Fuzzy Reasoning Chain (FRC): An Innovative Reasoning Framework from Fuzziness to Clarity

Ping Chen[1,2†] **Xiang Liu**[1,2†] **Zhaoxiang Liu**[1,2*] **Zezhou Chen**[1,2] **Xingpeng Zhang**[3]
Huan Hu[1,2] **Zipeng Wang**[1,2] **Kai Wang**[1,2] **Shuming Shi**[1,2] **Shiguo Lian**[1,2*]

[1] Data Science & AI Research Institute, China Unicom
[2] Unicom Data Intelligence, China Unicom
[3] School of Computer Science, Southwest Petroleum University
{chenp181, liux750, liuzx178, liansg}@chinaunicom.cn, xpzhang@swpu.edu.cn

Abstract

With the rapid advancement of large language models (LLMs), natural language processing (NLP) has achieved remarkable progress. Nonetheless, significant challenges remain in handling texts with ambiguity, polysemy, or uncertainty. We introduce the Fuzzy Reasoning Chain (FRC) framework, which integrates LLM semantic priors with continuous fuzzy membership degrees, creating an explicit interaction between probability-based reasoning and fuzzy membership reasoning. This transition allows ambiguous inputs to be gradually transformed into clear and interpretable decisions while capturing conflicting or uncertain signals that traditional probability-based methods cannot. We validate FRC on sentiment analysis tasks, where both theoretical analysis and empirical results show that it ensures stable reasoning and facilitates knowledge transfer across different model scales. These findings indicate that FRC provides a general mechanism for managing subtle and ambiguous expressions with improved interpretability and robustness.

1 Introduction

With the rapid advancement of large-scale language models (LLMs) (Guo et al., 2025; Devlin et al., 2019), natural language processing (NLP) has achieved remarkable progress across various domains. However, reasoning with texts that exhibit ambiguity, polysemy, and uncertainty remains a significant challenge. In tasks such as sentiment analysis, ethical judgment, and security review, texts often contain complex emotions, such as "***Though dissatisfied, still acceptable.***", which convey both negative sentiment and a degree of acceptance. This complexity makes traditional approaches, based on fixed labels or static rules, inadequate for capturing the multidimensional semantics (Pang et al., 2008; Bradley and Lang, 1994).

Current *fuzzy reasoning* approaches quantify ambiguity through membership degrees; however, their reliance on manually defined rules limits their adaptability to dynamic and evolving contexts (Taboada et al., 2011). Although *Chain-of-Thought* (CoT) reasoning enhances transparency, its discrete decision-making process often struggles with complex texts, such as those containing sarcasm and contradictions (Fei et al., 2023; Wei et al., 2022; Wang et al., 2023). Consequently, both fuzzy reasoning and CoT methods face limitations when addressing fuzzy and uncertain texts, underscoring the need for a more adaptive, stable, and transparent reasoning framework.

In this paper, we propose the *Fuzzy Reasoning Chain* (FRC) framework to address challenges arising from ambiguous and uncertain text, as illustrated in Fig. 1. FRC follows the standard step-by-step reasoning procedure typical of sentiment analysis and extends conventional chain-of-thought approaches by replacing discrete probability assignments with continuous fuzzy membership degrees. This transition from probability-based reasoning to fuzzy membership-based reasoning, which we refer to as the probability-to-membership collision, is the core methodological innovation. By capturing conflicting and ambiguous signals that probabilities alone cannot express, FRC enables a more nuanced and robust representation of sentiment while naturally identifying "Other" or unspecified categories. This fuzzy-to-clear reasoning paradigm ensures stability and transparency. We analyze the convergence behavior of FRC and corroborate its practical utility through comprehensive experiments, demonstrating its effectiveness on complex, fuzzy texts and potential for reasoning transfer across model scales.

The main contributions of FRC are as follows:

- **Core Methodological Innovation:**FRC integrates discrete probability reasoning with continuous fuzzy membership, enabling clearer interpretation of ambiguous inputs and capturing conflicting or uncertain signals that traditional probability-based methods cannot. This makes it the first framework to combine fuzzy membership with LLM reasoning in this way.
- **Convergence Analysis:** Ensures stable reasoning by analyzing robustness, monotonicity, and semantic completeness, supported by both theoretical analysis and experimental validation.
- **Empirical Validation and Generalization:** Experiments demonstrate that FRC enhances model

*Corresponding authors.
†Equal contribution.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 10230–10240

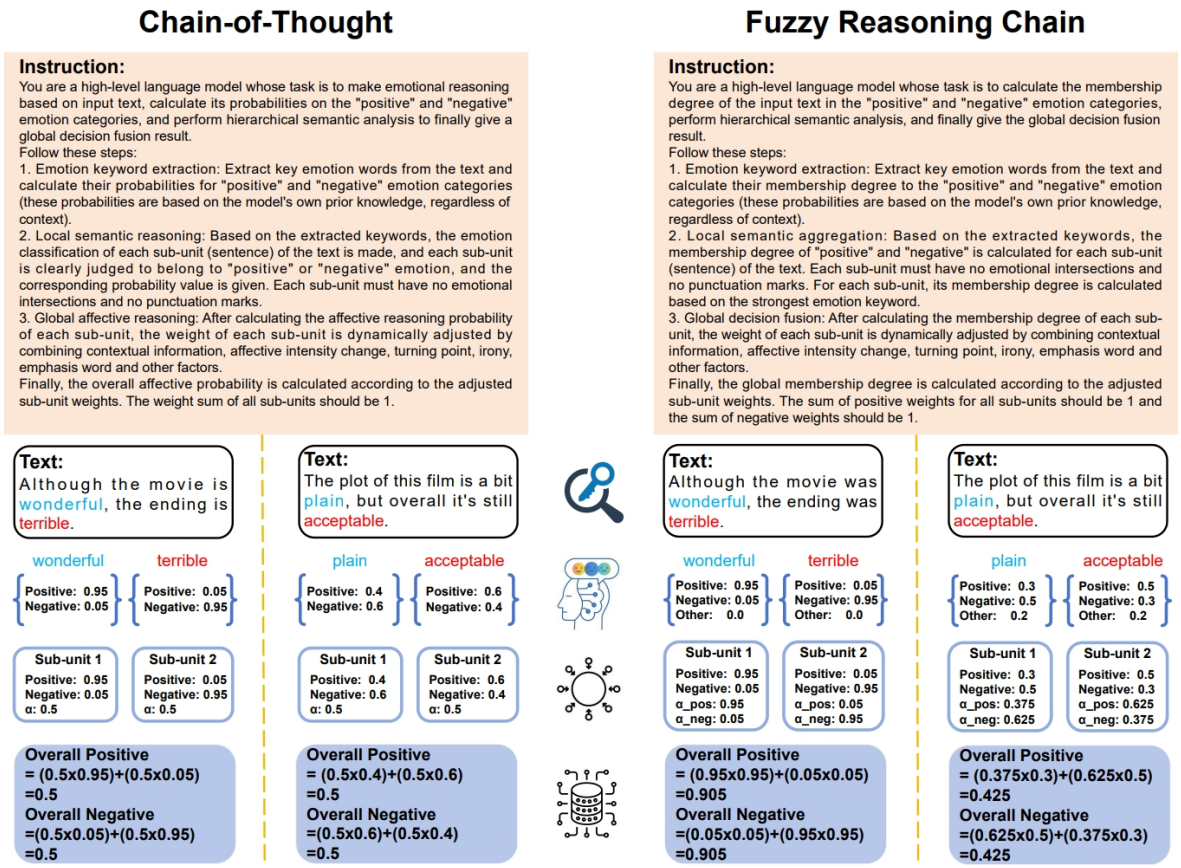

Figure 1: **Comparison of Chain-of-Thought (CoT, left) and Fuzzy Reasoning Chains (FRC, right) in sentiment analysis**. Both follow the same reasoning steps for fairness, reflecting typical sentiment analysis procedures and highlighting the interaction between probability-based reasoning and fuzzy membership. Unlike CoT, FRC membership degrees are unconstrained and do not need to sum to one, which allows any sentiment not explicitly covered in the prompt to be assigned to an "Other" category, capturing ambiguity and uncertainty that probability-based CoT cannot represent. FRC refines sentiment weights using fuzzy membership. In the first example, CoT shows balanced probabilities (0.5 each), while FRC captures strong conflicting sentiment (0.905 each). In the second example, CoT remains neutral (0.5 each), whereas FRC indicates a fuzzy sentiment state (0.425 each), enabling a transition from fuzzy to clear sentiment analysis.

performance on uncertain and ambiguous inputs, reflecting its generalization capacity and potential in lightweight knowledge transfer scenarios.

2 Related Work

2.1 Sentiment Analysis

Sentiment analysis is a fundamental task in NLP that aims to identify and classify the sentiment polarity of text (Pang et al., 2008). Research has explored sentiment dictionaries, machine learning, deep learning, and large model approaches. The use of emotion dictionaries enhances the original English (Whissell, 2009) and Chinese emotion dictionary, thereby increasing its applicability to natural language samples. Support Vector Machines (SVM), Bayesian methods, ridge regression, and other classical machine learning techniques are also employed in natural language analysis (Jemai et al., 2021; Hourrane et al., 2019). After the emergence of deep learning, it has become widely utilized in natural language sentiment analysis (Zhang et al., 2018; Tang et al., 2015; Felbo et al., 2017). Numerous methods employing Long Short-Term Memory (LSTM) networks and attention mechanisms have been proposed in this field. Regularized LSTM (Qian et al., 2017) and Sentiment-Aware Bidirectional LSTM (SAB-LSTM) (Kumar and Chinnalagu, 2020) integrate linguistic resources, such as sentiment dictionaries, negative words, and intensity words, into the LSTM framework to more accurately capture emotional nuances in sentences. Attention mechanisms effectively capture the significance of each contextual word in relation to a specific target aspect (Lei et al., 2016; Tiwari and Nagpal, 2022). The BERT (Devlin et al., 2019) model is a groundbreaking advancement in machine reading comprehension and can be applied across various fields, including sentiment analysis.

In recent years, large language models (LLMs) have propelled advancements in unsupervised and self-supervised sentiment analysis, significantly enhancing

performance on sentiment classification tasks (Devlin et al., 2019; Yinhan et al., 2019; Chen and Xiao, 2024; Zhang et al., 2024a). The literature (Zhang et al., 2024b) emphasizes the importance of contextual information in enhancing the sentiment estimation of LLMs. Deep prompt tuning and low-rank adaptation effectively enhance the performance of large models in sentiment analysis (Peng et al., 2024). Chain-of-Thought (CoT) are an effective strategy for enhancing the performance of LLMs and are also utilized for emotion classification (Fei et al., 2023; Duan and Wang, 2024). In addition to CoT reasoning that relies solely on textual information (Fei et al., 2023), audio and visual modalities are also taken into account to propose a cross-modal filtering and fusion (CMFF) module, which facilitates multi-modal sentiment analysis (Li et al., 2024). The generative CoT strategy is employed for zero-shot and few-shot emotion classification tasks (Gu et al., 2024; Wu et al., 2025; Liu et al., 2024b). However, these LLMs still predominantly depend on static labeling systems, making it difficult to handle complex sentiment conflicts and mixed emotions in texts effectively.

2.2 Fuzzy Reasoning

These sentiment analysis methods rely on discrete classification labels such as "positive", "negative", or "neutral", which struggle to address ambiguity, uncertainty, and mixed sentiment expressions (Bing, 2012). To overcome these limitations, fuzzy reasoning approaches have been explored. Multi-dimensional sentiment models have been employed to capture the complex emotional components of text (Bradley and Lang, 1994). Fuzzy clustering simulates the continuous nature of sentiment distributions, providing a more nuanced approach to sentiment analysis (Peizhuang, 1983). Lexicon-based sentiment classifiers have been enhanced with fuzzy logic to create sentiment distributions more accurately reflect human cognitive processes in uncertain cases (Taboada et al., 2011). There are also scholars who integrate the learning capabilities of deep learning with the uncertainty processing abilities of fuzzy logic to offer users more accurate emotional predictions (Chaturvedi et al., 2019; Sweidan et al., 2022; Do et al., 2024; Wang et al., 2025). The emotion enhancement inference model integrates word embedding, an emotion dictionary, and fuzzy inference (Yan et al., 2022). However, mainstream approaches still struggle with transparency in the reasoning process and lack dynamic adjustment of sentiment labels during inference.

2.3 Chain-of-Thought (CoT)

CoT reasoning has emerged as a technique to enhance the interpretability and performance of LLMs by facilitating step-by-step reasoning (Wei et al., 2022). This method has proven effective in alleviating data bottlenecks and improving model performance on complex tasks, such as mathematical reasoning, commonsense reasoning, and planning (Wei et al., 2022; Wang et al., 2023). CoT allows models to decompose problems into logical intermediate steps, providing greater interpretability compared to traditional end-to-end models (Santoro et al., 2017; Xu et al., 2024). Several techniques have been proposed to enhance CoT reasoning, including few-shot CoT (Wei et al., 2022), self-consistency (Wang et al., 2023), and Auto-CoT (Zhang et al., 2023). Furthermore, strategies like least-to-most prompting (Zhou et al., 2023), AutoHint (Sun et al., 2023), and result entropy optimization (Wan et al., 2023) have been developed to simplify problem-solving and enhance reasoning accuracy. Specialized approaches like MathPrompter (Imani et al., 2023) for mathematical problem-solving and Meta-prompting (Suzgun and Kalai, 2024) for a refined prompt generation have further expanded the capabilities of CoT.

Recent research, such as THOR (Fei et al., 2023), introduced a CoT reasoning approach to enhance the accuracy and interpretability of sentiment inference through step-by-step reasoning. However, within the realms of fuzzy reasoning and sentiment analysis, the application of CoT is still in its nascent stages. Recent investigations into Zero-Shot-CoT (Jin et al., 2024) have demonstrated that large models can engage in step-by-step reasoning even in the absence of task-specific fine-tuning, presenting a promising avenue for dynamic sentiment inference within a CoT framework. By utilizing CoT, sentiment evaluations can be adjusted dynamically during the inference process, rather than depending solely on static classification systems (Dong et al., 2024). Despite the potential of CoT for fuzzy sentiment analysis, several challenges persist, including the dependence on discrete reasoning and the insufficient integration with continuous fuzzy membership modeling (Liu et al., 2023; Liang et al., 2018).

3 Fuzzy Reasoning Chain

In this section, we introduce the Fuzzy Reasoning Chain (FRC) framework, which addresses the unclear reasoning in traditional Chain-of-Thought (CoT) methods and the rigidity of fixed-rule fuzzy reasoning, providing a fuzzy-to-clear inference approach. A comparison between CoT and FRC is shown in Fig. 1, where FRC refines sentiment analysis through fuzzy membership degrees, facilitating a smoother transition from fuzzy to clear sentiment analysis. Observations of LLMs(Guo et al., 2025; Liu et al., 2024a; Yang et al., 2024; Hurst et al., 2024) reveal key characteristics in the generated semantic membership degrees, particularly in sentiment analysis:

◇ ***Approximate Robustness***: Small changes in input result in bounded variations in membership degrees when context and syntax remain stable.

◇ ***Conditional Monotonicity***: Sentiment intensity correlates monotonically with membership degrees, except during abrupt context shifts.

◇ ***Dynamic Completeness***: Contextual mechanisms ensure full capture of essential semantic elements and their interactions.

Based on these characteristics, FRC consists of three main steps: Continuous Membership Degree, Multi-Granular Semantic Parsing, and Global Decision Fusion.

3.1 Continuous Membership Degree

Let C represents the sentiment class (e.g., positive or negative), and X denotes a given text unit. For sentiment analysis, large language models are capable of computing membership degrees reliably, which avoids the need for manually defined membership functions. The membership function $\mu_C(X)$ can therefore be computed from the output of a large language model as follows:

$$\mu_C(X) = f_{\text{LLM-prompt}}(X, C) \in [0, 1], \tag{1}$$

This ensures approximate robustness to input changes, where small variations in the input lead to bounded changes in the membership degree when the context and sentiment class remain stable.

3.2 Multi-Granular Semantic Parsing

We expect the large model to use hierarchical reasoning for semantic decomposition and aggregation, ensuring conditional monotonicity. The key steps are:

Keyword Membership Degree Calculation: We begin by extracting sentiment keywords $\{k_i\}$ from the input text X and calculating the corresponding membership degrees for each keyword:

$$\mu_C(k_i) = f_{\text{LLM-prompt}}(k_i, C) \in [0, 1], \tag{2}$$

where $\mu_C(k_i)$ represents the membership degree of keyword k_i in sentiment class C.

Local Semantic Aggregation: After extracting the keywords, the next step is to apply a weighted aggregation scheme to the sub-units X_j of the text. A sub-unit is defined as a portion of the text without emotional overlap, whose membership degree is computed as:

$$\mu_C(X_j) = \max_{k_i \in I_j} \mu_C(k_i), \tag{3}$$

where I_j is the set of keywords k_i in sub-unit X_j, and the maximum membership degree of the keywords is taken as the membership degree for the sub-unit. This ensures that the strongest sentiment influence from the keywords determines the sentiment of the sub-unit.

3.3 Global Decision Fusion

The large language model integrates information from sub-units X_j to determine the overall sentiment membership degree for input text X. Before the final decision, the model dynamically adjusts the weight based on context, sentiment intensity, and shifts. Key factors considered include, but are not limited to, the following:

◇ *Language Phenomena:* Shifts in tone, irony, or implication influencing sentiment strength or direction.

◇ *Sentiment Intensity:* Changes in intensity requiring weight adjustments.

◇ *Contextual Shifts:* Context changes, e.g., from descriptive to evaluative, triggering adjustments.

This adjustment ensures semantic completeness, capturing relevant elements and interactions for a final sentiment assessment, as follows:

$$\mu_C(X) = \sum_{j=1}^{m} \alpha_{j,C} \cdot \mu_C(X_j), \tag{4}$$

where m is the total number of sub-units X_j, $\alpha_{j,C}$ is class-specific dynamic weight for sub-unit X_j under the class C, satisfying $\sum_{j=1}^{m} \alpha_{j,C} = 1$ for each C.

By defining class-specific weight $\alpha_{j,C}$ for sub-units X_j and ensuring independence of membership degrees, we can quantify fine-grained sentence emotions and distinguish neutral or conflicting texts, which traditional probability methods cannot do.

4 Convergence Analysis

This section presents the convergence properties of the FRC framework, focusing on three key characteristics: Approximate Robustness, Conditional Monotonicity, and Dynamic Completeness. These properties contribute to the stability and interpretability of sentiment analysis results, even when the input data is uncertain or imprecise.

4.1 Approximate Robustness

During the *Keyword Membership Degree Calculation* and *Local Semantic Aggregation* stages, the sentiment membership degree $\mu_C(X)$ for each subunit is determined based on the strength of sentiment-bearing keywords. Minor input changes, such as syntactic variations or synonym substitutions, lead to smooth, bounded changes in $\mu_C(X)$, thereby ensuring stable and bounded sentiment analysis.

Bounded Analysis: Consider two input texts, X and X', with $d(X, X')$ representing their semantic distance. Assuming that the LLM preserves local smoothness in its semantic mappings, and referencing Eq.(4), the difference in membership degrees between X and X' can be expressed as:

$$|\mu_C(X) - \mu_C(X')| = \left| \sum_{j=1}^{m} g(\alpha_{j,C}, X_j) - g(\alpha'_{j,C}, X'_j) \right|$$

where $g(\alpha_C, X) = \alpha_C \cdot \mu_C(X)$. Focusing on the difference in membership degrees for each subunit X_j, we recognize that while the output of the language model for each keyword k_i may not strictly adhere to a continuity constraint, it generally exhibits approximate robustness to small input perturbations. This behavior is especially prominent when the context and syntax of the input remain stable. We hypothesize that the language model demonstrates local stability with respect to minor input variations, so these changes do not cause disproportionately large shifts in the output. Specifically, for

each subunit X_j, the difference in its membership degrees relative to X'_j can be bounded as:

$$|\mu_C(X_j) - \mu_C(X'_j)| \leq L \cdot d(X_j, X'_j), \quad (5)$$

where L is a constant that reflects the model's stability with respect to small input variations, depending on the properties of both the keywords and the context. By summing over all subunits and accounting for the weight adjustments, we derive the final relationship:

$$|\mu_C(X) - \mu_C(X')| \leq K \cdot d(X, X'), \quad (6)$$

where $K = \sum_{j=1}^{m} \alpha_{j,C} \cdot L$. This bounded property enables FRC to exhibit approximate Lipschitz continuity, a phenomenon commonly observed in deep learning models (Fazlyab et al., 2019; Kim et al., 2021; Shang et al., 2021), indicating that the FRC framework demonstrates stability and robustness to small perturbations.

4.2 Conditional Monotonicity

In the *Local Semantic Aggregation* stage, the membership degree $\mu_C(X_j)$ for each subunit is determined by the sentiment strength $s(X_j)$ of the relevant keywords. As sentiment strength increases, the membership degree also increases, maintaining a consistent, monotonic relationship. This is preserved in the *Global Decision Fusion*, where the final sentiment membership degree $\mu_C(X)$ is derived by aggregating the weighted membership degrees of each subunit, as defined in Eq. 4.

Given the text X, consider binary sentiment analysis (positive and negative). Traditional Chain-of-Thought (CoT) methods output probabilities, as shown in Fig. 1 (left), where the relationship between sentiment strength and the final output is linearly predictable. Specifically, for a positive sentiment change $\Delta s_{\text{positive}}$, the probabilities of both positive and negative classes change proportionally:

$$\Delta P(\text{positive}) = f_{\text{CoT}}(\Delta s_{\text{positive}}),$$
$$\Delta P(\text{negative}) = f_{\text{CoT}}(\Delta s_{\text{positive}}),$$

where f_{CoT} is a linear function. The changes are symmetric and linear, ensuring predictable monotonicity.

In contrast, the Fuzzy Reasoning Chain (FRC) independently assigns weights to subunits for each sentiment class, as shown in Fig. 1 (right). The membership degrees for the positive and negative classes $\mu_{\text{positive}}(X)$ and $\mu_{\text{negative}}(X)$ are determined separately, which may result in different responses to sentiment changes:

$$\Delta\mu_{\text{positive}}(X) = f_{\text{FRC-positive}}(\Delta s_{\text{positive}}),$$
$$\Delta\mu_{\text{negative}}(X) = f_{\text{FRC-negative}}(\Delta s_{\text{negative}}).$$

Here, $f_{\text{FRC-positive}}$ and $f_{\text{FRC-negative}}$ are monotonic mapping functions specific to each sentiment class. As sentiment changes, the impact on the positive and negative classes may differ, leading to less predictable relationships between sentiment intensity and global membership degree. This suggests that FRC's monotonicity might not always follow a linear pattern and could vary across classes.

While CoT typically exhibits a linear, predictable monotonicity, FRC offers more flexibility, potentially allowing for more nuanced relationships between sentiment and class membership. This flexibility could enable FRC to better capture complex sentiment dynamics, although its global monotonicity may not be as straightforward as that of traditional CoT methods

4.3 Dynamic Completeness

The *Dynamic Completeness* property ensures that all essential semantic elements are captured and appropriately weighted in the final sentiment decision. In FRC, each subunit X_j represents a critical semantic fragment, and its importance is determined dynamically based on its context. The weights $\alpha_{j,C}$ are adjusted based on the relevance of each subunit, ensuring that all relevant semantic information contributes to the final membership degree $\mu_C(X)$. This is reflected in Eq.(4). By doing so, FRC captures not only individual sentiment influences but also the interactions between these influences, ensuring a comprehensive sentiment analysis.

4.4 Summary

Through the analysis of *Approximate Robustness*, *Conditional Monotonicity*, and *Dynamic Completeness*, the FRC framework appears to effectively handle fuzzy inputs, facilitating stable and consistent sentiment analysis. These properties contribute to smooth transitions from uncertainty to clarity, supporting reliable and interpretable results.

5 Experiments

To validate the effectiveness of the Fuzzy Reasoning Chain (FRC) framework in sentiment analysis, we conduct experiments on both English and Chinese datasets. The evaluation focuses on robustness, monotonicity, and classification performance in comparison with CoT-based and direct prompting methods.

5.1 Datasets

In this section, we describe the datasets used in our evaluation. The datasets are divided into two categories: the original datasets used for standard sentiment classification, and the perturbed datasets, which are generated by extending the original datasets for the analysis of FRC convergence and do not require class labels.

5.1.1 Original Datasets

We adopt two well-known datasets for sentiment analysis tasks, selected to ensure both linguistic diversity and domain representation:

- ***SemEval-2016 Task 4 (English)***: A widely used benchmark for sentiment analysis, specifically focused on the evaluation of fine-grained sentiment in social media. It contains a collection of labeled tweets, spanning multiple sentiment classes (5 classes, ranging from negative to positive). The dataset includes over 9,000

labeled tweets, providing a balanced mix of different sentiment categories and domains. The test set consists of 1,791 tweets.

- ***Takeout Review Dataset (Chinese)***: This proprietary dataset is collected from online food delivery reviews and consists of over 10,000 labeled reviews in Chinese. It includes text paired with sentiment labels (positive and negative), focusing on the sentiment expressed in the context of food and restaurant experiences. The large size of this dataset allows for comprehensive training and testing in a real-world application setting.

5.1.2 Perturbed Datasets

To evaluate the robustness and monotonicity of our FRC, we generate perturbed versions of the original datasets using the GPT-4o (Hurst et al., 2024) API through prompt engineering techniques. These perturbations are designed to test the FRC's ability to handle different levels of changes in the input while preserving core sentiment or adjusting sentiment intensity. The perturbations are divided into two main categories:

- ***Robustness Perturbations*** These perturbations are designed to preserve the original sentiment and meaning while modifying surface-level expressions. Three levels of perturbations are applied:

Low: Simple synonym replacements (1-2 words) to test the FRC's resilience to slight lexical changes.

Medium: Sentence restructuring and multiple word replacements to evaluate the FRC's ability to handle moderate perturbations.

High: Full sentence rewriting that retains the sentiment but significantly alters sentence structure.

- ***Monotonicity Perturbations***: These perturbations modify sentiment intensity by replacing sentiment-bearing words or adding adverbial modifiers to shift sentiment intensity. The goal is to evaluate the FRC's sensitivity to changes in sentiment intensity, either positive or negative. The labels are used as follows: -1 indicates a more negative sentiment, +1 indicates a more positive sentiment, and 0 indicates no change in sentiment.

- ***Quality Assurance***: GPT-4o's strong performance, combined with manual spot checks, ensures the high quality of the perturbed datasets.

5.2 Evaluation Metrics

To assess the effectiveness of FRC, we propose the following metrics:

- ***Robustness Score (RS)***: Measures the stability of membership degrees or probabilities under perturbations. Given an original text X and its perturbed counterpart X', we define:

$$RS = 1 - \frac{1}{N}\sum_{i=1}^{N} \left|\mu_C(X_i) - \mu_C(X_i')\right|,$$

where $\mu_C(X)$ is the sentiment membership degree (or probability for CoT) of text X, and N is total test samples. Higher values indicate greater robustness.

- ***Monotonicity Score (MS)***: Evaluates whether increasing sentiment intensity in perturbations leads to a corresponding increase (or decrease) in membership degree. It is computed as:

$$MS = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(sgn(\mu_C(X_i') - \mu_C(X_i)) = Y_{i,C}),$$

where $Y_{i,C}$ is the sentiment shift direction label corresponding to the monotonic perturbation data relative to class C (+1 for more aligned with C, -1 for moving away from C, 0 for no change), $\mathbb{I}(\cdot)$ is the indicator function (1 if the condition is holds, 0 otherwise). A higher MS value indicates better performance, as it signifies that the model consistently reflects the expected sentiment shift direction.

- ***F1-score:*** Standard metric for sentiment classification.

5.3 Baseline Methods

We compare FRC against two baselines:

- ***Direct Prompting (DP)***: Asking LLMs directly for sentiment classification without reasoning steps.

- ***CoT***: A CoT-base method that follows reasoning steps similar to FRC for fairness but outputs probabilities instead of membership degrees. The prompt design is referenced in the left part of Fig. 1.

5.4 Experimental Convergence Analysis

Theoretical Foundation: In Sec.4, we performed a preliminary convergence analysis of FRC, which indicated that, underpinned by robustness and monotonicity, the FRC framework inherently possesses dynamic completeness.

Experiment Setup: Building upon the theoretical foundation, we aim to further investigate the empirical aspects of robustness and monotonicity through targeted experiments. For this purpose, we selected three versions of DeepSeek R1 (32b, 14b, and 7b)(Guo et al., 2025) alongside Qwen2.5 32b (Yang et al., 2024) for evaluation. To ensure a more accurate assessment of FRC's convergence, we employed a mixed-language dataset, incorporating both source and perturbed data in Chinese and English, thereby eliminating language as a confounding factor.

Comparison Objective: CoT was chosen as a fairness comparison method to highlight FRC's unique advantages in the context of perturbation-induced sentiment shifts. ***Metrics***: The Robustness Score (RS) was used as the evaluation metric, with higher values indicating better performance.

5.4.1 Robustness Evaluation

Following the approximate robustness analysis in Sec.4.1, we focus on the experimental examination of FRC's robustness. Robustness ensures that small perturbations in input data do not cause significant fluctuations

Table 1: Robustness Comparison across Models via **RS** Metric - Higher values indicate better performance.

Model	Method	Low	Medium	High	Avg
Qwen2.5-32b	CoT	0.91	0.81	0.68	0.80
	FRC	0.94	0.87	0.80	0.87
DeepSeek-32b	CoT	0.92	0.83	0.72	0.82
	FRC	**0.95**	**0.89**	**0.82**	**0.89**
DeepSeek-14b	CoT	0.89	0.78	0.65	0.77
	FRC	0.93	0.85	0.78	0.85

Table 2: Monotonicity Evaluation across Models - Higher Monotonicity Score (MS) indicates better performance.

Model	Method	Positive	Negative	Avg MS
Qwen2.5-32b	CoT	0.84	0.82	0.83
	FRC	0.93	0.89	0.91
DeepSeek-32b	CoT	0.86	0.84	0.85
	FRC	**0.94**	**0.90**	**0.92**
DeepSeek-14b	CoT	0.81	0.79	0.80
	FRC	0.91	0.87	0.89

in the output, reinforcing the stability of FRC's reasoning process. As shown in Table 1, FRC consistently outperforms CoT across all models and perturbation levels (Low, Medium, High). The average robustness scores reveal a clear performance advantage for FRC.

For example, DeepSeek-32b achieved an average score of 0.89 with FRC, compared to 0.82 for CoT, resulting in an 8.5% improvement. Similarly, DeepSeek-14b showed a 10.4% improvement with FRC (0.85) over CoT (0.77), while Qwen2.5-32b demonstrated an 8.8% improvement with FRC (0.87) over CoT (0.80).

These results are consistent across models with different parameter sizes, showcasing FRC's robustness even when scaling down to smaller models like DeepSeek-14b. While the theoretical analysis highlighted FRC's inherent convergence from fuzziness to clarity, these experimental results further validate its superior robustness, even though the convergence is not absolute. Overall, the findings emphasize FRC's stability and its ability to handle perturbations that challenge traditional methods like CoT, with particular strength in the robustness of the approximation.

5.4.2 Monotonicity Evaluation

In addition to robustness, monotonicity is another key characteristic that influences FRC's convergence properties. This section investigates how the sentiment intensity of input perturbations correlates monotonically with the changes in membership degrees, and how FRC leverages this relationship to refine its reasoning and sentiment analysis.

The results for monotonicity evaluation across models are provided in Table 2. FRC demonstrates superior performance in capturing both positive and negative sentiment shifts, as shown by higher average MS scores across all models. For example, DeepSeek-32b with FRC achieved an average MS of 0.92, compared to 0.85 for CoT, indicating a significant improvement in the model's ability to reflect sentiment changes. Similarly, DeepSeek-14b and Qwen2.5-32b showed similar gains in MS with FRC.

5.5 Sentiment Classification

We evaluate DP, CoT, and FRC on the SemEval-2016 Task 4 and Takeout Review datasets, which provide positive and negative labels, with Task 4 also including neutral cases. Some samples exhibit inherently ambiguous sentiment, leading conventional prompt-based methods to produce unstable or neutral predictions. Using FRC's coarse positive/negative assessment, we observed that the membership degree differences for these samples are typically within 0.3. Based on this, we divide the data into *Clear* cases, with differences above 0.3, and *Ambiguous* cases, with differences at or below 0.3. This allows us to retain the original labels while enabling more detailed analysis. Even without relying on the ground-truth labels, finer distinctions, such as strong versus weak sentiment conflicts, can be examined by comparing membership degrees. For evaluation, the predicted label is assigned according to the higher membership degree, with equal values considered neutral. The *Average* category then reports overall performance across both *Clear* and *Ambiguous* cases.

Table 3: Sentiment Classification Results on SemEval-2016 Task 4 (English) using F1 Score, divided into Clear, Ambiguous, and Average categories.

Model	Method	Clear	Ambiguous	Avg
Qwen2.5-32b	DP	0.86	0.81	0.84
	CoT	0.88	0.82	0.85
	FRC	**0.90**	**0.85**	**0.88**
DeepSeek-32b	DP	0.83	0.75	0.79
	CoT	0.85	0.77	0.81
	FRC	**0.89**	**0.84**	**0.87**
DeepSeek-14b	DP	0.77	0.72	0.75
	CoT	0.79	0.73	0.76
	FRC	**0.83**	**0.78**	**0.81**

Table 4: Sentiment Classification Results on Takeout Review Dataset (Chinese) using F1 Score, divided into Clear, Ambiguous, and Average categories.

Model	Method	Clear	Ambiguous	Avg
Qwen2.5-32b	DP	0.84	0.79	0.81
	CoT	0.86	0.80	0.83
	FRC	**0.88**	**0.83**	**0.86**
DeepSeek-32b	DP	0.81	0.74	0.77
	CoT	0.84	0.76	0.80
	FRC	**0.89**	**0.84**	**0.87**
DeepSeek-14b	DP	0.73	0.68	0.70
	CoT	0.75	0.71	0.74
	FRC	**0.79**	**0.74**	**0.77**

The experimental results demonstrate that FRC consistently outperforms both DP and CoT across all models and datasets.

SemEval-2016 Task 4: As shown in Table 3. , FRC achieves the highest F1 scores in both the *Clear* and *Average* categories, showing substantial improvements over CoT and DP.

Takeout Review Dataset: On the Takeout Review Dataset (Chinese), as shown in Table 4, FRC also outperforms the baseline methods in both *Clear* and *Average* categories, indicating its superior ability to distinguish sentiment in diverse languages.

The results show that the difference in performance between FRC and the other methods is particularly pronounced in the *Clear* category, where FRC demonstrates more robust sentiment classification. This is consistent with FRC's ability to process sentiment shifts and handle fine-grained sentiment distinctions more effectively.

Furthermore, the model size appears to have a positive impact on FRC's performance, with larger models such as DeepSeek-32b and Qwen2.5-32b consistently achieving better performance across both datasets.

5.6 Knowledge Transfer from Large to Small Models

To evaluate the knowledge transfer capability of the FRC framework from large to small models, we conduct prompt-based experiments by injecting intermediate reasoning results, specifically keyword knowledge and sub-unit knowledge, extracted from the DeepSeek-R1 32b model into the prompts of smaller models, including the 7b and 1.5b models. This strategy, similar to retrieval-augmented generation, provides structural reasoning guidance without modifying any model parameters. As shown in Table 5, the injected knowledge leads to substantial improvements in the 7b model's F1-score.

Table 5: Knowledge Transfer from DeepSeek-R1 32b to Smaller Models via FRC (F1-Score Results)

Prompt Configuration	1.5b	7b
No injection (baseline)	0.62	0.76
Injecting Keyword Knowledge	0.68	0.79
Injecting Sub-unit Knowledge	0.72	0.81
Injecting Keyword & Sub-unit Knowledge	**0.75**	**0.83**

Analysis: The baseline 1.5b and 7b models achieve F1-scores of 0.62 and 0.76, which are substantially lower than the 32b model, which reaches 0.87. This gap indicates the limitations of smaller models in handling long prompts, maintaining global context, and performing complex reasoning. By injecting fine-grained keyword and sub-unit knowledge derived from the 32b model, these deficiencies are effectively mitigated. In particular, the 1.5b model reaches 0.75 F1-score, corresponding to a relative improvement of 21 percent, while the 7b model achieves 0.83 F1-score. These results demonstrate that smaller models can narrow the performance gap without any parameter updates.

The improvements obtained by injecting keyword knowledge alone, sub-unit knowledge alone, and their combination indicate that multi-level knowledge components complement each other in enhancing reasoning performance. Moreover, this offline knowledge transfer mechanism enables small models to process new sentences independently by querying a pre-built knowledge base or leveraging offline-enhanced training, without requiring online invocation of large models. This approach allows for efficient and low-cost deployment while maintaining robust reasoning capability.

6 Conclusion

In this work, we propose the Fuzzy Reasoning Chain (FRC) framework to address the challenge of reasoning over ambiguous and uncertain texts. FRC extends conventional probability-based reasoning by integrating continuous fuzzy membership degrees, providing a structured mechanism for translating fuzzy information into interpretable decisions. We provide a preliminary exploration of the framework's properties, and empirical results demonstrate its effectiveness on sentiment analysis tasks. The transition from probabilities to membership degrees allows FRC to capture conflicting or uncertain signals that traditional approaches cannot, and also facilitates knowledge transfer between models of different scales. Future work may explore applying FRC to other fuzzy reasoning tasks such as security review and ethical judgment, integrating richer external knowledge sources, and further improving robustness and efficiency for practical deployments.

7 Limitations

While the proposed Fuzzy Reasoning Chain (FRC) framework shows promising results, several limitations remain to be addressed in future work.

First, current large language models still face challenges in precisely executing the FRC prompting process end-to-end. Similar to many existing prompt-based tasks(Kojima et al., 2022), practical deployment often requires manual interpretation of intermediate outputs or multiple rounds of interaction with the model , which may limit efficiency and scalability.

Second, although we focus on sentiment analysis for evaluation, other fuzzy reasoning tasks—such as security review, ethical judgment, and culturally sensitive scenarios—pose additional complexities due to their context-dependent semantics and domain-specific knowledge requirements. Careful task-specific adaptation and thorough empirical validation are needed before applying FRC to these areas.

Third, the generalization of FRC beyond sentiment analysis remains to be systematically explored. Different fuzzy tasks may have distinct ambiguity types and semantic structures, which could impact the framework's reasoning stability and interpretability.

Fourth, the reliance on large language models as semantic priors implies a dependency on the inherent biases and knowledge limitations of these models. This

may affect the robustness of FRC's reasoning in scenarios with underrepresented or evolving language phenomena.

Finally, while FRC facilitates knowledge transfer via prompt injection, the granularity and optimal design of transferable reasoning components warrant deeper investigation to maximize cross-model efficacy.

Addressing these limitations will be crucial for further enhancing FRC's applicability and robustness in diverse fuzzy reasoning domains.

8 Ethics and Impact Statement

This paper presents the Fuzzy Reasoning Chain (FRC) framework designed to improve the interpretability and robustness of sentiment analysis through fuzzy logic-based reasoning. By enabling models to handle ambiguous or uncertain emotional expressions more effectively, our work aims to enhance natural language understanding in real-world scenarios.

While we do not identify immediate ethical concerns inherent to our methodology, we acknowledge that sentiment analysis technologies can influence user experience and decision-making in various applications such as content moderation, recommendation systems, and social media analysis. It is therefore important that future work applying FRC continues to consider fairness, privacy, and the avoidance of biased or misleading interpretations.

Our goal is to support the development of more nuanced and reliable sentiment analysis tools that better reflect human emotional complexity, ultimately benefiting users and society.

References

Liu Bing. 2012. Sentiment analysis and opinion mining (synthesis lectures on human language technologies). *University of Illinois: Chicago, IL, USA*.

Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59.

Iti Chaturvedi, Ranjan Satapathy, Sandro Cavallari, and Erik Cambria. 2019. Fuzzy commonsense reasoning for multimodal sentiment analysis. *Pattern Recognit. Lett.*, 125:264–270.

Yuyan Chen and Yanghua Xiao. 2024. Recent advancement of emotion cognition in large language models. *arXiv preprint arXiv: 2409.13354*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186.

Hoang Nam Do, Huyen Trang Phan, and Ngoc Thanh Nguyen. 2024. Multimodal sentiment analysis using deep learning and fuzzy logic: A comprehensive survey. *Appl. Soft Comput.*, 167:112279.

Ximing Dong, Shaowei Wang, Dayi Lin, Gopi Krishnan Rajbahadur, Boquan Zhou, Shichao Liu, and Ahmed E Hassan. 2024. Promptexp: Multi-granularity prompt explanation of large language models. *arXiv preprint arXiv:2410.13073*.

Zhihua Duan and Jialin Wang. 2024. Implicit sentiment analysis based on chain of thought prompting. *arXiv preprint arXiv: 2408.12157*.

Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George Pappas. 2019. Efficient and accurate estimation of lipschitz constants for deep neural networks. *Advances in neural information processing systems*, 32.

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. 2023. Reasoning implicit sentiment with chain-of-thought prompting. In *Association for Computational Linguistics*, pages 1171–1182.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625.

Xu Gu, Xiaoliang Chen, Peng Lu, Zonggen Li, Yajun Du, and Xianyong Li. 2024. Agcvt-prompt for sentiment classification: Automatically generating chain of thought and verbalizer in prompt learning. *Eng. Appl. Artif. Intell.*, 132:107907.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv: 2501.12948*.

Oumaima Hourrane, Nouhaila Idrissi, and El Habib Benlahmar. 2019. Sentiment classification on movie reviews and twitter: An experimental study of supervised learning models. In *International Conference on Smart Systems and Data Science*, pages 1–6.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Shima Imani, Liang Du, and Harsh Shrivastava. 2023. Mathprompter: Mathematical reasoning using large language models. In *Association for Computational Linguistics: Industry Track*, pages 37–42.

Fatma Jemai, Mohamed Hayouni, and Sahbi Baccar. 2021. Sentiment analysis using machine learning algorithms. In *International Wireless Communications and Mobile Computing*, pages 775–779.

Feihu Jin, Yifan Liu, and Ying Tan. 2024. Zero-shot chain-of-thought reasoning guided by evolutionary algorithms in large language models. *arXiv preprint arXiv: 2402.05376*.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. 2021. The lipschitz constant of self-attention. In *International Conference on Machine Learning*, pages 5562–5571.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

D. Ashok Kumar and Anandan Chinnalagu. 2020. Sentiment and emotion in social media covid-19 conversations: Sab-lstm approach. In *International Conference System Modeling and Advancement in Research Trends*, pages 463–467.

Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2016. Rationalizing neural predictions. In *Conference on Empirical Methods in Natural Language Processing*, pages 107–117.

Yan Li, Xiangyuan Lan, Haifeng Chen, Ke Lu, and Dongmei Jiang. 2024. Multimodal pear chain-of-thought reasoning for multimodal sentiment analysis. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(9):1–23.

Xiaodan Liang, Zhiting Hu, Hao Zhang, Liang Lin, and Eric P Xing. 2018. Symbolic graph reasoning meets convolutions. *Advances in neural information processing systems*, 31.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv: 2412.19437*.

Liang Liu, Dong Zhang, Suyang Zhu, and Shoushan Li. 2024b. Generative chain-of-thought for zero-shot cognitive reasoning. In *International Conference on Artificial Neural Networks*, pages 324–339.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135.

Wang Peizhuang. 1983. Pattern recognition with fuzzy objective function algorithms (james c. bezdek). *Siam Review*, 25(3):442.

Liyizhe Peng, Zixing Zhang, Tao Pang, Jing Han, Huan Zhao, Hao Chen, and Björn W. Schuller. 2024. Customising general large language models for specialised emotion recognition tasks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 11326–11330.

Qiao Qian, Minlie Huang, Jinhao Lei, and Xiaoyan Zhu. 2017. Linguistically regularized LSTM for sentiment classification. In *Association for Computational Linguistics*, pages 1679–1689.

Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. 2017. A simple neural network module for relational reasoning. *Advances in neural information processing systems*, 30.

Yuzhang Shang, Bin Duan, Ziliang Zong, Liqiang Nie, and Yan Yan. 2021. Lipschitz continuity guided knowledge distillation. In *IEEE/CVF international conference on computer vision*, pages 10675–10684.

Hong Sun, Xue Li, Yinchuan Xu, Youkow Homma, Qi Cao, Min Wu, Jian Jiao, and Denis Charles. 2023. Autohint: Automatic prompt optimization with hint generation. *arXiv preprint arXiv:2307.07415*.

Mirac Suzgun and Adam Tauman Kalai. 2024. Meta-prompting: Enhancing language models with task-agnostic scaffolding. *arXiv preprint arXiv.2401.12954*.

Asmaa Hashem Sweidan, Nashwa El-Bendary, and Haytham Al-Feel. 2022. Word embeddings with fuzzy ontology reasoning for feature learning in aspect sentiment analysis. In *International Conference on Artificial Neural Networks*, pages 320–331.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *Conference on Empirical Methods in Natural Language Processing*, page 1422–1432.

Dimple Tiwari and Bharti Nagpal. 2022. KEAHT: A knowledge-enriched attention-based hybrid transformer model for social sentiment analysis. *New Gener. Comput.*, 40(4):1165–1202.

Xingchen Wan, Ruoxi Sun, Hanjun Dai, Sercan O Arik, and Tomas Pfister. 2023. Better zero-shot reasoning with self-adaptive prompting. In *Association for Computational Linguistics*, pages 3493–3514.

Xin Wang, Jianhui Lyu, Byung-Gyu Kim, Parameshachari Bidare Divakarachari, Keqin Li, and Qing Li. 2025. Exploring multimodal multiscale features for sentiment analysis using fuzzy-deep neural network learning. *IEEE Trans. Fuzzy Syst.*, 33(1):28–42.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou,

et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Cynthia Whissell. 2009. Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language. *Psychological Reports*, 105(2):509–521.

Hao Wu, Danping Yang, Peng Liu, and Xianxian Li. 2025. Chain of thought guided few-shot fine-tuning of llms for multimodal aspect-based sentiment classification. In *International Conference on Multimedia Modeling*, pages 182–194.

Weijia Xu, Andrzej Banburski-Fahey, and Nebojsa Jojic. 2024. Reprompting: Automated chain-of-thought prompt inference through gibbs sampling. In *International Conference on Machine Learning*.

Ruiteng Yan, Yan Yu, and Dong Qiu. 2022. Emotion-enhanced classification based on fuzzy reasoning. *Int. J. Mach. Learn. Cybern.*, 13(3):839–850.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Liu Yinhan, Ott Myle, Goyal Naman, Du Jingfei, Joshi Mandar, Chen Danqi, Levy Omer, and Lewis Mike. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *WIREs Data Mining Knowl. Discov.*, 8(4).

Yiqun Zhang, Xiaocui Yang, Xingle Xu, Zeran Gao, Yijie Huang, Shiyi Mu, Shi Feng, Daling Wang, Yifei Zhang, Kaisong Song, and Ge Yu. 2024a. Affective computing in the era of large language models: A survey from the NLP perspective. *arXiv preprint arXiv: 2408.04638*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *International Conference on Learning Representations*.

Zixing Zhang, Liyizhe Peng, Tao Pang, Jing Han, Huan Zhao, and Björn W. Schuller. 2024b. Refashioning emotion recognition modeling: The advent of generalized large models. *IEEE Transactions on Computational Social Systems*, 11(5):6690–6704.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2023. Least-to-most prompting enables complex reasoning in large language models. In *International Conference on Learning Representations*.