*, 90–93. New York, NY, USA.

Inácio, M. L.; and Oliveira, H. G. 2024. Exploring Multimodal Models for Humor Recognition in Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese, PROPOR 2024, Santiago de Compostela, Galicia/Spain, March 12-15, 2024, Volume 1*, 568–574.

Kim, H.; Han, B.; Kim, J.; Lubis, M. F. S.; Kim, G. J.; and Hwang, J.-I. 2024a. Engaged and Affective Virtual Agents: Their Impact on Social Presence, Trustworthiness, and Decision-Making in the Group Discussion. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24.

Kim, H.; Park, J.; Choi, Y.; Lee, S.; and Lee, J. 2024b. BayesNAM: Leveraging Inconsistency for Reliable Explanations. *arXiv preprint arXiv:2411.06367*.

Kim, S.; Joo, S. J.; Jang, Y.; Chae, H.; and Yeo, J. 2023. CoTEVer: Chain of Thought Prompting Annotation Toolkit for Explanation Verification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 195–208.

Levitan, S. I.; Maredia, A.; and Hirschberg, J. 2018. Linguistic Cues to Deception and Perceived Deception in Interview Dialogues. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1941–1950.

Liang, Y.; Ju, Y.; Zeng, X.-J.; Li, H.; Dong, P.; and Ju, T. 2025. A user-generated content-based social network large-scale group decision-making approach in healthcare service: Case study of general practitioners selection in UK. *Expert Systems with Applications*, 261: 125542.

Liu, X.; Zhang, J.; Shang, H.; Guo, S.; Yang, C.; and Zhu, Q. 2025. Exploring Prosocial Irrationality for LLM Agents: A Social Cognition View. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan,

S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30*, 4765–4774.

Mo, S.; and Xin, M. 2024. Tree of Uncertain Thoughts Reasoning for Large Language Models. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, 12742–12746.

Naeem, M. 2020. Uncovering the role of social motivational factors as a tool for enhancing brand-related content. *Qualitative Market Research: An International Journal*, 23(2): 287–307.

Qin, C.; Zhang, A.; Zhang, Z.; Chen, J.; Yasunaga, M.; and Yang, D. 2023. Is ChatGPT a General-Purpose Natural Language Processing Task Solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, 1339–1384.

Rasool, Z.; Barnett, S.; Kurniawan, S.; and et al. 2023. Evaluating LLMs on Document-Based QA: Exact Answer Selection and Numerical Extraction using Cogtale dataset. *CoRR*, abs/2311.07878.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Shapley, L. S. 2016. *A Value for n-Person Games*, volume 2, 307–318.

Shi, F.; Suzgun, M.; Freitag, M.; Wang, X.; and et al. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2695–2709.

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning important features through propagating activation differences. In *International conference on machine learning*, 3145–3153.

Strumbelj, E.; and Kononenko, I. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.*, 41(3): 647–665.

Tarjan, R. E.; and Zwick, U. 2024. Finding strong components using depth-first search. *European Journal of Combinatorics*, 119: 103815.

Thapa, S.; Rauniyar, K.; Jafri, F.; and et al. 2024. Stance and Hate Event Detection in Tweets Related to Climate Activism - Shared Task at CASE 2024. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text, CASE 2024, St. Julians, Malta, March 22, 2024*, 234–247.

van Aken, B.; Winter, B.; Löser, A.; and Gers, F. A. 2019. How Does BERT Answer Questions?: A Layer-Wise Analysis of Transformer Representations. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, 1823–1832.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q. V.; and et al. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5*.

Wei, J.; Bosma, M.; Zhao, V. Y.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2022a. Finetuned Language Models are Zero-Shot Learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2022b. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 24824–24837.

Wu, X.; Li, L.; Tao, X.; Yuan, J.; and Xie, H. 2025. Towards the Explanation Consistency of Citizen Groups in Happiness Prediction via Factor Decorrelation. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 9(2): 1392–1405.

Xing, F. 2025. Designing Heterogeneous LLM Agents for Financial Sentiment Analysis. *ACM Trans. Manage. Inf. Syst.*, 16(1): 1 – 24.

Xu, B.; Li, L.; Luo, W.; Naseriparsa, M.; Zhao, Z.; Lin, H.; and Xia, F. 2024a. Beyond Linguistic Cues: Fine-grained Conversational Emotion Recognition via Belief-Desire Modelling. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2318–2328.

Xu, L.; Peng, N.; Zhou, D.; Ng, S.; and Fu, J. 2024b. Chain of Thought Explanation for Dialogue State Tracking. *CoRR*, abs/2403.04656.

Yang, L.; Li, S.; Luo, X.; Xu, B.; Geng, Y.; Zeng, Z.; Zhang, F.; and Lin, H. 2022. Computational personality: a survey. *Soft Computing*, 26: 9587–9605.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.; Cao, Y.; and Narasimhan, K. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 11809–11822.

Zhang, J.; Wang, Y.; Sun, M.; and Zhang, N. 2022. Two-Stage Bootstrap Sampling for Probabilistic Load Forecasting. *IEEE Transactions on Engineering Management*, 69(3): 720–728.