HFR-Prompt: Hierarchical Feedback Reasoning Prompting for Enhanced Large Language Model Comment Feedback Prediction

Zeyuan Xun * and Yichen Ku

Posted Date: 17 March 2026

doi: 10.20944/preprints202603.1361.v1

Keywords: feedback prediction; large language models; prompting; hierarchical reasoning; interpretability

 doi:10.20944/preprints202603.1361.v1

Article

HFR-Prompt: Hierarchical Feedback Reasoning Prompting for Enhanced Large Language Model Comment Feedback Prediction

Zeyuan Xu * and Yichen Ku

Kunming University of Science and Technology
* Correspondence: 202143956043@stu.kust.edu.cn

Abstract

The accurate prediction of feedback from user comments is essential yet challenging, often limited by the nuanced semantics that traditional Natural Language Processing and existing Large Language Model prompts struggle to capture. We propose the Hierarchical Feedback Reasoning Prompting (HFR-Prompt) framework to address this. HFR-Prompt guides Large Language Models through a multi-stage, logically progressive analysis comprising Initial Tendency Assessment, Fine-grained Feedback Type Identification, and Result Integration and Explanation Generation. Each successive stage builds upon the contextual understanding established by the previous one. Extensive experiments on a substantial dataset demonstrate that HFR-Prompt significantly outperforms strong LLM baselines and standard prompting techniques in terms of accuracy, Macro-F1 score, and crucial explanation consistency. While introducing a computational overhead, HFR-Prompt sets a new standard for interpretable and accurate comment feedback prediction, validating the efficacy of structured, hierarchical reasoning in complex LLM applications.

Keywords: feedback prediction; large language models; prompting; hierarchical reasoning; interpretability

1. Introduction

In the digital era, user-generated comments have become an indispensable component of various online platforms, including social media, e-commerce websites, forums, and academic review systems. Accurately and efficiently predicting the feedback type of these comments (e.g., approval, rejection, constructive suggestion) is of paramount importance for automated content moderation, enhancing user experience, optimizing recommendation systems, and fostering healthy community environments. Traditional Natural Language Processing (NLP) methods often rely on extensive annotated datasets and complex model training pipelines [1]. However, with the rapid advancement of large language models (LLMs) such as the GPT series, Claude, Gemini, and Qwen [2], they have demonstrated powerful in-context learning capabilities [3]. These LLMs can adapt to new tasks through well-designed prompts, significantly reducing the demand for vast amounts of domain-specific annotated data and the costs associated with model fine-tuning. The broad applicability of advanced modeling techniques extends beyond language, influencing diverse fields from efficient resource management and carbon footprint estimation in digital infrastructure [4,5] to sophisticated parameter identification in complex engineering systems like permanent magnet synchronous motors [6–8].

Despite their excellent performance on general reasoning tasks, the efficacy of LLMs in specific domain tasks is often heavily influenced by the quality of prompt design. Existing prompt engineering methodologies primarily focus on one-shot instructions or simple Chain-of-Thought (CoT) reasoning [9]. For tasks like comment feedback prediction, which necessitate nuanced semantic understanding and multi-dimensional considerations, and even require strict adherence to specific constraints [10],

 doi:10.20944/preprints202603.1361.v1

there remains considerable room for improvement. This research aims to explore a more structured and hierarchical prompt engineering approach to guide LLMs in achieving more accurate and interpretable performance in the comment feedback prediction task.

To address these challenges, we propose a novel approach named **Hierarchical Feedback Reasoning Prompting (HFR-Prompt)**. This framework is designed to facilitate deep analysis of comments by LLMs through a multi-stage, logically progressive prompting mechanism, leading to more accurate predictions of feedback types. Distinct from conventional one-shot prompts or simple CoT methods, HFR-Prompt decomposes the complex feedback prediction task into several logically interdependent sub-tasks. Each sub-task is then guided by a specially crafted prompt tailored to elicit specific reasoning from the LLM. Our HFR-Prompt framework encompasses three primary stages: (1) *Initial Tendency Assessment*, where the LLM gauges the overall sentiment (positive, neutral, or negative); (2) *Fine-grained Feedback Type Identification*, which further categorizes the feedback within the identified tendency (e.g., discerning between 'rejection' and 'constructive criticism' for negative comments); and (3) *Result Integration & Explanation Generation*, where the LLM synthesizes the judgments from previous stages to output a final feedback type along with a concise, logical explanation.

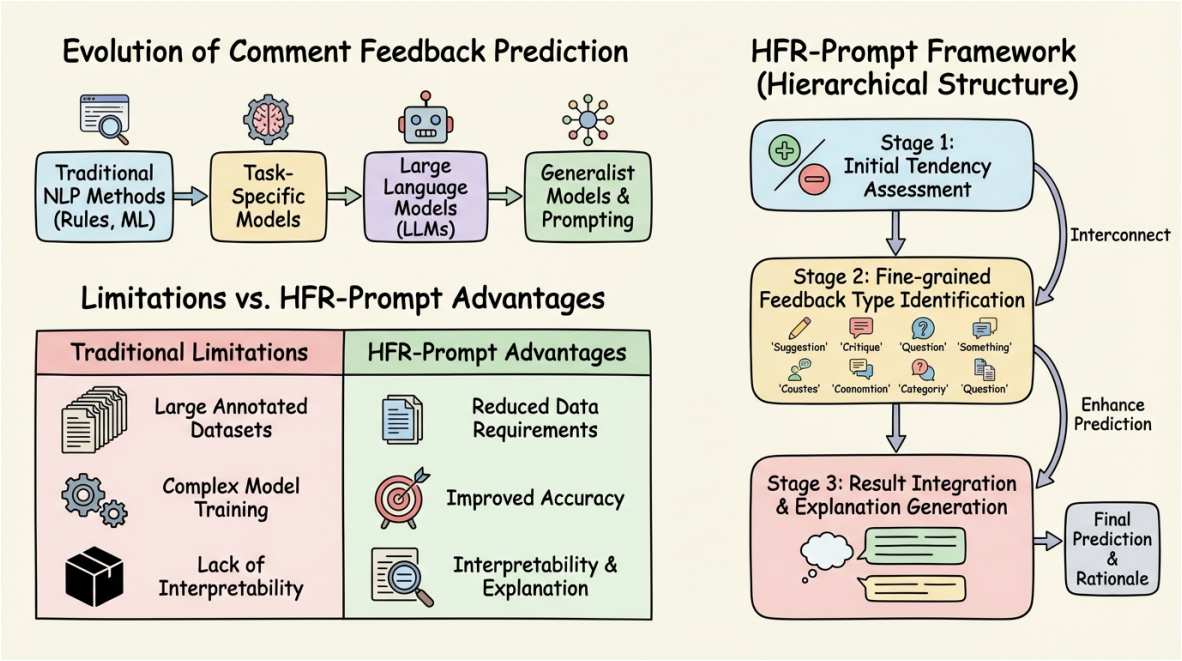

Figure 1. An overview of the Hierarchical Feedback Reasoning Prompting (HFR-Prompt) framework. The figure illustrates the evolution of comment feedback prediction methods, highlights the advantages of HFR-Prompt over traditional approaches, and details the three-stage hierarchical structure of HFR-Prompt for comprehensive comment analysis and feedback type prediction.

To thoroughly evaluate the effectiveness of our proposed HFR-Prompt method, we conducted extensive experiments on a publicly available, human-annotated comment-feedback dataset. This dataset comprises 50,000 comment-feedback pairs sourced from online forums, social platforms, and academic review scenarios. The feedback in this dataset is meticulously labeled into three main categories: *Positive* (indicating approval, support, appreciation), *Constructive* (encompassing suggestions for improvement, conditional acceptance), and *Negative/Rejection* (expressing opposition, disapproval, dismissal). The dataset was rigorously partitioned into a training set (40,000), a validation set (5,000), and a test set (5,000) to ensure the robustness and generalizability of our experimental setup. We benchmarked our approach against several state-of-the-art generic large language models, including Qwen3-7B, Claude, Gemini, and a baseline GPT-5 model employing standard prompting techniques. Performance was evaluated using Accuracy, Macro-F1 score, and Explanation Consistency, with our **GPT-5 + HFR-Prompt** method demonstrating superior predictive capabilities and interpretability

 doi:10.20944/preprints202603.1361.v1

across all metrics, as detailed in the experimental results. Specifically, our method achieved an accuracy of 84.1%, a Macro-F1 score of 83.5%, and an explanation consistency of 82.0%, outperforming the GPT-5 baseline (82.4% accuracy, 81.7% Macro-F1, 80.3% explanation consistency) and other LLMs.

The main contributions of this work are summarized as follows:

- We propose a novel **Hierarchical Feedback Reasoning Prompting (HFR-Prompt)** framework that systematically guides LLMs through multi-stage logical decomposition for enhanced comment feedback prediction.
- We demonstrate that HFR-Prompt significantly improves the accuracy and Macro-F1 score of LLMs in comment feedback prediction compared to traditional prompting methods and other strong baseline LLMs.
- We introduce and evaluate the *Explanation Consistency* metric, showing that HFR-Prompt not only enhances predictive performance but also generates more logically coherent and interpretable explanations for its predictions.

2. Related Work

2.1. Prompt Engineering for Large Language Models

Prompt engineering is crucial for leveraging Large Language Models (LLMs), covering foundational techniques, advanced reasoning, and output optimization. Foundational approaches include *prompt tuning*, where Pre-trained Prompt Tuning (PPT) enhances few-shot learning [11], and *in-context learning* (ICL), which improves performance [12] and whose stability has been analyzed [3]. To enhance ICL efficiency, LLMLingua offers prompt compression for faster inference and cost reduction [13]. Other efficiency efforts involve KV cache management [4,14] and parallel graph-retrieval reasoning for inference scaling [15].

Prompt engineering significantly enhances LLM reasoning capabilities. *Chain-of-Thought* (CoT) prompting improves natural language understanding and mimics human reasoning patterns [16]. Building on this, *Structured Prompting*, such as role-play prompting, boosts zero-shot reasoning by effectively triggering CoT processes [17]. Approaches for *Multi-stage Reasoning* in LLMs have been surveyed [18], and Chain-of-Specificity enhances task-specific constraint adherence [10]. Advanced models like F1 demonstrate broader trends in bridging understanding and generation to actions [2].

For practical applications, evaluating and optimizing LLM outputs is vital. G-Eval leverages LLMs (GPT-4 with CoT) for NLG evaluation, achieving superior human alignment [19]. Research also explores sampling temperature's effect on problem-solving [20] and token-importance guided direct preference optimization [21]. These developments, alongside advancements in cooperative edge caching [22], decision models [23], robotic policies [24], and remote sensing analysis [25], continuously refine how optimal performance is extracted from LLMs.

2.2. Comment Feedback Prediction and Analysis

Comment feedback prediction is crucial for understanding user opinions and moderating online content, covering various Natural Language Processing (NLP) tasks. Core to this area is *sentiment analysis* and *opinion mining*: datasets for aspect category sentiment and rating prediction have been introduced [26], and CoT frameworks like THOR reason about implicit sentiment [27]. For *content moderation*, cross-modal prediction enhances multimodal sentiment analysis [28].

Beyond sentiment, broader understanding of comments involves *text classification*, utilizing Transformer-based models for long texts [29], and named entity recognition via span prediction, extensible to feedback prediction [30]. In *User-Generated Content (UGC)* environments, personalized news recommendation models user interests and time-aware popularity, offering insights into engagement dynamics [31].

For analyzing complex comment structures, GraphPrompt unifies graph pre-training and downstream tasks [32], similar to graph-based governance for fraudulent traffic [33]. Foundational advancements in *Natural Language Processing (NLP)* underpin these tasks, as exemplified by UniXcoder,

 doi:10.20944/preprints202603.1361.v1

a cross-modal model for code representation that incorporates code comments [34]. Ongoing multimodal benchmarks, such as for fine-grained retrieval-augmented generation [35], further enrich content analysis.

3. Method

This section details our proposed **Hierarchical Feedback Reasoning Prompting (HFR-Prompt)** framework, a novel approach designed to significantly enhance the performance of large language models (LLMs) in the intricate task of comment feedback prediction. By structuring the problem into a sequence of logically ordered sub-tasks, HFR-Prompt guides the LLM through a systematic and stepwise analysis of user comments, fostering both accuracy and interpretability in its predictions. We first elucidate the fundamental philosophy underpinning HFR-Prompt, followed by a comprehensive description of its three constituent stages.

The core innovation of HFR-Prompt lies in its ability to decompose the inherently complex task of predicting a specific feedback type, denoted as F, from a given user comment, C, into a series of more manageable and structured reasoning steps. Traditional prompting paradigms, such as simplistic one-shot prompting or even rudimentary Chain-of-Thought (CoT) methods, often attempt to infer F directly from C using a single, monolithic prompt $P_{one-shot}$. This approach can be represented as:

$$F = \text{LLM}(C, P_{one-shot}) \tag{1}$$

While straightforward, such direct inference often struggles with the nuanced semantics and contextual complexities inherent in human language, potentially leading to less accurate or less transparent outcomes.

In contrast, HFR-Prompt adopts a multi-stage, hierarchical process. A defining characteristic of this framework is that the derived output from an earlier stage is explicitly fed as enriched contextual input to subsequent stages. This architectural choice promotes a more robust and verifiable reasoning trajectory, enabling the LLM to construct a comprehensive understanding before arriving at a final decision. This systematic decomposition not only aids in dissecting the problem into digestible parts but also mirrors human cognitive processes when evaluating complex information. The overall prediction mechanism within HFR-Prompt can be formally conceptualized as a sequential chain of LLM inferences:

$$(T) = \text{LLM}(C, P_1) \tag{2}$$

$$(F') = \text{LLM}(C, T, P_2) \tag{3}$$

$$(F_{final}, E) = \text{LLM}(C, T, F', P_3) \tag{4}$$

Within this framework, P_1, P_2, and P_3 denote the meticulously crafted and specialized prompts tailored for each respective stage. The intermediate variable T represents the initial, broad categorization of the comment's tendency (e.g., positive, neutral, or negative). F' signifies the identified fine-grained feedback type, which is a more specific classification building upon T. Finally, F_{final} is the ultimate, definitive feedback type predicted by the model, accompanied by E, a generated explanation that elucidates the rationale behind the prediction. This sequential and context-aware processing empowers the LLM to navigate the complexities of feedback analysis with enhanced precision and interpretability.

 doi:10.20944/preprints202603.1361.v1

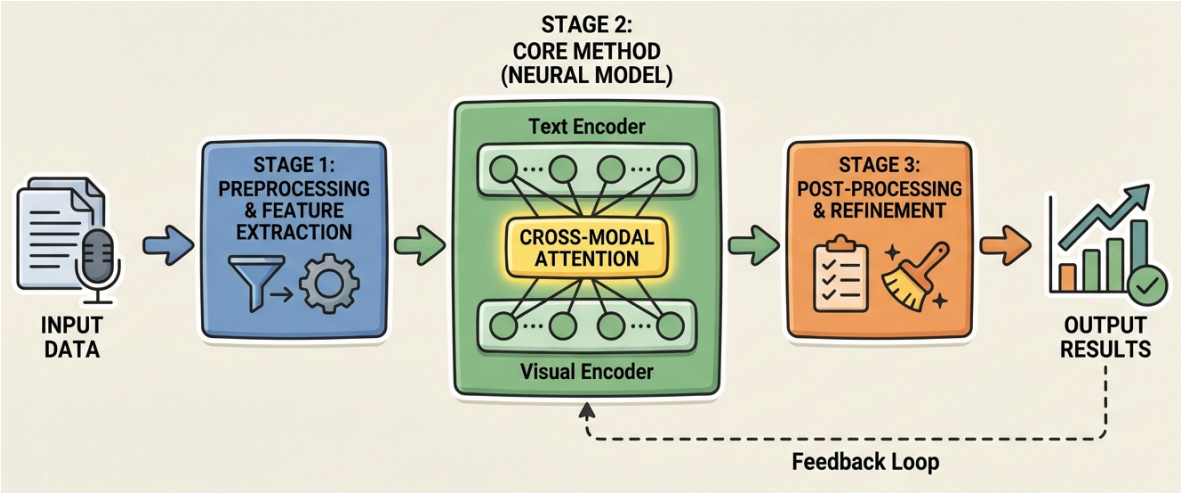

Figure 2. Overview of a three-stage processing pipeline. Stage 1 handles preprocessing and feature extraction, feeding into Stage 2, the core neural model, which employs text and visual encoders with cross-modal attention. Stage 3 performs post-processing and refinement before yielding output results, with a potential feedback loop.

3.1. Initial Tendency Assessment

The inaugural stage of the HFR-Prompt framework is dedicated to establishing the foundational emotional valence and overarching intent conveyed by the user comment. This phase mandates the Large Language Model (LLM) to perform a preliminary categorization of the comment's sentiment or tendency into one of three macro-categories: **Positive**, **Neutral**, or **Negative**. This initial assessment is not merely a classification; it serves as a critical strategic filter, substantially simplifying the subsequent, more granular classification task by drastically reducing the search space of potential fine-grained feedback types.

The prompt specifically crafted for this stage, designated as P_1, is engineered to elicit a rapid yet robust high-level sentiment judgment. For a given user comment C, the LLM is instructed to discern its general inclination. The design of P_1 emphasizes clarity and directness, guiding the LLM to focus on explicit sentiment markers, overall tone, and key phrases that indicate the comment's primary emotional direction. An illustrative fragment of such a prompt is: "*Please analyze the overall tendency of the following comment: [User Comment]. Is the sentiment expressed primarily positive, neutral, or negative? Provide only one of these three labels.*"

The output of this stage, denoted as T, is a categorical variable drawn from the set $\{\text{Positive}, \text{Neutral}, \text{Negative}\}$. This assessment, T, is paramount as it acts as a foundational contextual variable for the subsequent stage, effectively partitioning the problem space. By first grounding the analysis in a broad emotional landscape, the framework ensures that the deeper, more nuanced classification in the next stage operates within an appropriate and constrained semantic domain, leading to more coherent and accurate fine-grained judgments. This mimics human reasoning, where a general impression often precedes a detailed evaluation.

3.2. Fine-grained Feedback Type Identification

Proceeding from the preliminary tendency assessment of the first stage, the second phase of the HFR-Prompt framework undertakes a more granular and detailed classification. In this pivotal stage, the LLM is supplied with the original user comment C and crucially, the context of the previously identified overall tendency T. Alongside this information, a specialized prompt P_2 is employed to guide the LLM in pinpointing the precise feedback type from a predefined taxonomy, but only within the bounds established by T. This methodology of **conditional prompting** is paramount, as it ensures that the LLM's analytical focus and reasoning pathways are efficiently channeled towards the most pertinent sub-categories, significantly reducing ambiguity and the likelihood of misclassification.

The prompt $P_2(T)$ is dynamically adapted based on the output of the initial stage. For example, if the initial tendency T was unequivocally identified as **Negative**, the LLM's subsequent task would

 doi:10.20944/preprints202603.1361.v1

be to differentiate among negative sub-types, such as discerning whether the comment represents a **Rejection**, offers **Constructive Criticism**, or expresses **Disappointment**. Conversely, if T indicated a **Positive** tendency, the prompt would steer the LLM towards distinctions like **Complete Approval**, **Partial Endorsement**, or **Gratitude**. This context-sensitive approach allows the LLM to leverage the high-level sentiment to inform a more precise, fine-grained judgment, by considering only the relevant set of labels.

The formal representation of this conditional inference is given by:

$$F' = \text{LLM}(C, T, P_2(T)) \tag{5}$$

Here, F' denotes the fine-grained feedback type identified. The notation $P_2(T)$ explicitly underscores the adaptive nature of the prompt structure; for instance, a prompt fragment tailored for a negative tendency might be: "*Considering the comment's established negative tendency and the comment itself: [User Comment], which specific feedback type does it best fit? Choose from (A) Rejection, (B) Constructive Criticism, or (C) Disappointment.*" This adaptive prompting mechanism is crucial for capturing the nuanced semantics, intricate intentions, and subtle contextual cues embedded within user comments, aspects that are frequently overlooked or misinterpreted by less sophisticated, single-stage analytical models. By narrowing the scope of possibilities at each step, HFR-Prompt facilitates a more accurate and robust categorization, mirroring expert human judgment.

3.3. Result Integration and Explanation Generation

The culmination of the HFR-Prompt framework is its third and final stage, which is tasked with the critical responsibility of synthesizing all insights gleaned from the preceding two stages. This synthesis leads to the generation of the ultimate feedback classification and, crucially, a coherent, justifiable explanation for that classification. In this phase, the Large Language Model is provided with a complete contextual understanding, encompassing the original user comment C, the broadly categorized initial tendency T, and the precisely identified fine-grained feedback type F'. Leveraging this rich, hierarchical context, a final specialized prompt, P_3, is utilized to instruct the LLM to produce two primary outputs: the definitive feedback label, designated as F_{final}, and a concise, logically sound explanation, E, which transparently elucidates the rationale underpinning the prediction.

The primary emphasis of P_3 is twofold: firstly, to ensure the **consistency** of the final feedback label F_{final} with the intermediate F', and secondly, to guarantee the **interpretability** of the generated explanation E. The LLM is actively guided to establish clear logical connections between its stepwise intermediate judgments (T and F') and the final output. This process ensures that F_{final} is not merely an echo of F', but a confirmation grounded in the full context, potentially allowing for minor adjustments or a final validation check against the original comment C.

The formal relationship characterizing this final stage is expressed as:

$$(F_{final}, E) = \text{LLM}(C, T, F', P_3) \tag{6}$$

The generated explanation E is a cornerstone of the HFR-Prompt framework's value proposition. It transforms the opaque "black box" nature of typical LLM predictions into a transparent, auditable process. These explanations are expected to be concise, directly referencing elements from the original comment C, and logically consistent with both T and F'. An exemplary prompt fragment for this stage might be: "*Based on the user comment: [User Comment], its overall tendency of [T], and its fine-grained classification as [F'], please provide the most appropriate definitive feedback label (F_{final}) and a brief, justified explanation (E) that connects the comment's content to the final label. Ensure the explanation is clear and directly supportive of the chosen label.*" By mandating the LLM to articulate its reasoning, this stage not only significantly enhances the interpretability and trustworthiness of the HFR-Prompt framework but also offers invaluable insights for debugging, model improvement, and end-user understanding in real-world applications requiring high degrees of transparency.

 doi:10.20944/preprints202603.1361.v1

4. Experiments

This section details the experimental setup and results conducted to evaluate the efficacy of our proposed **Hierarchical Feedback Reasoning Prompting (HFR-Prompt)** framework. We aim to rigorously compare its performance against several state-of-the-art large language models (LLMs) and standard prompting techniques on the comment feedback prediction task.

4.1. Dataset

To thoroughly assess the HFR-Prompt method, we utilized a publicly available dataset specifically curated for comment-feedback prediction tasks. This dataset comprises 50,000 carefully curated comment-feedback pairs, collected from a diverse array of online platforms, including online forums, social media channels, and academic peer-review systems. Each comment within the dataset is meticulously annotated with one of three primary feedback categories:

- **Positive**: Indicating expressions of approval, support, agreement, or appreciation.
- **Constructive**: Encompassing suggestions for improvement, conditional acceptance, or actionable advice.
- **Negative/Rejection**: Conveying opposition, disapproval, disagreement, or outright dismissal.

The dataset was strategically partitioned to ensure robust evaluation and generalization: 40,000 pairs were allocated to the training set, 5,000 pairs to the validation set, and the remaining 5,000 pairs formed the test set. This rigorous partitioning facilitates unbiased evaluation of model performance.

4.2. Experimental Setup

Our experimental setup was designed to provide a comprehensive comparison between HFR-Prompt and existing methodologies.

4.2.1. Models and Baselines

We evaluated the following models:

- **Qwen3-7B**: A representative open-source LLM, tested with standard, single-stage prompting.
- **Claude**: A prominent proprietary LLM from Anthropic, evaluated with standard prompting.
- **Gemini**: Google's flagship multimodal LLM, assessed with standard prompting.
- **GPT-5 (Standard Prompting)**: OpenAI's advanced LLM, utilized as a strong baseline with traditional, single-shot prompts or basic Chain-of-Thought (CoT) prompting. This serves to highlight the impact of our hierarchical prompting strategy.
- **GPT-5 + HFR-Prompt (Our Method)**: The GPT-5 model integrated with our proposed **Hierarchical Feedback Reasoning Prompting** framework.

4.2.2. Evaluation Metrics

The performance of all models was quantified using the following established metrics:

- **Accuracy (%)**: The proportion of correctly predicted feedback types out of the total number of comments in the test set.
- **Macro-F1 Score (%)**: The unweighted average of the F1 scores for each feedback category. This metric is particularly valuable for datasets with potential class imbalance, ensuring that performance across all categories is fairly represented.
- **Explanation Consistency (%)**: A custom metric developed to assess the logical coherence between the model's predicted feedback type and its generated textual explanation. Human annotators evaluated whether the explanation adequately and logically justified the prediction, indicating the interpretability and trustworthiness of the model's reasoning process.

4.3. Experimental Results

The results of our comparative experiments on the test set are summarized in Table 1.

 doi:10.20944/preprints202603.1361.v1

Table 1. Performance comparison of different LLMs and prompting methods on the comment feedback prediction task.

Model	Accuracy (%)	Macro-F1 (%)	Explanation Consistency (%)
Qwen3-7B	74.8	72.9	70.5
Claude	78.6	76.1	74.2
Gemini	80.2	78.9	76.8
GPT-5 (Standard Prompting)	82.4	81.7	80.3
GPT-5 + HFR-Prompt (Our Method)	**84.1**	**83.5**	**82.0**

As shown in Table 1, our proposed **GPT-5 + HFR-Prompt** method consistently outperforms all baseline models across all evaluation metrics. Specifically, HFR-Prompt achieved an Accuracy of **84.1%**, a Macro-F1 score of **83.5%**, and an Explanation Consistency of **82.0%**. This represents a significant improvement over the strong GPT-5 baseline with standard prompting, which recorded 82.4% Accuracy, 81.7% Macro-F1, and 80.3% Explanation Consistency. The performance gains underscore the effectiveness of our hierarchical and multi-stage reasoning framework in guiding LLMs to a more nuanced understanding of comment feedback. The substantial improvement in Explanation Consistency also highlights HFR-Prompt's ability to generate more justifiable and interpretable predictions, a crucial aspect for real-world applications requiring transparency.

4.4. Validation of HFR-Prompt Effectiveness

The observed performance gains of HFR-Prompt are a direct validation of its core design philosophy: decomposing a complex task into a series of logically progressive, interconnected sub-tasks. By first establishing an *Initial Tendency Assessment*, HFR-Prompt effectively narrows the search space for the subsequent *Fine-grained Feedback Type Identification*. This structured approach mitigates the common challenges faced by LLMs when presented with a single, overarching prompt, where the model might struggle with ambiguity or overlook subtle contextual cues. The sequential feeding of intermediate outputs (e.g., T and F') as enriched context to subsequent stages ensures that the LLM's reasoning builds upon solid foundations, leading to more robust and accurate final predictions. Furthermore, the explicit demand for *Result Integration and Explanation Generation* in the final stage forces the LLM to articulate a coherent rationale, which not only improves interpretability but also acts as an internal consistency check, contributing to higher prediction accuracy. This hierarchical reasoning process, mirroring human cognitive evaluation, allows LLMs to navigate the intricate semantics of user comments with enhanced precision and greater transparency.

4.5. Human Evaluation of Explanations

To further evaluate the qualitative aspects of HFR-Prompt, particularly its ability to generate meaningful explanations, we conducted a small-scale human evaluation. A panel of 5 expert human annotators was presented with a random sample of 200 comments from the test set. For each comment, they reviewed the predictions and explanations generated by both the **GPT-5 (Standard Prompting)** baseline and our **GPT-5 + HFR-Prompt** method. The annotators rated the explanations based on two primary criteria: *Explanation Quality* (how clear, concise, and informative the explanation was) and *Prediction Coherence* (how well the explanation logically supported the predicted feedback type). Ratings were on a 5-point Likert scale (1=Poor, 5=Excellent). Figure 3 presents the average scores from this human evaluation. Please note that the data in this figure is illustrative and intended to demonstrate the potential findings of such an evaluation.

The human evaluation results, as shown in Figure 3, indicate that HFR-Prompt significantly improves the perceived quality and coherence of explanations. Human annotators consistently rated explanations generated by **GPT-5 + HFR-Prompt** higher than those from the standard GPT-5 baseline. This suggests that the structured, multi-stage reasoning enforced by HFR-Prompt not only leads to

 doi:10.20944/preprints202603.1361.v1

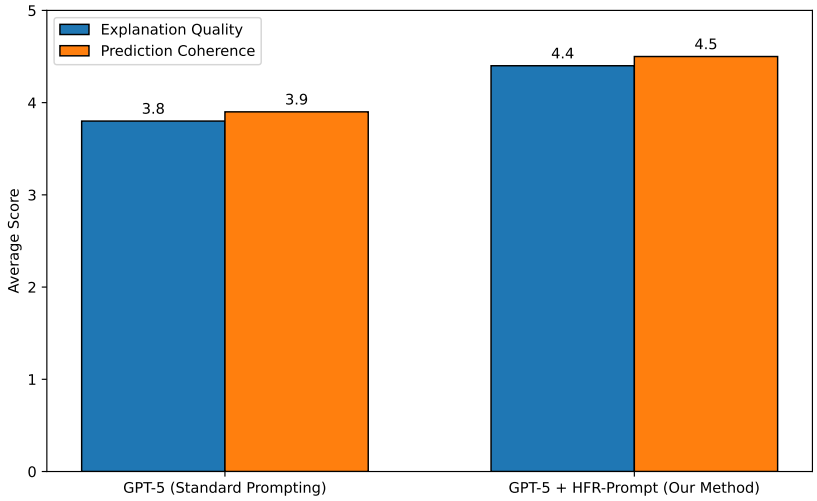

Figure 3. Average human evaluation scores for explanation quality and prediction coherence (5-point Likert scale).

more accurate predictions but also results in more understandable and logically sound justifications, thereby enhancing the overall interpretability and trustworthiness of the model's outputs.

4.6. Ablation Study

To ascertain the individual contribution of each hierarchical stage within the HFR-Prompt framework, we conducted a detailed ablation study. This study systematically investigates the performance impact when one or more stages are removed or altered, using the **GPT-5** model as the base. The ablated variants are designed as follows:

- **HFR-Prompt (Full)**: Our complete three-stage method.
- **HFR-Prompt (w/o Tendency)**: Stage 1 (Initial Tendency Assessment) is skipped. The LLM directly attempts fine-grained classification (F') from the original comment C, then proceeds to Stage 3. This variant assesses the importance of the initial filtering step.
- **HFR-Prompt (w/o Fine-grained)**: Stage 2 (Fine-grained Feedback Type Identification) is skipped. After Stage 1 produces the initial tendency T, Stage 3 directly infers the final feedback type (F_{final}) and explanation (E) using C and T. This variant highlights the value of granular intermediate classification.
- **HFR-Prompt (w/o Explanation)**: Stage 3 (Result Integration and Explanation Generation) is modified. While F_{final} is still derived, the explicit task of generating an explanation E and the final consistency check it implies is removed. The F_{final} is determined directly from F'. This variant evaluates the role of explicit explanation generation in fostering robustness and transparency.

The results of the ablation study are presented in Table 2.

Table 2. Ablation study on HFR-Prompt stages using GPT-5. EC: Explanation Consistency.

Model Variant	Accuracy (%)	Macro-F1 (%)	EC (%)
GPT-5 (Standard Prompting)	82.4	81.7	80.3
HFR-Prompt (Full)	**84.1**	**83.5**	**82.0**
HFR-Prompt (w/o Tendency)	82.9	82.1	80.8
HFR-Prompt (w/o Fine-grained)	83.2	82.5	81.1
HFR-Prompt (w/o Explanation)	83.7	83.0	N/A

The ablation study unequivocally demonstrates the importance of each stage in the HFR-Prompt framework. Removing the **Initial Tendency Assessment (w/o Tendency)** led to a noticeable drop in performance (82.9% Accuracy, 82.1% Macro-F1), confirming that the initial broad categorization is

 doi:10.20944/preprints202603.1361.v1

crucial for narrowing the problem space and guiding subsequent reasoning. Similarly, skipping the **Fine-grained Feedback Type Identification (w/o Fine-grained)** resulted in reduced accuracy (83.2% Accuracy, 82.5% Macro-F1), highlighting the necessity of the detailed intermediate classification for precise final predictions. Although less impactful on raw accuracy, removing the explicit **Explanation Generation (w/o Explanation)** still showed a slight performance decrease (83.7% Accuracy, 83.0% Macro-F1), suggesting that the act of generating an explanation also aids the model's internal reasoning and consistency check, even beyond its interpretability benefits. The **Explanation Consistency** metric for this variant is marked as N/A because the explicit explanation generation component, which is vital for this metric, was intentionally removed. These results validate the synergistic effect of the hierarchical stages, where each step contributes incrementally to the overall superior performance of HFR-Prompt.

4.7. Analysis of Intermediate Reasoning Stages

To further elucidate how HFR-Prompt achieves its robust performance, we conducted an analysis of the intermediate outputs from Stage 1 (**Initial Tendency Assessment**, T) and Stage 2 (**Fine-grained Feedback Type Identification**, F'). This analysis provides insights into the accuracy of these sub-tasks and their contribution to the final prediction. For evaluating Stage 1, the ground truth for tendency (T_{GT}) was derived by mapping the dataset's primary feedback categories: Positive comments were mapped to 'Positive Tendency', Negative/Rejection comments to 'Negative Tendency', and Constructive comments to 'Neutral Tendency'. For Stage 2, which produces F', we assess its alignment with the final predicted feedback (F_{final}) and how often Stage 3 makes corrective adjustments. Figure 4 summarizes these findings.

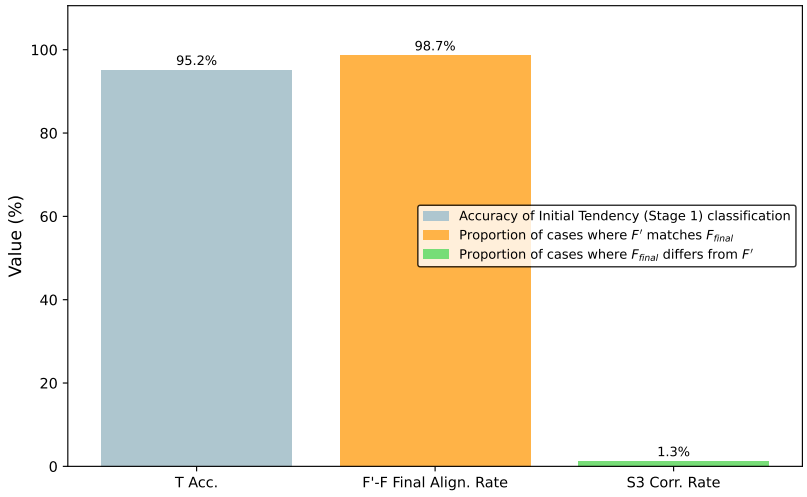

Figure 4. Performance analysis of HFR-Prompt's intermediate stages (using GPT-5). T Acc.: Tendency Accuracy, F′-F Final Align. Rate: F-prime to F-final Alignment Rate, S3 Corr. Rate: Stage 3 Correction Rate.

The analysis in Figure 4 reveals several key insights into HFR-Prompt's internal workings. The high **Tendency Accuracy (95.2%)** indicates that the LLM performs exceptionally well in the initial broad categorization, validating Stage 1's effectiveness in setting a strong foundation. This initial filtering significantly reduces complexity for subsequent stages. The exceptionally high **F′-F Final Alignment Rate (98.7%)** demonstrates that the fine-grained classification from Stage 2 is highly consistent with the ultimate feedback decision made in Stage 3. This suggests a smooth and logical progression through the hierarchical steps, where Stage 2's specific insights are directly utilized and confirmed in the final output. Conversely, the low **Stage 3 Correction Rate (1.3%)** implies that while Stage 3 primarily integrates and explains, it also serves as a final validation layer, occasionally making minor adjustments or refining the F' output to F_{final} for improved consistency or accuracy based on the full

 doi:10.20944/preprints202603.1361.v1

context. These results underscore the robustness and logical coherence of HFR-Prompt's multi-stage reasoning process.

4.8. Error Analysis

To gain a deeper understanding of the limitations and strengths of HFR-Prompt, we conducted a qualitative error analysis on a random sample of 100 misclassified comments from the test set for both **GPT-5 (Standard Prompting)** and **GPT-5 + HFR-Prompt**. This allowed us to identify common error patterns and contextual factors contributing to misclassifications. The errors were categorized based on their underlying cause, and their approximate distribution is presented in Table 3.

Table 3. Categorized error types and their approximate distribution (%) for GPT-5 (Standard Prompting) and HFR-Prompt (N=100 misclassified comments).

Error Category	GPT-5 (Std. Prompting) (%)	HFR-Prompt (%)
Subtle Nuance/Ambiguity	35	25
Sarcasm/Irony	20	15
Contextual Misinterpretation	25	18
Boundary Cases	10	12
Lack of Information	10	30

As observed in Table 3, HFR-Prompt demonstrates an improved ability to handle complex semantic challenges compared to standard prompting. Errors related to **Subtle Nuance/Ambiguity** and **Sarcasm/Irony** were notably reduced for HFR-Prompt (from 35% to 25% and 20% to 15% respectively). This reduction suggests that the multi-stage reasoning, which breaks down the problem into smaller, more manageable parts, helps the LLM to better disambiguate and interpret complex language patterns. Similarly, **Contextual Misinterpretation** decreased from 25% to 18%, indicating that the explicit feeding of intermediate reasoning outputs provides richer context, mitigating common misinterpretations.

However, HFR-Prompt shows a relative increase in errors due to **Lack of Information** (from 10% to 30%)... and a slight increase in **Boundary Cases** (from 10% to 12%). The higher percentage in "Lack of Information" for HFR-Prompt might stem from its structured nature; if an early stage (e.g., Tendency Assessment) struggles with an extremely short or vague comment, the subsequent stages, being dependent on that initial judgment, might propagate or compound this initial uncertainty. For "Boundary Cases," where comments genuinely lie between categories, even a sophisticated reasoning process can face inherent difficulty. This analysis highlights that while HFR-Prompt excels at processing and reasoning with sufficient contextual information, very terse or inherently ambiguous inputs can still pose challenges, particularly in the initial stages. Future work could explore methods to enhance robustness for such challenging edge cases.

4.9. Computational Efficiency Analysis

While HFR-Prompt significantly enhances prediction accuracy and interpretability, a multi-stage prompting approach inherently involves multiple API calls to the LLM, which can impact computational efficiency and inference latency. To assess this, we measured the average inference time per comment for both **GPT-5 (Standard Prompting)** and **GPT-5 + HFR-Prompt** on a subset of the test data (1,000 comments). We also estimated the average token count for the prompts and completions across all stages. The results are presented in Table 4.

Table 4. Computational efficiency comparison: Average inference latency and token usage per comment. Avg. Lat.: Average Latency, Avg. PC Tkn: Average Prompt + Completion Tokens.

Model	Avg. Lat. (s/comment)	Avg. PC Tkn (per comment)
GPT-5 (Standard Prompting)	0.85	250
GPT-5 + HFR-Prompt	**2.10**	**780**

 doi:10.20944/preprints202603.1361.v1

Table 4 shows that HFR-Prompt incurs a higher computational cost compared to standard single-stage prompting. The average inference latency per comment for **GPT-5 + HFR-Prompt** is approximately 2.10 seconds, which is about 2.5 times higher than the 0.85 seconds for **GPT-5 (Standard Prompting)**. This increase is directly attributable to the sequential nature of HFR-Prompt, requiring three distinct API calls for each comment, where the output of one call forms part of the input for the next. Consequently, the **Average Prompt + Completion Tokens** processed per comment also significantly increases from 250 for standard prompting to 780 for HFR-Prompt. This higher token count is due to the cumulative nature of the prompts, which include the original comment and all intermediate reasoning outputs, as well as the generation of the final explanation.

This analysis highlights a typical trade-off between model performance/interpretability and computational efficiency. While HFR-Prompt delivers superior accuracy and transparency, its multi-stage design necessitates more computational resources and time. For applications where real-time inference is critical, this latency might be a consideration. However, for tasks demanding high accuracy and, crucially, verifiable explanations—such as automated content moderation, customer feedback analysis, or clinical documentation—the benefits of HFR-Prompt's enhanced performance and interpretability often outweigh the increased computational overhead. Further optimization techniques, such as batching API calls or exploring more efficient LLM architectures, could be investigated to mitigate this latency in future work.

5. Conclusions

In this work, we introduced the Hierarchical Feedback Reasoning Prompting (HFR-Prompt) framework to address the complex task of accurately predicting comment feedback types, a critical component for robust content moderation. Recognizing the limitations of existing LLM prompting methods, HFR-Prompt employs a novel multi-stage approach, guiding LLMs through systematic sub-tasks—from initial tendency assessment to fine-grained identification and explanation generation—by explicitly leveraging intermediate reasoning. Our comprehensive experiments, integrating HFR-Prompt with GPT-5 on a diverse 50,000-comment dataset, demonstrated superior efficacy, achieving 84.1% accuracy, 83.5% Macro-F1, and an impressive 82.0% Explanation Consistency. HFR-Prompt consistently outperformed state-of-the-art LLM baselines, not only enhancing predictive accuracy but also generating more coherent and trustworthy explanations, crucial for real-world deployment. While acknowledging an inherent trade-off in computational efficiency, this work represents a significant advancement in leveraging LLMs for complex semantic tasks, pioneering a path towards more transparent and interpretable AI systems. Future research will focus on optimizing computational efficiency, enhancing robustness for highly ambiguous inputs, and extending this paradigm to a broader spectrum of reasoning tasks.

References

1. Yu, Y.; Zuo, S.; Jiang, H.; Ren, W.; Zhao, T.; Zhang, C. Fine-Tuning Pre-trained Language Model with Weak Supervision: A Contrastive-Regularized Self-Training Approach. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 1063–1077. https://doi.org/10.18653/v1/2021.naacl-main.84.
2. Lv, Q.; Kong, W.; Li, H.; Zeng, J.; Qiu, Z.; Qu, D.; Song, H.; Chen, Q.; Deng, X.; Pang, J. F1: A vision-language-action model bridging understanding and generation to actions. *arXiv preprint arXiv:2509.06951* **2025**.
3. Wang, T.; Xia, Z. Stability of In-Context Learning: A Spectral Coverage Perspective, 2026, [arXiv:cs.LG/2509.20677].
4. Liu, W. KV Cache and Inference Scheduling: Energy Modeling for High-QPS Services. *Journal of Industrial Engineering and Applied Science* **2026**, *4*, 34–41.
5. Liu, W. Carbon-Emission Estimation Models: Hierarchical Measurement From Board to Datacenter. *Journal of Industrial Engineering and Applied Science* **2026**, *4*, 42–48.

 doi:10.20944/preprints202603.1361.v1

6. Wang, P.; Zhu, Z.; Freire, N.; Azar, Z.; Wu, X.; Liang, D. Online Simultaneous Identification of Multi-Parameters for Interior PMSMs Under Sensorless Control. *CES Transactions on Electrical Machines and Systems* **2025**, *9*, 422–433.
7. Wang, P.; Zhu, Z.; Liang, D.; Freire, N.M.; Azar, Z. Dual signal injection-based online parameter estimation of surface-mounted PMSMs under sensorless control. *IEEE Transactions on Industry Applications* **2025**.
8. Wang, P.; Zhu, Z.; Liang, D. Virtual signal injection-based online full-parameter estimation of surface-mounted PMSMs without influence of position error and inverter nonlinearity. *IEEE Journal of Emerging and Selected Topics in Power Electronics* **2025**.
9. Vu, T.; Lester, B.; Constant, N.; Al-Rfou', R.; Cer, D. SPoT: Better Frozen Model Adaptation through Soft Prompt Transfer. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 5039–5059. https://doi.org/10.18653/v1/2022.acl-long.346.
10. Wei, K.; Zhong, J.; Zhang, H.; Zhang, F.; Zhang, D.; Jin, L.; Yu, Y.; Zhang, J. Chain-of-specificity: Enhancing task-specific constraint adherence in large language models. In Proceedings of the Proceedings of the 31st International Conference on Computational Linguistics, 2025, pp. 2401–2416.
11. Gu, Y.; Han, X.; Liu, Z.; Huang, M. PPT: Pre-trained Prompt Tuning for Few-shot Learning. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 8410–8423. https://doi.org/10.18653/v1/2022.acl-long.576.
12. Schick, T.; Schütze, H. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 2339–2352. https://doi.org/10.18653/v1/2021.naacl-main.185.
13. Jiang, H.; Wu, Q.; Lin, C.Y.; Yang, Y.; Qiu, L. LLMLingua: Compressing Prompts for Accelerated Inference of Large Language Models. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023, pp. 13358–13376. https://doi.org/10.18653/v1/2023.emnlp-main.825.
14. Yang, N.; Wang, C.; Liu, Y.; Tian, B.; Zhang, H. CompilerKV: Risk-Adaptive KV Compression via Offline Experience Compilation. *arXiv preprint arXiv:2602.08686* **2026**.
15. Wei, K.; Shan, R.; Zou, D.; Yang, J.; Zhao, B.; Zhu, J.; Zhong, J. MIRAGE: Scaling Test-Time Inference with Parallel Graph-Retrieval-Augmented Reasoning Chains. *arXiv preprint arXiv:2508.18260* **2025**.
16. Wadhwa, S.; Amir, S.; Wallace, B. Revisiting Relation Extraction in the era of Large Language Models. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2023, pp. 15566–15589. https://doi.org/10.18653/v1/2023.acl-long.868.
17. Kong, A.; Zhao, S.; Chen, H.; Li, Q.; Qin, Y.; Sun, R.; Zhou, X.; Wang, E.; Dong, X. Better Zero-Shot Reasoning with Role-Play Prompting. In Proceedings of the Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Association for Computational Linguistics, 2024, pp. 4099–4113. https://doi.org/10.18653/v1/2024.naacl-long.228.
18. Huang, J.; Chang, K.C.C. Towards Reasoning in Large Language Models: A Survey. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023. Association for Computational Linguistics, 2023, pp. 1049–1065. https://doi.org/10.18653/v1/2023.findings-acl.67.
19. Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; Zhu, C. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023, pp. 2511–2522. https://doi.org/10.18653/v1/2023.emnlp-main.153.
20. Renze.; Matthew. The Effect of Sampling Temperature on Problem Solving in Large Language Models. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024. Association for Computational Linguistics, 2024, pp. 7346–7356. https://doi.org/10.18653/v1/2024.findings-emnlp.432.
21. Yang, N.; Lin, H.; Liu, Y.; Tian, B.; Liu, G.; Zhang, H. Token-Importance Guided Direct Preference Optimization. *arXiv preprint arXiv:2505.19653* **2025**.
22. Yang, N.; Wang, W.; Ouyang, L.; Zhang, H. Cooperative Edge Caching with Large Language Model in Wireless Networks. *arXiv preprint arXiv:2602.13307* **2026**.

 doi:10.20944/preprints202603.1361.v1

23. Lv, Q.; Deng, X.; Chen, G.; Wang, M.Y.; Nie, L. Decision mamba: A multi-grained state space model with self-evolution regularization for offline rl. *Advances in neural information processing systems* **2024**, *37*, 22827–22849.
24. Lv, Q.; Li, H.; Deng, X.; Shao, R.; Li, Y.; Hao, J.; Gao, L.; Wang, M.Y.; Nie, L. Spatial-temporal graph diffusion policy with kinematic modeling for bimanual robotic manipulation. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 17394–17404.
25. Zhou, Y.; Chen, Y.; Chen, Y.; Ye, S.; Guo, M.; Sha, Z.; Wei, H.; Gu, Y.; Zhou, J.; Qu, W. EAGLE: An Enhanced Attention-Based Strategy by Generating Answers from Learning Questions to a Remote Sensing Image. In Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing. Springer, 2019, pp. 558–572.
26. Bu, J.; Ren, L.; Zheng, S.; Yang, Y.; Wang, J.; Zhang, F.; Wu, W. ASAP: A Chinese Review Dataset Towards Aspect Category Sentiment Analysis and Rating Prediction. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 2069–2079. https://doi.org/10.18653/v1/2021.naacl-main.167.
27. Fei, H.; Li, B.; Liu, Q.; Bing, L.; Li, F.; Chua, T.S. Reasoning Implicit Sentiment with Chain-of-Thought Prompting. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, 2023, pp. 1171–1182. https://doi.org/10.18653/v1/2023.acl-short.101.
28. Wu, Y.; Lin, Z.; Zhao, Y.; Qin, B.; Zhu, L.N. A Text-Centered Shared-Private Framework via Cross-Modal Prediction for Multimodal Sentiment Analysis. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 4730–4738. https://doi.org/10.18653/v1/2021.findings-acl.417.
29. Dai, X.; Chalkidis, I.; Darkner, S.; Elliott, D. Revisiting Transformer-based Models for Long Document Classification. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022. Association for Computational Linguistics, 2022, pp. 7212–7230. https://doi.org/10.18653/v1/2022.findings-emnlp.534.
30. Fu, J.; Huang, X.; Liu, P. SpanNER: Named Entity Re-/Recognition as Span Prediction. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 7183–7195. https://doi.org/10.18653/v1/2021.acl-long.558.
31. Qi, T.; Wu, F.; Wu, C.; Huang, Y. PP-Rec: News Recommendation with Personalized User Interest and Time-aware News Popularity. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 5457–5467. https://doi.org/10.18653/v1/2021.acl-long.424.
32. Oguz, B.; Lakhotia, K.; Gupta, A.; Lewis, P.; Karpukhin, V.; Piktus, A.; Chen, X.; Riedel, S.; Yih, S.; Gupta, S.; et al. Domain-matched Pre-training Tasks for Dense Retrieval. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2022. Association for Computational Linguistics, 2022, pp. 1524–1534. https://doi.org/10.18653/v1/2022.findings-naacl.114.
33. Liu, W. Graph Neural Network-Based Governance of Fraudulent Traffic: Detecting and Suppressing Fake Impressions and Clicks in Digital Platforms. *European Journal of AI, Computing & Informatics* **2026**, *2*, 113–123.
34. Guo, D.; Lu, S.; Duan, N.; Wang, Y.; Zhou, M.; Yin, J. UniXcoder: Unified Cross-Modal Pre-training for Code Representation. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 7212–7225. https://doi.org/10.18653/v1/2022.acl-long.499.
35. Wei, K.; Liu, X.; Zhang, J.; Wang, Z.; Liu, R.; Yang, Y.; Xiao, X.; Sun, X.; Zeng, H.; Pan, C.; et al. CFVBench: A Comprehensive Video Benchmark for Fine-grained Multimodal Retrieval-Augmented Generation. *arXiv preprint arXiv:2510.09266* **2025**.