Beyond the Retweet: A Large-Scale Dataset of Reader Emotion, Propagation Intent, and Behavior

Yujin Zheng[1], **Jian Liao**[1,3*], **Suge Wang**[2], **Jianxing Zheng**[2], **Deyu Li**[2], **Qian Chen**[1]
[1]School of Computer and Information Technology, Shanxi University, China
[2]Institute of Intelligent Information Processing, Shanxi University, China
[3] Joint Laboratory of Tourism Big Data in Shanxi Province, Taiyuan, China
Correspondence: liaoj@sxu.edu.cn

Abstract

Current emotion analysis in social media is predominantly author-centric, failing to capture the subjective nature of emotional responses across diverse readers. This paradigm overlooks the crucial link between individual perception, implicit propagation intent, propagation behavior, and the underlying social network. To bridge this gap, we introduce EmoIntenNet, a novel bilingual (English and Chinese) large-scale benchmark that uniquely integrates six critical dimensions for social perception: author-created content, actual readers' emotional feedback (derived from their comments), implicit propagation intent, explicit propagation behavior, user attributes, and the social relationship graph. This benchmark advances the shift towards more granular, reader-centric analysis, where different readers' emotional responses to the same content are grounded in their actual social interactions rather than simulated annotations. By annotating implicit emotions from reader comments and synchronously capturing implicit multi-label propagation intents, EmoIntenNet provides a unique resource to model the intrinsic triadic coupling among implicit emotion, implicit intent, and explicit behavior, grounded in social context. We establish a comprehensive evaluation protocol, testing state-of-the-art methods, including large language models (LLMs) with advanced reasoning enhancement. Our findings reveal significant shortcomings in existing approaches when handling this multifaceted, user-aware task. EmoIntenNet offers a foundational resource and clear direction for future research in socially-intelligent NLP, pushing models towards a more unified understanding of emotion on social media.

1 Introduction

Social media has become the primary arena for emotional expression and information propagation. The analysis of emotions within user-generated content is crucial for understanding public opinion and providing personalized services. However, a fundamental limitation persists: mainstream methodologies predominantly operate under an "author-centric" (Wei et al., 2025; Hu and Collier, 2025) or "content-centric" (Li et al., 2025; Jian et al., 2025) paradigm, assigning emotion labels to an entire post. This overlooks the highly subjective nature of emotion perception, failing to capture that the same content can evoke disparate emotional responses and subsequent behaviors among different readers. For instance, a discussion about an emerging technology product might elicit "excitement" and a propensity to share among tech enthusiasts, while triggering "anxiety" and critical comments from skeptical users (as illustrated in Fig. 1). This exposes the core blind spot: a single label detached from the "reader's perspective" cannot capture the true dynamics of emotion, nor can it explain the cognitive mechanisms through which internal affective states translate into specific external social actions.

To address this, annotating emotions on individual-specific reader comments preserves personalized responses, shifting analysis from post-level to comment-level. Unlike datasets with isolated post-comment pairs, EmoIntenNet links a single post to multiple genuine reader comments. Crucially, instead of simulating reader emotions, EmoIntenNet leverages natural interactions: interacting users are treated as "real readers," with labels based on their genuine responses. This grounds personalization in authentic behavior, linking user profiles to emotional reactions.

Although recent studies have begun to transcend conventional paradigms, their limitations remain significant. Ye et al. (2025); Jun and Lee (2025) enhance personalized emotion understanding by incorporating author dynamics or user history, yet remain within an "author-centric" framework, lacking authentic reader feedback. Liao et al. (2025)

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 19506–19542
October 24-29, 2026 ©2026 Association for Computational Linguistics

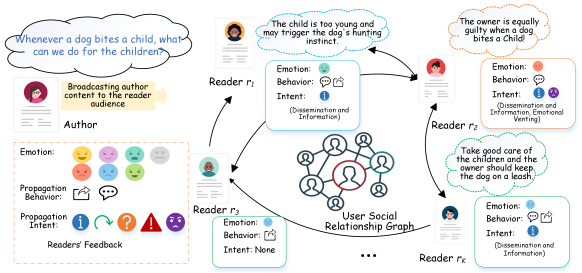


Figure 1: Illustration of personalized emotions and differential propagation behaviors evoked by the same social media content from different reader perspectives.

introduce an innovative approach integrating simulated reader agents, but treat emotion prediction and behavior simulation as separate tasks, failing to capture their intrinsic coupling. More critically, these approaches typically overlook a crucial mediating variable: implicit propagation intent. Specifically, propagation intent serves as the intrinsic reason that drives readers to engage in differential propagation behaviors. We distinguish this concept from prior emotional intent (Tu et al., 2026); while the latter explains why an emotion arises, the former explains why one engages in propagation behaviors such as reposting.

Without modeling this implicit intent, the causal chain from emotion to behavior remains incomplete, leaving models unable to explain why the same emotion leads to vastly different behaviors. Furthermore, most existing models treat users as isolated nodes (Lyu et al., 2023; Zhao et al., 2024), ignoring social media's relational nature. *Emotion Contagion Theory* (Chen et al., 2024) posits that emotions spread within groups, profoundly influenced by network structures. Consequently, without the social relationship graph, models cannot answer the core question: "How do emotions propagate and evolve within a community?"

These limitations lead to three interconnected core challenges:

(1) How can we achieve personalized emotion annotation from a real reader's perspective? Existing paradigms fail to capture inter-individual differences. A shift is required from labeling content itself to jointly annotating emotion and propagation intent for "user-content" interaction pairs.

(2) How can we effectively model the triadic coupling among implicit emotions, implicit propagation intents, and explicit propagation behaviors? A causal theory-driven framework is required to map these internal states to concrete social choices, elucidating how emotion and intent jointly drive behavioral decisions.

(3) How can we integrate user attributes with social structures to reveal group-level emotion dynamics? By embedding the social relationship graph with user profiles, we can trace propagation pathways and emotion evolution within complex networks.

To address these challenges, we construct and release EmoIntenNet, a large-scale, heterogeneous, bilingual (English and Chinese) benchmark from Weibo and Twitter. It integrates six dimensions of social perception: author content, reader-specific emotional feedback, implicit propagation intent, explicit propagation behavior, user attributes, and the social relationship graph. The main contributions are summarized as follows:

(1) We construct the EmoIntenNet dataset, integrating six core aspects (author, reader, intent, behavior, profiles, relationships), providing a solid foundation for studying emotion analysis and propagation modeling. The dataset is available at: https://github.com/sxu-nlp/EmoIntenNet.

(2) We introduce a novel reader-grounded annotation strategy that annotates emotions from actual reader comments, achieving truly personalized emotion analysis by capturing real emotional responses.

(3) We establish a fine-grained triadic mapping among implicit emotions, multi-label propagation intents, and explicit behaviors by synchronously annotating emotions and intents, offering a valuable resource for understanding how internal states drive external actions.

(4) We provide a comprehensive evaluation benchmark, systematically assessing SOTA methods, including LLMs with chain-of-reasoning and debate-based reasoning, to reveal current shortcomings and future directions.

2 Related Work

Sentiment/Emotion Datasets Standard sentiment/emotion analysis benchmarks such as SemEval-14 (Pontiki et al., 2014), SemEval-17 (Rosenthal et al., 2017), GoEmotions (Demszky et al., 2020), OATS-ABSA (Chebolu et al., 2024), M-ABSA (Wu et al., 2025), BRIGHTER (Muhammad et al., 2025), and XED (Öhman et al., 2020) provide large-scale annotations, yet they still follow an author-centric paradigm by assigning a single emotion label to each post and ignoring reader sub-

Dataset	Emotion Subject	Domain	Language	Intent	Behavior	User Profile	Social Graph	Scale	Applicable Tasks
SemEval-14	author	product reviews	EN	×	×	×	×	7.6k	ABSA
ECC	author	conversation	EN	×	×	×	×	2.4k	emotion cause extraction
SemEval-17	author	Twitter	EN, AR	×	×	×	×	72k	ABSA
GoEmotions	author	Reddit	EN	×	×	×	×	58k	ABSA
OATS-ABSA	author	social media	EN	×	×	×	×	44k	open-target ABSA
M-ABSA	author	social media	EN	×	×	×	×	14k	multimodal ABSA
BRIGHTER	author	social media	multiple	×	×	×	×	100k	sentiment analysis
XED	author	movie subtitles	multiple	×	×	×	×	55k	sentiment analysis
RAPPIE	author	social media	EN, CN	×	×	✓	CN only	29k	personalized emotion analysis
EmoBank	average reader	news	EN	×	×	×	×	10k	dimensional emotion analysis
MultiPICo	annotator	social media	multiple	×	×	506 annotators'	×	18.8k	irony detection
INEWS	annotator	news	EN	×	×	291 annotators'	×	2.9k	emotional respond simulation
EmoIntenNet	individual reader	social media	EN, CN	✓	✓	39349 real reader users	✓	113k	personalized emotion/intent/behavior analysis, emotion propagation

Figure 2: Comparison of EmoIntenNet with existing sentiment analysis datasets.

jectivity. Additionally, the WASSA-2017 shared task (Mohammad and Bravo-Marquez, 2017) introduced intensity scores for four basic emotions, also adopting an author-centric perspective. Although RAPPIE (Liao et al., 2025) explores personalization, it remains limited to a "content-centric" setting and fails to jointly annotate implicit emotions, explicit behaviors, and propagation intent. See Fig. 2 for an overview and Table 6 in Appendix A.1 for details.

Reader Perspective and Subjectivity Recent work has explored subjective annotation, as shown in Fig. 2. Early studies have investigated appraisal theories (Troiano et al., 2023) and perspectivist modeling (Basile et al., 2022; Xu and Jurgens, 2026). Unlike EmoBank (Buechel and Hahn, 2017), which evaluates emotions evoked in an average reader and thus captures only group-level feedback, EmoIntenNet is distinctive in achieving precise, individual-level reader emotion annotation. Datasets such as GoodNewsEveryone (Bostan et al., 2020), MultiPICo (Casola et al., 2024), and INEWS (Hu and Collier, 2025), although annotating from a reader's perspective, all require annotators to imagine or simulate the reader's potential feelings rather than drawing from real social interactions. EmoIntenNet leverages natural social interactions by identifying users who actively engage with a post—through commenting or forwarding—as "real readers," and integrates their content, emotions, behaviors, profiles, and social relationship graphs.

Propagation Intent and Propagation Motivation. Understanding *why* users disseminate content is central to propagation research. While Uses and Gratifications theory links sharing to specific needs (Xu et al., 2025; Hai and Xiong, 2025), and empirical studies show emotions fuel these motivations (Ali et al., 2022), these triggers are mediated by *implicit intentions*. Behavioral Reasoning Theory emphasizes that intentions act as the cognitive bridge translating reasons and impulses into concrete sharing choices (Kumar et al., 2023); even the decision not to share is a strategic choice driven by self-presentation intent (Ihm and mee Kim, 2025). However, current NLP benchmarks overlook this crucial *why* dimension, modeling engagement as mere explicit actions rather than choices shaped by underlying intent. Without modeling implicit propagation intent, models cannot explain why the same emotion leads to vastly different behaviors, leaving the causal chain from emotion to behavior fundamentally incomplete.

Personalization and Human Behavior Simulation. Effective personalization relies on user models, evolving from traditional features (Zhou et al., 2023) to persona-based LLMs like LaMP (Salemi et al., 2024) and PPlug (Liu et al., 2025), though often hindered by profile dependence and privacy concerns (Hellwig et al., 2025). Recently, LLMs have enabled new behavior simulation approaches, such as prompt-based techniques (Wang et al., 2025a), behavior-realism alignment (Wang et al., 2025c), and cognitive-behavioral fixation (Wang et al., 2025b). These methods have proven valuable for downstream tasks like recommendation systems (Bougie and Watanabe, 2025) and social simulation (Mou et al., 2024). However, evaluating these sophisticated models remains constrained by the lack of datasets that simultaneously provide personalization and rich behavioral context necessary for a comprehensive assessment of social intelligence.

3 Dataset Construction: EmoIntenNet

To address the three core challenges, we construct and release EmoIntenNet, a large-scale, heterogeneous, bilingual benchmark for studying emotion

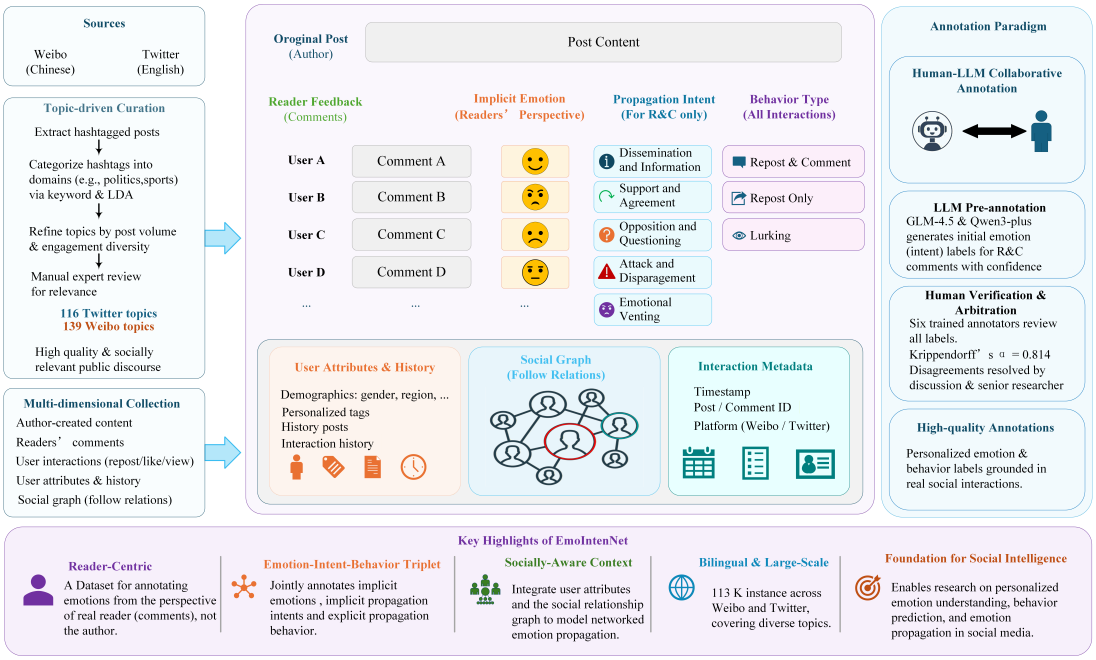


Figure 3: The overall architecture and construction process of EmoIntenNet.

propagation from a personalized and networked perspective. It operationalizes our theoretical commitments by integrating the reader's viewpoint, modeling the emotion-intent-behavior coupling, and embedding social structures. The overall architecture and construction process of EmoIntenNet are summarized in Fig. 3.

3.1 Data Collection and Curation

Our data collection employs a **topic-driven** curation strategy, ensuring every instance is grounded in authentic public discourse.

Sources and Scope. We collected data from Weibo (Chinese) and Twitter (English) to capture culturally diverse perspectives. The 2009–2023 temporal scope offers a dual advantage: ensuring the data is free from AI-generated content contamination (pre-LLM era), and enabling longitudinal analysis of emotion origin and propagation across socio-cultural developments.

Topic Filtering. To ensure relevance beyond keyword matching, we applied a rigorous multi-step pipeline combining algorithmic clustering, engagement filtering, and manual noise removal. This process retained 116 Twitter and 139 Weibo topics, ensuring EmoIntenNet is a curated archive of meaningful public conversations (see Appendix A.3 for parameters).

Multi-Dimensional Heterogeneity. To actualize our **six-dimensional** framework, we systematically gathered:

(1) Author-created Content: Original posts initiating discussions.

(2) Readers' Emotional Feedback: The textual comments from real users responding to the original posts. This is the core data unit for our reader-centric analysis. Collectively, they reflect the diverse emotional feedback of different users to the original post.

(3) Propagation Intent: Underlying propagation motivations identified from R&C interactions, formulated as multi-label annotations (detailed in Section 3.2).

(4) Propagation Behaviors: Explicit engagement categorized as *Repost & Comment* (R&C), *Repost Only* (R), and *Lurking* (L). Lurking is operationally defined by the structural absence of public interaction: specifically, a user is categorized as Lurking if they have a following relationship with the original author but exhibit neither Repost nor Comment.

(5) User Attributes: Desensitized profile information for all participating users (e.g., gender, region), along with historical posts.

(6) Social relationship Graph: "*following*" relationships to model the social network structure.

This approach preserves the inherent heterogeneity and relational context of social media interactions. Standard preprocessing (URL removal, tag cleaning, spam filtering) was applied to ensure data quality.

3.2 Emotion-Intent-Behavior Annotation Paradigm

To eliminate ambiguity, we first define our key actors: the **Author** creates the original post; the **Reader** is a real user who views and interacts with the post, possessing a unique profile and history; the **Annotator** is a trained human or LLM that infers states from these interactions.

The core innovation of EmoIntenNet is its annotation strategy modeling the triadic coupling among personalized emotional states, implicit intents, and explicit behaviors. Instead of annotating content in isolation, we annotate *Reader-Content* interaction pairs. This grounds personalization in authentic behavioral traces rather than annotator imagination: because different Readers possess distinct profiles and interaction histories, their emotional responses and behaviors to the same post naturally diverge. Each R&C interaction is annotated on three concurrent fronts:

(1) Personalized Emotion (Reader's Perspective): We classify the Reader's emotional reaction into seven categories (Ekman's six basic emotions (Ekman, 1992; Ekman and Friesen, 1971) plus *Neutral*). Crucially, annotators infer the emotion from the *Reader's own comment* text (not the post), and this emotion label is tightly coupled with that specific Reader's profile and history. While this is an inference rather than a self-report, it captures the manifested affective state of a *real individual*. Unlike prior datasets where annotators imagine a generic reader's reaction, we preserve disaggregated, user-specific responses (different Readers yielding different labels for the same post), aligning with perspectivist principles and achieving genuine personalization.

(2) Implicit Propagation Intent: To bridge emotion and action, we introduce multi-label intent annotation for R&C comments based on Behavioral Reasoning Theory (Westaby, 2005; Kumar et al., 2023). Five categories are used: *Dissemination and Information* (DI), *Support and Agreement* (SA), *Opposition and Questioning* (OQ), *Attack and Disparagement* (AD), and *Emotional Venting* (EV). For the specific meanings of propagation intent, see Appendix A.5. The multi-label formulation acknowledges that comments may simultaneously serve multiple pragmatic purposes.

(3) Explicit Propagation Behavior: Reader actions are classified as R&C (deep engagement), R (information diffusion), or L (passive consumption/lurking), offering nuance beyond binary actions.

By creating an Emotion-Intent-Behavior Triad for each R&C instance, EmoIntenNet operationalizes Social Exchange Theory (Ahmad et al., 2023), enabling direct analysis of how personalized emotional states are cognitively translated into intents and manifest as behavioral choices.

3.3 Human-LLM Collaborative Annotation

To ensure scalability and reliability, we employed a human-in-the-loop pipeline prioritizing human judgment. Acknowledging recent findings on potential "LLM hacking" (anchoring bias) in subjective tasks (Baumann et al., 2025), we specifically designed our protocol to mitigate this risk.

(1) LLM-Based Pre-annotation. We utilized GLM-4.5 (GLM et al., 2024) and Qwen3-plus (Yang et al., 2025) to generate initial suggestions. For multi-label intent, we employed a dual-model union strategy with a confidence threshold to deliberately maximize recall, surfacing all plausible intents while deferring precision to human verification (see Appendix B for prompts).

(2) Human Verification and Arbitration. Trained annotators conducted a mandatory 100% review. To counteract anchoring bias, guidelines strictly prioritized human judgment over LLM suggestions and emphasized the reader's perspective. Inter-annotator agreement reached Krippendorff's $\alpha = 0.814$ for single-label emotion and Jaccard Index 0.76 for multi-label intent, indicating substantial consensus. Disagreements were resolved via discussion and senior arbitration.

For specific annotation details, please refer to Appendix A.4.

3.4 Dataset Statistics and Analysis

EmoIntenNet comprises user-post interactions across three behaviors (*R&C*, *R*, *L*), seven comment emotions (*Happy* (Ha), *Anger* (An), *Sad* (Sa), *Disgust* (Di), *Fear* (Fe), *Surprise* (Su), *Neutral* (Ne)), and five multi-label propagation intents (*Dissemination and Information* (DI), *Support and Agreement* (SA), *Opposition and Questioning* (OQ), *Attack and Disparagement* (AD), *Emotional Venting* (EV)), as detailed in Table 1 and Table 2. To address inherent class imbalance, we applied LLM-based data augmentation (DA) to the training set following Liao et al. (2025). Table 3 summarizes user profiles and the social relationship graph. Notably, the "Follow" graph $\mathcal{G}_f$ (with edges E_f and

average degree D/U) reflects real-world network sparsity, providing a realistic testbed for relational models motivated by Emotion Contagion Theory.

Table 1: Distribution of behaviors and emotions.

	Propagation Behavior							
	R&C		R		L		Total	
Twitter	22580		22215		21975		66770	
Weibo	15663		15358		15639		46660	
	Emotion Distribution of Comments							
	Ha	An	Sa	Fe	Su	Di	Ne	Total
Twitter	12338	1645	947	446	1862	614	4728	22580
Twitter$_{DA}$	12338	8838	9310	8817	8792	8864	8414	65373
Weibo	6769	1293	1170	462	912	386	4671	15663
Weibo$_{DA}$	6769	4976	4566	4687	4776	4370	4671	34815

Table 2: Distribution of intent.

Platform	DI	SA	OQ	EV	AD
Twitter	9,622	16,923	3,614	2,395	722
Weibo	10,348	12,069	2,109	3,880	465

Table 3: Statistics of user information

	Profiles		History		Total	E_f	D/U
	w/	w/o	w/	w/o			
Twitter	18331	5142	8552	14921	23473	47127	2.01
Weibo	12139	3737	9946	5930	15876	540530	34.05

4 Tasks and Evaluation Protocol

4.1 Task Definitions for Social Perception

EmoIntenNet introduces five progressively defined tasks to systematically evaluate models' social perception capacity.

Task A: Personalized Reader Emotion Recognition. Classifies a reader's emotional reaction into seven categories (**Challenge 1**). Evaluated under *w/o Profile* (text only) and *w/ Profile* (adding user attributes/history), the performance gap Δ_{Profile} measures personalization impact.

Task B: Implicit Propagation Intent Prediction. Predicts multi-label propagation motivations from a reader's comment, probing the cognitive "why" bridging emotions and behaviors (**Challenge 2**).

Task C: Reader Behavior Prediction. Predicts a user's action (*R&C*, *R*, *L*) given a post and author/reader profiles, probing the link between internal states and observable actions (**Challenge 2**).

Task D: Joint Emotion-Behavior Modeling. Requires simultaneous inference of emotion and behavior, evaluating both individual accuracy and the *consistency* of the predicted pair to ground models in unified social cognition (**Challenge 3**).

Task E: In-context Comment Generation. Given a post and user profiles, generates a persona-consistent comment, testing the ultimate goal of socially-intelligent interactive generation (**Challenge 3**).

4.2 Overview of Evaluation Protocol

We evaluate EmoIntenNet using an 8:1:1 data split.

Baselines We benchmark against SOTA baselines with complementary reasoning strengths: advanced LLMs (**GLM4.5-plus** (GLM et al., 2024), **Qwen3-plus** (Yang et al., 2025), **DeepSeek-V3.2** (DeepSeek-AI et al., 2025)), reasoning paradigms (**THOR** (Fei et al., 2023), **TOC** (Weinzierl and Harabagiu, 2024), **Debate** (Fang et al., 2025)), and foundation models (**BERT** (Devlin et al., 2018), **GLM-3-6B** (Du et al., 2022)). All are tested on Tasks A-C; for Tasks D-E, only the three foundation LLMs are used to prevent error propagation and label leakage. See Appendix C for details.

Metrics Single-label tasks (A, C) use Accuracy and Macro-F1 (plus Δ_{Profile} for Task A). Multi-label Task B uses Subset Accuracy, Hamming Loss, and Micro-F1. Task D assesses joint consistency via confusion matrices (Emotion-Emotion, Behavior-Behavior, Emotion-Behavior). Task E employs n-gram metrics (BLEU-4, ROUGE-L) and LLM-based judgments (User Intent Consistency, Propagation Function Consistency, Emotional Expression Consistency). Prompts are in Appendix E.

5 Benchmark Experiments Analysis

Our comprehensive evaluation on the EmoIntenNet benchmark yields several critical insights, answering the following core research questions:

RQ1: Does a reader-centric, emotion-intent-behavior joint-modeling benchmark expose fundamental limitations in current models' ability for personalized social understanding? (supported by **Tasks A–D**)

RQ2: How do SOTA modeling approaches fare in capturing the intricate coupling among emotion, implicit intent, and explicit behavior? (supported by **Tasks B, C and D**)

RQ3: What are the key takeaways and future directions for building socially intelligent NLP systems, as highlighted by our evaluation? (supported by **Tasks A–E**)

5.1 Results and Analysis of Task A

Table 4 shows Task A's overall performance (per-emotion details under *w/o Profile* and *w/ Profile*

are in Tables 7–10, Appendix D). Our evaluation yields three pivotal findings.

First, models exhibit "striking platform dependency" (e.g., Qwen excels on Weibo but not Twitter), validating our cross-platform design. Second, chain-of-reasoning or debate-based reasoning often degrades performance, challenging its universal applicability in socially nuanced tasks. Fine-grained analysis reveals models excel at unambiguous emotions like "Happy" (Tables 7, 8 in Appendix D) but struggle with context-dependent ones (e.g., "Disgust"), exposing limitations in capturing nuanced expressions.

Furthermore, in profile-aware emotion analysis, debate-based models outperform baselines on Twitter, likely due to emergent "dynamic user profiling": arguing diverse interpretations helps construct context-aware mental models of real readers. This underscores that effective personalization requires sophisticated reasoning deeply integrating context, not just more data. Δ_{Profile} scores are predominantly negative (Table 4), revealing that naive profile concatenation often degrades performance, highlighting architectures' inability to transcend shallow feature fusion.

Finally, a counter-intuitive phenomenon emerges regarding user activity: stratifying users into high-activity (top 20%) and low-activity (bottom 20%) groups shows models consistently achieve higher M-F1 scores on low-activity users across both datasets (Fig. 6 in Appendix D), challenging the assumption that more data invariably improves personalization. Grounded in social psychology, we posit high-activity users engage in complex self-presentation (Kim et al., 2024), curating diverse personas that yield noisy profiles and complicate emotional propensity distillation. Conversely, sparse interactions from low-activity users carry stronger signals, faithfully reflecting core dispositions.

5.2 Results and Analysis of Task B

Figures 7 and 8 in Appendix D present Task B (implicit propagation intent) results using Subset Accuracy, Hamming Loss, and Micro F1. Qwen with TOC consistently leads across both datasets (e.g., Micro F1 of 0.850 on Weibo and 0.816 on Twitter), achieving the lowest Hamming Loss and highest Subset Accuracy. Conversely, the DEBATE paradigm degrades performance here, indicating its unsuitability for multi-label intent classification.

Two broader insights emerge. First, models exhibit platform dependency, performing better on Weibo than Twitter, likely due to distinct real user behavior patterns. Second, TOC consistently outperforms other paradigms, highlighting its effectiveness for multi-label tasks where exploring diverse reasoning paths is beneficial.

Table 4: Overall performance on Task A (w/o Profile and w/ Profile)

Models	Twitter		Weibo		Δ_{Profile} (M-F1)	
	M-F1	Acc	M-F1	Acc	Twitter	Weibo
BERT	0.3707	0.6488	0.3516	0.4923	-0.2423	-0.1718
GLM-3-6B	0.3753	0.5815	0.2819	0.4323	-0.1926	+0.0109
GLM	0.5346	0.7436	0.4684	0.6437	-0.0439	-0.0213
Qwen	0.5429	0.7387	0.5205	0.7056	+0.0027	-0.0253
DeepSeek	0.5334	0.7520	0.4916	0.6756	-0.0108	-0.0391
GLM_TOC	0.5096	0.7201	0.4430	0.6341	+0.0030	-0.0112
Qwen_TOC	0.5186	0.7161	0.4856	0.6584	-0.0101	-0.0015
DeepSeek_TOC	0.5238	0.7334	0.4807	0.6724	-0.0329	-0.0064
GLM_THOR	0.4671	0.6953	0.4164	0.6086	-0.0002	+0.0402
Qwen_THOR	0.5092	0.7068	0.4872	0.6328	+0.0499	-0.0006
DeepSeek_THOR	0.4940	0.7112	0.4425	0.6220	-0.0044	-0.0380
GLM_Debate	0.4833	0.7121	0.4557	0.6424	+0.0493	+0.0266
Qwen_Debate	0.5369	0.7383	0.5046	0.6711	+0.0852	-0.0312
DeepSeek_Debate	0.5343	0.7378	0.4745	0.6379	+0.0529	-0.0463

Ultimately, these results underscore the complexity of intent prediction. Unlike emotion or behavior, intent is a cognitive construct requiring deep motivational understanding, which current models struggle with. The substantial gap from perfect scores reveals significant room for improvement, particularly in disambiguating semantically proximate intents.

5.3 Results and Analysis of Task C

Table 5 presents Task C results (per-behavior details in Table 12 in Appendix D), revealing significant challenges for current SOTA models in predicting explicit social behaviors. Echoing prior tasks, we confirm a strong platform dependency. Crucially, advanced reasoning methods exhibit nuanced trade-offs: TOC dramatically boosts detection of high-engagement behaviors (e.g., nearly doubling F1 for R&C on Twitter) but degrades performance on other classes. This indicates such frameworks act as coarse proxies rather than achieving genuine behavioral understanding.

Contrasting Section 5.1, our user-activity analysis (Fig. 9 in Appendix D) shows high-activity users are predicted significantly better than low-activity ones. This supports the hypothesis that behavioral patterns, reflecting underlying personality, are more consistent than volatile emotions, making extensive user history conducive to robust profiling. The accuracy gap highlights current limitations in data-sparse scenarios, where insufficient history prevents reliable profile formation. Ulti-

mately, Task C demonstrates that models fail to effectively integrate nuanced social signals, underscoring the need for architectures unifying user profiles, content, and relational semantics.

Table 5: Overall performance on Task C

	Twitter		Weibo	
models	**M-F1**	**Acc**	**M-F1**	**Acc**
BERT	0.2237	0.3570	0.2601	0.3837
GLM-3-6B	0.3388	0.3482	0.3174	0.3761
GLM	0.3003	0.4047	**0.3701**	**0.4896**
Qwen	**0.3738**	**0.5320**	0.2678	0.3167
DeepSeek	0.3500	0.4745	0.2654	0.3384
GLM_TOC	0.3112	0.4304	0.3660	0.4879
Qwen_TOC	0.3622	0.4874	0.2955	0.4227
DeepSeek_TOC	0.3693	0.4925	0.3416	0.4666
GLM_THOR	0.2589	0.3764	0.3529	0.4542
Qwen_THOR	0.3234	0.4576	0.3536	0.4587
DeepSeek_THOR	0.3675	0.4819	0.3202	0.4417
GLM_Debate	0.2440	0.3511	0.1947	0.3135
Qwen_Debate	0.3404	0.4497	0.3384	0.4689
DeepSeek_Debate	0.3513	0.4614	0.3365	0.4464

5.4 Results and Analysis of Task D

Task D results (Figs. 10–15 in Appendix D) reveal a fundamental limitation in current multi-task paradigms: a striking lack of inter-task consistency when predicting emotion and behavior simultaneously. Despite moderate performance on individual sub-tasks, we observe systemic decoupling—emotion predictions rely on superficial lexical cues (e.g., misclassifying high-arousal "Anger" as "Happy"), while behavior predictions regress to simplistic intensity ranking (e.g., collapsing "Reply & Comment" into "Reply").

This failure underscores a core gap: models lack a unified, theory-driven framework for the emotion-to-behavior causal chain. As posited by *Social Exchange Theory* (Ahmad et al., 2023), emotion is not a direct trigger but an internal variable informing a cost-benefit analysis; decoupled prediction heads fail to capture this mediating cognitive process.

Furthermore, our benchmark exposes the profound challenge of modeling the "silent majority" (lurkers). Drawing from the *Spiral of Silence Effect* (Sohn, 2022), their silence is not neutrality but a passive state shaped by EmoIntenNetd social pressure. Neglecting this cohort skews group sentiment understanding. Their latent affect is critical for modeling *group-level emotion evolution*, as its accumulation can trigger sudden shifts in public opinion. Future work must advance beyond explicit signals to develop mechanisms for *latent affect inference* and predict the activation threshold of this passive population, which is crucial for holistic emotion dynamics on social media.

5.5 Results and Analysis of Task E

Task E's low scores on traditional n-gram metrics (BLEU-4, ROUGE-L; Appendix D, Fig. 16) are expected and indeed validating, confirming that surface similarity is a poor proxy for socially-aware generation. Because the core objective is semantic and pragmatic fidelity in persona-consistent responses rather than replicating a single ground-truth, our proposed LLM-based metrics are indispensable.

As shown in Fig. 17 in Appendix D, while models exhibit baseline competence (3.5–4.5 scores), a clear performance ceiling persists, revealing struggles in internalizing complex profiles and social contexts. Grounded in *Uses and Gratifications Theory* (Xu et al., 2025), users actively seek media to fulfill specific needs; our results indicate models can simulate emotional gratification (EEC) but struggle with strategic, goal-oriented use (UIC) and social functioning (PFC). The lower PFC scores versus EEC confirm that generating text with coherent social purpose is more challenging than merely matching emotion. Furthermore, cross-platform variance underscores that models have yet to master the platform-specific cultural scripts and interaction norms governing online propagation.

6 Conclusion

This paper introduces EmoIntenNet, a novel benchmark shifting social media emotion analysis from an author-centric to a reader-centric perspective. Crucially, EmoIntenNet integrates six interdependent dimensions—author content, reader emotion, implicit intent, behavior, user profiles, and the social relationship graph—to capture the full lifecycle of online emotion. This multidimensional design ensures that the causal chain from content exposure to emotional reaction, and further to motivated behavior, is explicitly modeled within the context of individual traits and social structures. Our evaluation reveals critical gaps: SOTA LLMs struggle to capture personalized emotional responses and fail to model the intrinsic triadic coupling among emotion, intent, and behavior. These findings underscore the urgent need for theory-driven models that fuse content, user, and graph-structural information. Ultimately, EmoIntenNet serves as a catalyst for this new line of inquiry, driving the development of models that understand not just what people feel, but why they act and how emotions propagate through social networks.

Limitations

While our work introduces the novel EmoIntenNet benchmark, several limitations open exciting avenues for future research. First, our evaluation relies on LLMs, which, while scalable, may introduce inherent implicit biases and do not fully account for ethical considerations. Furthermore, our analysis lacks a deep dive into the distinct cultural and platform-specific dynamics of both emotion and implicit intent propagation on Weibo and Twitter. The manifestation and interpretation of the Emotion-Intent-Behavior triad may vary significantly across these socio-cultural contexts. Future work should focus on developing bias-aware evaluation protocols and context-aware models that can reason about these cross-platform nuances. Second, our dataset, like most based on observable interactions, is subject to the "Spiral of Silence" effect. While EmoIntenNet explicitly categorizes passive engagement as Lurking (L), it primarily captures the explicit emotions and intents of active participants, lacking the latent affect and underlying motivations of the vast population of lurkers. Consequently, current models cannot predict the activation threshold of this passive population—a crucial transition from passive consumption to active engagement. A promising direction is to integrate passive signals (e.g., dwell time) or hybrid methodologies to infer the latent affective states and intents of the silent majority, enabling a more holistic and dynamic view of emotion propagation. Third, despite providing user attributes and historical posts, EmoIntenNet inevitably suffers from inherent data sparsity, as a portion of user profiles contain missing fields. This inevitably introduces bias into profile-aware modeling. Future work could explore robust representation learning to better utilize incomplete user information. Fourth, our dataset and task are not suitable for temporal prediction, as they are formulated as emotion and intent recognition in static scenarios without involving temporal modeling. However, we will include timestamps to provide a foundation for future time-series-related research.

Ethical Considerations

This research adheres to strict ethical guidelines concerning data, human annotators, and potential misuse. All data used in the EmoIntenNet benchmark underwent rigorous anonymization. All personally identifiable information and traceable user identifiers were removed before public release to ensure user privacy. All annotators involved in this project were compensated fairly. While this work aims to have a positive impact by advancing our understanding of social dynamics, we strongly condemn any malicious applications. We explicitly prohibit the use of our dataset and derivative models for emotion manipulation, behavioral intervention, non-consensual user tracking, or political manipulation. This work is intended strictly for academic research to prevent its misuse.

Acknowledgements

The works described in this paper are supported by the National Natural Science Foundation of China (U24A20335, 62476162, 62376143, 62473241, 62272286), the Fundamental Research Program of Shanxi Province, China (202303021211021, 202503021211239). Thanks to Bayou Tech (Hong Kong) Limited for providing the data used in this paper free of charge.

References

Rehan Ahmad, Muhammad Rafay Nawaz, Muhammad Ishtiaq Ishaq, Mumtaz Muhammad Khan, and Hafiz Ahmad Ashraf. 2023. Social exchange theory: Systematic review and future directions. *Frontiers in Psychology*, Volume 13 - 2022.

Khudejah Ali, Cong Li, Khawaja Zain ul abdin, and Syed Ali Muqtadir. 2022. The effects of emotions, individual attitudes towards vaccination, and social endorsements on perceived fake news credibility and sharing motivations. *Computers in Human Behavior*, 134:107307.

Valerio Basile, Tommaso Caselli, Alexandra Balahur, and Lun-Wei Ku. 2022. Editorial: Bias, subjectivity and perspectives in natural language processing. *Frontiers in Artificial Intelligence*, Volume 5 - 2022.

Joachim Baumann, Paul Röttger, Aleksandra Urman, Albert Wendsjö, Flor Miriam Plaza del Arco, Johannes B. Gruber, and Dirk Hovy. 2025. Large language model hacking: Quantifying the hidden risks of using llms for text annotation.

Laura Bostan, Evgeny Kim, and Roman Klinger. 2020. Goodnewseveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception.

Nicolas Bougie and Narimawa Watanabe. 2025. SimUSER: Simulating user behavior with large language models for recommender system evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume*

6: Industry Track), pages 43–60, Vienna, Austria. Association for Computational Linguistics.

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Silvia Casola, Simona Frenda, Soda Marem Lo, Erhan Sezerer, Antonio Uva, Valerio Basile, Cristina Bosco, Alessandro Pedrani, Chiara Rubagotti, Viviana Patti, and Davide Bernardi. 2024. MultiPICo: Multilingual perspectivist irony corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16008–16021, Bangkok, Thailand. Association for Computational Linguistics.

Siva Uday Sampreeth Chebolu, Franck Dernoncourt, Nedim Lipka, and Thamar Solorio. 2024. OATS: A challenge dataset for opinion aspect target sentiment joint detection for aspect-based sentiment analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12336–12347, Torino, Italia. ELRA and ICCL.

Chien-Wen Chen, Duong Thuy Trang Nguyen, Mingchang Chih, and Pei-Ying Chen. 2024. Fostering youtube followers' stickiness through social contagion: The role of digital influencer' characteristics and followers' compensation psychology. *Computers in Human Behavior*, 158:108304.

DeepSeek-AI, Daya Guo, Dejian Yang, and et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Nature*, 645:633–638.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200.

Paul Ekman and Wallace V. Friesen. 1971. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129.

Yi Fang, Moxin Li, Wenjie Wang, Lin Hui, and Fuli Feng. 2025. Counterfactual debating with preset stances for hallucination elimination of LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10554–10568, Abu Dhabi, UAE. Association for Computational Linguistics.

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. 2023. Reasoning implicit sentiment with chain-of-thought prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1171–1182, Toronto, Canada. Association for Computational Linguistics.

Team GLM, Aohan Zeng, Bin Xu, and et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Le Dinh Hai and Xiao Yan Xiong. 2025. Motivations behind gen z's news sharing on social media: a pls-sem study in vietnam. *Frontiers in Psychology*, Volume 16 - 2025.

Nils Constantin Hellwig, Jakob Fehle, Udo Kruschwitz, and Christian Wolff. 2025. Do we still need human annotators? prompting large language models for aspect sentiment quad prediction. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*, pages 153–172, Vienna, Austria. Association for Computational Linguistics.

Tiancheng Hu and Nigel Collier. 2025. iNews: A multimodal dataset for modeling personalized affective responses to news. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25000–25040, Vienna, Austria. Association for Computational Linguistics.

Jennifer Ihm and Eun mee Kim. 2025. Why do news sharers choose not to share news? a self-presentation model of deliberate news withholding on social media. *SOCIAL MEDIA + SOCIETY*, 11(4).

Zhongquan Jian, Yanhao Chen, Jiajian Li, Shaopan Wang, Xiangjian Zeng, Junfeng Yao, Xinying An, and Qingqiang Wu. 2025. Simrp: Syntactic and semantic similarity retrieval prompting enhances aspect sentiment quad prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24248–24256.

Yonghyun Jun and Hwanhee Lee. 2025. Exploring persona sentiment sensitivity in personalized dialogue generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18384–18402, Vienna, Austria. Association for Computational Linguistics.

JaeWon Kim, Robert Wolfe, Ishita Chordia, Katie Davis, and Alexis Hiniker. 2024. "sharing, not showing off": How bereal approaches authentic self-presentation on social media through its design. In *Proceedings of the ACM on Human-Computer Interaction*, volume 8, pages 1–32, New York, NY, USA. Association for Computing Machinery.

Aman Kumar, Amit Shankar, Abhishek Behl, Varsha Arya, and Nakul Gupta. 2023. Should i share it? factors influencing fake news-sharing behaviour: A behavioural reasoning theory perspective. *Technological Forecasting and Social Change*, 193:122647.

Shichen Li, Zhongqing Wang, Zheyu Zhao, Yue Zhang, and Peifeng Li. 2025. Exploring model editing for llm-based aspect-based sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24467–24475.

Jian Liao, Yu Feng, Yujin Zheng, Jun Zhao, Suge Wang, and Jianxing Zheng. 2025. My words imply your opinion: Reader agent-based propagation enhancement for personalized implicit emotion analysis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vienna, Austria. Association for Computational Linguistics.

Jiongnan Liu, Yutao Zhu, Shuting Wang, Xiaochi Wei, Erxue Min, Yu Lu, Shuaiqiang Wang, Dawei Yin, and Zhicheng Dou. 2025. LLMs + persona-plug = personalized LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9373–9385, Vienna, Austria. Association for Computational Linguistics.

Chenyang Lyu, Linyi Yang, Yue Zhang, Yvette Graham, and Jennifer Foster. 2023. Exploiting rich textual user-product context for improving personalized sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1419–1429, Toronto, Canada. Association for Computational Linguistics.

Saif Mohammad and Felipe Bravo-Marquez. 2017. WASSA-2017 shared task on emotion intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 34–49, Copenhagen, Denmark. Association for Computational Linguistics.

Xinyi Mou, Zhongyu Wei, and Xuanjing Huang. 2024. Unveiling the truth and facilitating change: Towards agent-based large-scale social movement simulation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4789–4809, Bangkok, Thailand. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, Guanqun Cao, Rodrigo Tufiño, Rendi Chevi, Chiamaka Ijeoma Chukwuneke, Alexandra Ciobotaru, Daryna Dementieva, Murja Sani Gadanya, Robert Geislinger, Bela Gipp, Oumaima Hourrane, Oana Ignat, Falalu Ibrahim Lawan, Rooweither Mabuya, Rahmad Mahendra, Vukosi Marivate, Alexander Panchenko, Andrew Piper, Charles Henrique Porto Ferreira, Vitaly Protasov, Samuel Rutunda, Manish Shrivastava, Aura Cristina Udrea, Lilian Diana Awuor Wanzare, Sophie Wu, Florian Valentin Wunderlich, Hanif Muhammad Zhafran, Tianhui Zhang, Yi Zhou, and Saif M. Mohammad. 2025. BRIGHTER: BRIdging the gap in human-annotated textual emotion recognition datasets for 28 languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8895–8916, Vienna, Austria. Association for Computational Linguistics.

Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. 2020. XED: A multilingual dataset for sentiment analysis and emotion detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6542–6552, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392, Bangkok, Thailand. Association for Computational Linguistics.

Dongyoung Sohn. 2022. Spiral of silence in the social media era: A simulation approach to the interplay between social networks and mass media. *Communication Research*, 49(1):139–166.

Enrica Troiano, Laura Oberländer, and Roman Klinger. 2023. Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1):1–72.

Geng Tu, Jun Wang, Li Yang, Bin Liang, Erik Cambria, Wenjie Li, and Ruifeng Xu. 2026. Multi-task

mutual learning for multimodal emotion-cause pair extraction in conversations. *Information Fusion*, 127:103877.

Jianan Wang, Bin Li, Jingtao Qi, Xueying Wang, Fu Li, and Lihanxun. 2025a. BeSimulator: A large language model powered text-based behavior simulator. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4736–4754, Suzhou, China. Association for Computational Linguistics.

Yujie Wang, Yunwei Zhao, Jing Yang, Han Han, Shiguang Shan, and Jie Zhang. 2025b. Evaluating cognitive-behavioral fixation via multimodal user viewing patterns on social media. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19571–19583, Suzhou, China. Association for Computational Linguistics.

Yunzhe Wang, Gale Lucas, Burcin Becerik-Gerber, and Volkan Ustun. 2025c. Implicit behavioral alignment of language agents in high-stakes crowd simulations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30669–30686, Suzhou, China. Association for Computational Linguistics.

Yangbo Wei, Zhen Huang, Fangzhou Zhao, Qi Feng, and Wei W. Xing. 2025. MECoT: Markov emotional chain-of-thought for personality-consistent role-playing. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8297–8314, Vienna, Austria. Association for Computational Linguistics.

Maxwell Weinzierl and Sanda Harabagiu. 2024. Tree-of-counterfactual prompting for zero-shot stance detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–880, Bangkok, Thailand. Association for Computational Linguistics.

James D. Westaby. 2005. Behavioral reasoning theory: Identifying new linkages underlying intentions and behavior. *Organizational Behavior and Human Decision Processes*, 98(2):97–120.

ChengYan Wu, Bolei Ma, Yihong Liu, Zheyu Zhang, Ningyuan Deng, Yanshu Li, Baolan Chen, Yi Zhang, Yun Xue, and Barbara Plank. 2025. M-ABSA: A multilingual dataset for aspect-based sentiment analysis. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2530–2557, Suzhou, China. Association for Computational Linguistics.

Chunze Xu, Fengqiang Gao, and Lei Han. 2025. Enhancing user information disclosure intention in dynamic conversations of intelligent recommendation systems based on large language models: A perspective of user gratification and privacy calculus. *International Journal of Human-Computer Studies*, 200:103511.

Yinuo Xu and David Jurgens. 2026. Beyond consensus: Perspectivist modeling and evaluation of annotator disagreement in nlp.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Jing Ye, Lu Xiang, Yaping Zhang, and Chengqing Zong. 2025. From generic empathy to personalized emotional support: A self-evolution framework for user preference alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18826–18853, Suzhou, China. Association for Computational Linguistics.

Xusheng Zhao, Hao Peng, Qiong Dai, Xu Bai, Huailiang Peng, Yanbing Liu, Qinglang Guo, and Philip S Yu. 2024. Rdgcn: Reinforced dependency graph convolutional network for aspect-based sentiment analysis. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 976–984.

Wangchunshu Zhou, Qifei Li, and Chenle Li. 2023. Learning to predict persona information for dialogue personalization without explicit persona description. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2979–2991, Toronto, Canada. Association for Computational Linguistics.

A Dataset Details

A.1 Dataset Comparison

Table 6 provides a detailed comparison of EmoIntenNet with existing datasets.

A.2 Definition of Lurking

We clarify the definition of lurker and its implications for our experimental results:

(1) Due to the inherent limitations of data collection processes and social media platform APIs, it is infeasible to precisely determine users' actual browsing status (e.g., whether a post appeared on their feed but was not actually viewed). Therefore, distinguishing between users who "viewed but did

Table 6: Comparison of EmoIntenNet with existing sentiment analysis datasets.

Dataset	Emotion Subject	Domain	Language	Intent	Behavior	User Profile	Social Graph	Scale	Applicable Tasks
SemEval-14	author	product reviews	EN	×	×	×	×	7.6k	ABSA
ECC	author	conversation	EN	×	×	×	×	2.4k	emotion cause extraction
SemEval-17	author	Twitter	EN, AR	×	×	×	×	72k	ABSA
GoEmotions	author	Reddit	EN	×	×	×	×	58k	ABSA
OATS-ABSA	author	social media	EN	×	×	×	×	44k	open-target ABSA
M-ABSA	author	social media	EN	×	×	×	×	14k	multimodal ABSA
BRIGHTER	author	social media	multiple	×	×	×	×	100k	sentiment analysis
XED	author	movie subtitles	multiple	×	×	×	×	55k	sentiment analysis
RAPPIE	author	social media	EN, CN	×	×	✓	CN only	29k	personalized emotion analysis
EmoBank	average reader	news	EN	×	×	×	×	10k	dimensional emotion analysis
MultiPICo	annotator	social media	multiple	×	×	506 annotators'	×	18.8k	irony detection
INEWS	annotator	news	EN	×	×	291 annotators'	×	2.9k	emotional respond simulation
EmoIntenNet	individual reader	social media	EN, CN	✓	✓	39349 real reader users	✓	113k	personalized emotion/intent/behavior analysis, emotion propagation

not interact" and those who "did not view" the content remains extremely challenging under publicly available data constraints. To this end, we adopt a broad yet operable definition: a user is categorized as a Lurker if they follow the post author (thus having the theoretical potential to view the post) but produce no explicit behaviors (neither repost nor comment). This represents the most reasonable approximation given the data at hand.

(2) As noted above, lurker identification itself is a non-trivial task. EmoIntenNet provides foundational data resources—including following relationships, user historical posts, and user profiles—to support future research aimed at more fine-grained lurker modeling.

(3) Our primary research focus is on users' observable behaviors (repost and comment), emotions, and propagation intents. The Lurker category primarily serves as a reference baseline in the three-class behavior prediction task, rather than as the core object for analyzing internal differences within the silent population.

(4) This undifferentiated lurker definition may affect the interpretation of silent users' true status to some extent: users who were never exposed to the content may be misclassified as "choosing to remain silent," potentially underestimating genuine emotional resonance within the group and weakening the credibility of activation threshold analysis under the Spiral of Silence effect. However, the impact on our modeling of observable behavior–emotion–intent triads remains limited, since explicit behavior modeling does not heavily rely on features of the silent group. The training and evaluation of these tasks primarily draw from samples with explicit interaction signals, whereas lurkers—lacking comments or other observable data—offer limited meaningful features and are therefore not the core modeling target of this work.

A.3 Detailed Topic Filtering Strategy

To ensure social relevance beyond keyword matching, we employed a rigorous multi-step process:

(1) Extracting hashtagged posts as initial seeds;

(2) Grouping multiple raw hashtags into broader semantic domains using keyword matching and Latent Dirichlet Allocation (LDA)—a probabilistic topic model that clusters co-occurring words into latent domains, where we set the number of domains $K = 50$;

(3) Within these domains, refining specific conversation topics based on volume and engagement diversity, carefully retaining niche topics with high research value while discarding those dominated by meaningless noise (e.g., spam or repetitive replies); and

(4) Manual expert review to ensure thematic coherence.

This process retained 116 Twitter topics and 139 Weibo topics, ensuring EmoIntenNet is a curated archive of meaningful public conversations.

It should be noted that LDA is employed in this work only as a lightweight coarse-grained topic filtering tool, accompanied by a rigorous manual review process. Therefore, its known limitations on short texts have little impact on the final research results. LDA is used only in the data sampling stage to merge a large number of hashtagged posts from raw labels into several broader latent topic domains, serving as a lexical-level preliminary topic filter rather than the final topic judgment. Although LDA generally performs weaker on short texts than on long texts, given that this step is only a pre-filter, we also conducted strict manual reviews for each topic group to ensure topic relevance. Hence, the precision of LDA has limited influence on the final tasks, and the manual review effectively compensates for the shortcomings of LDA in short-text modeling.

A.4 Annotation Protocol Details

LLM Pre-annotation Settings. We utilized GLM-4.5 (GLM et al., 2024) and Qwen3-plus (Yang et al., 2025) for pre-annotation. For single-label emotion, GLM-4.5 provided labels with confidence scores. For multi-label intent, we employed an ensemble, retaining the **union of labels with confidence** $>$ 0.5 as the candidate set.

Human Annotation Details. Six trained postgraduate annotators (compensated at standard institutional research assistant rates) conducted the mandatory review. To explicitly counteract anchoring bias, annotators were instructed that LLM suggestions were unverified hypotheses, not ground truth. Guidelines provided precise definitions of edge cases and emphasized the absolute prioritization of human judgment over LLM suggestions.

A.5 Propagation Intent

The propagation intents are defined as follows:

- **Dissemination and Information:** Conveying facts, supplementing details, or sharing information, aiming at information diffusion and sharing.
- **Support and Agreement:** Expressing alignment, endorsement, or solidarity, aiming at reinforcing a stance or viewpoint.
- **Opposition and Questioning:** Expressing disagreement, doubt, or challenge, aiming at correcting, arguing, or showing dissent.
- **Attack and Disparagement:** Using insulting, mocking, or demeaning language, aiming at causing psychological harm or destroying reputation.
- **Emotional Venting:** Releasing strong personal emotions without primarily seeking truth or attack, aiming at the direct outward expression of feelings.

B Prompt for LLM-based Pre-annotation

The prompt for LLM-based pre-annotation is illustrated in Fig. 4 and Fig. 5.

C Detailed Evaluation Settings

C.1 Task-Specific Evaluation Metrics

The metrics are meticulously chosen to align with the objective of each task.

- **Task A, B, & C:** We report **Accuracy** and **Macro-F1**. Accuracy measures overall correctness, while Macro-F1 ensures balanced performance across all emotion/behavior categories, including minorities. We also report the F1 score for individual emotions/behaviors. The performance gain from Task A to B is a key indicator of a model's effectiveness in utilizing user profiles for personalization.
- **Task D:** We report **Confusion Matrix (CM)** for the emotion-emotion (EECM), behavior-behavior (BBCM), and emotion-behavior (EBCM) prediction independently. Recognizing that emotion annotations are exclusive to the *R&C* behavior, we compute the EBCM specifically for this category. This provides a controlled and meaningful assessment of a model's joint emotion-behavior reasoning capabilities, focusing on the interaction type with the richest semantic signal.

Prompt for LLM-Based Pre-annotation

```
{"role": "system",
 "content":"You are an emotion analysis expert. Your task is to analyze the user's reply [reply_text] based_on the
            original Twitter [retweet_text] and the user's reply content [reply_text], and determine which of the
            seven emotion categories it belongs to: [Happy, Anger, Neutral, Fear, Disgust, Sad, Surprise].
            For each emotion categories, you must provide:
            Analysis="a brief explanation of why the user's reply may or may not express that emotion";
            Possibility="a probability score between 0 and 1 indicating the likelihood of belonging to that category";
            Targets="the target words in the sentence corresponding to the emotion (nouns or noun phrases related to the emotion
            in the sentence), if any; if none, leave blank".
            Then, based on your analysis, determine the final emotion classification and provide a confidence score for the final
            opinion, reflecting your certainty about the overall classification.
            The output must strictly adhere to the following format:
            {'Text':'user's reply [reply_text]',
            'Happy':{'Analysis':'analysis of the Happy emotion category for the given text', 'Possibility':'score','Targets':[]},
            'Anger':{'Analysis':'analysis of the Anger emotion category for the given text','Possibility':'score','Targets':[]},
            'Neutral':{'Analysis':'analysis of the Neutral emotion category for the given text','Possibility':'score','Targets':[],},
            'Fear':{'Analysis':'analysis of the Fear emotion category for thr given text','Possibility':'score','Targets':[]},
            'Disgust':{'Analysis':'analysis of the Disgust emotion category for the given text','Possibility':'score','Targets':[]},
            'Sad':{'Analysis':'analysis of the Sad emotion category for the given text','Possibility':'score','Targets':[]},
            'Surprise':{'Analysis':'analysis of the Surprise emotion category for the given text', 'Possibility':'score','Targets':[]},
            'Sentiment':'emotion','Confidence':'confidence score'}.
            The given example is:
            input:[retweet_text='TOGO 3D PRINTER - YouTube: :)',reply_text='Well done Affate, pride of #Africa'],
            output:[{'Text': 'Well done Affate, pride of #Africa',
            'Happy': {'Analysis': 'The user's reply use positive language such as 'Well done' and  'pride', which is typically associated
            with positive emotions,expressing recognition and pride in someone's achievement.','Possibility': 0.9,'Targets':
            ['Affate',]},
           'Anger': {'Analysis': 'The text shows no expression of anger or negative emotions; the tone is positive and encouraging.',
           'Possibility': 0.05,'Targets': []},
           'Neutral': {'Analysis': 'Although the sentence is concise, its wording carries positive emotion and dose not align well
           with a neutral emotion', 'Possibility': 0.02,'Targets': []},
           'Fear': {'Analysis': 'There is no content or implication related to fear in the text.','Possibility': 0.01,'Targets': []},
           'Disgust': {'Analysis': 'No expression of disgust or aversion is present.','Possibility': 0.01,'Targets': []},
           'Sad': {'Analysis': 'The reply does not express sadness or a sense of loss.','Possibility': 0.01,'Targets': []},
           'Surprise': {'Analysis': 'The reply does not show a tone ofsurprise or express unexpectedness.','Possibility': 0.01,
           'Targets': []},
           'Sentiment': 'Happy','Confidence':0.9}]"},
{"role":"user",
"content":"retweet_text:[{}];reply_text:[{}]"}
```

Figure 4: Prompt for LLM-based Emotion pre-annotation.

- **Task E:** We employ standard automatic metrics, including **BLEU-4** and **ROUGE-L**, to measure the n-gram overlap between the generated comments and the ground truth. Furthermore, we leverage GLM4.5-plus, Qwen3-plus, and DeepSeek-V3.2 to evaluate the generated content through the corresponding LLMs along three distinct dimensions: User Intent Consistency (**UIC**), Propagation Function Consistency (**PFC**), and Emotional Expression Consistency (**EEC**). Each dimension is rated on a 1-to-5 scale, with the mean score serving as the final evaluation metric.

C.2 Baselines

To ensure a rigorous evaluation, we select two foundation models, along with a complementary set of representative state-of-the-art LLMs and reasoning paradigms, to benchmark on our datasets. **(1) Foundation Models as Backbones:** We include two representative foundation models as the most basic baselines, providing a reference for evaluating the reasoning capabilities of larger LLMs.

- **BERT** (Devlin et al., 2018): A bidirectional Transformer pre-trained on large-scale text corpora, widely adopted as a standard backbone for sentence-level classification tasks.
- **GLM-3-6B** (Du et al., 2022): A bilingual (English-Chinese) pre-trained model with 6 billion parameters, serving as a representative medium-sized backbone for cross-lingual social media understanding.

(2) Representative Online LLMs: Our selection encompasses leading models recognized for their advanced reasoning capabilities.

- **GLM4.5-plus** (GLM et al., 2024): A model distinguished by its native integration of complex reasoning, coding, and agent capabilities, making it a robust baseline for multifaceted tasks.
- **Qwen3-plus** (Yang et al., 2025): An open-weight "thinking" model that demonstrates state-of-the-art performance on tasks requiring expert-level, multi-step reasoning.
- **DeepSeek-V3.2** (DeepSeek-AI et al., 2025): An architecture architected with a specific focus on deep, step-by-step logical deduction ("thinking" ability), excelling in problems that demand extended reasoning chains.

(3) Advanced Reasoning Paradigms: We further investigate established frameworks designed to augment the intrinsic reasoning abilities of LLMs.

- **THOR** (Fei et al., 2023): We employ a standard Chain-of-Thought prompting template, which guides the model to decompose the problem into a sequence of intermediate reasoning steps before yielding a final answer.
- **TOC** (Weinzierl and Harabagiu, 2024): This method models the reasoning process as a tree structure, systematically exploring multiple solution paths and using self-evaluation to identify the most promising line of thought.
- **Debate-based Reasoning (Debate)** (Fang et al., 2025): A multi-agent framework that enhances reasoning by compelling models to defend counterfactual stances, thereby mitigating inherent biases and stimulating more thorough justification.

D Detailed Results and Visualization

This section provides a comprehensive breakdown of the evaluation results for Tasks A-E. Table 7–10 presents the per-category and overall F1 scores for Task A (emotion prediction); Table 11, Figs. 7 and 8 present the Task B (implicit propagation intent prediction) results; Table 12 presents the per-category and overall F1 scores for Task C (behavior prediction). Fig. 9 further analyzes the performance of Task C by comparing high- and low-activity users. Figs. 10-15 present a comparative analysis of different foundation LLMs (GLM, Qwen, DeepSeek) on our core metrics (EECM, BBCM, EBCM). The quality of generated content for Task E is evaluated in Fig. 16 using BLEU and Rouge-L scores.

E The Prompts for LLM-based Evaluations

The prompts for LLM-based evaluations of each task are illustrated in Figs. 18-31. Note that these prompts are designed for emotion annotation; the prompts for propagation intent annotation are analogous and thus omitted here. Fig. 32 presents the prompt of UIC, PFC, and EEC measurements for in-context comment generation evaluation.

Three LLM-based judgments designed to capture different aspects of generation quality:

(1) User Intent Consistency (UIC): Measures whether the generated reply aligns with the user intent and core information of the real reply.

(2) Propagation Function Consistency (PFC): Assesses whether the generated reply exhibits similar propagation functions (e.g., triggering interactions, reinforcing viewpoints) as the real reply.

(3) Emotion Expression Consistency (EEC): Evaluates whether the generated reply maintains consistency with the real reply in terms of emotional tendency, intensity, and expression style.

Table 7: Detailed results for Task A in Twitter dataset (w/o Profile)

Models	Happy	Sad	Anger	Surprise	Disgust	Fear	Neutral	M-F1	Acc
BERT	0.7820	0.1951	0.5221	0.3408	0.1892	0.0426	0.5229	0.3707	0.6488
GLM-3-6B	0.7274	0.2655	0.5166	0.2259	0.1882	0.2456	0.4579	0.3753	0.5815
GLM	0.8661	0.7296	0.4455	0.5648	0.3733	0.6842	0.6137	0.5346	0.7436
Qwen	0.8550	0.6832	0.5188	0.5955	0.4556	0.7308	0.5043	0.5429	0.7387
DeepSeek	0.8687	0.7219	0.6473	0.5250	0.2963	0.6250	0.5833	0.5334	0.7520
GLM_TOC	0.8411	0.6447	0.6424	0.5567	0.3636	0.6111	0.4170	0.5096	0.7201
Qwen_TOC	0.8391	0.7421	0.5420	0.5816	0.4645	0.6279	0.3516	0.5186	0.7161
DeepSeek_TOC	0.8575	0.7052	0.4416	0.5872	0.4277	0.6750	0.4958	0.5238	0.7334
GLM_THOR	0.8415	0.6410	0.3318	0.4910	0.3760	0.6301	0.4251	0.4671	0.6953
Qwen_THOR	0.8401	0.6826	0.4367	0.5743	0.4362	0.7126	0.3910	0.5092	0.7068
DeepSeek_THOR	0.8430	0.6864	0.5973	0.5229	0.3392	0.5672	0.3962	0.4940	0.7112
GLM_DEBATE	0.8381	0.5972	0.3256	0.5054	0.3944	0.6667	0.5393	0.4833	0.7121
Qwen_DEBATE	0.8624	0.6905	0.5936	0.5493	0.4316	0.6667	0.5008	0.5369	0.7383
DeepSeek_DEBATE	0.8501	0.6709	0.4769	0.6163	0.4468	0.5846	0.6288	0.5343	0.7378

Table 8: Detailed results for Task A in Weibo dataset (w/o Profile)

Models	Happy	Sad	Anger	Surprise	Disgust	Fear	Neutral	M-F1	Acc
BERT	0.6084	0.3162	0.5363	0.1101	0.1154	0.3294	0.4456	0.3516	0.4923
GLM-3-6B	0.6134	0.2375	0.7519	0.0855	0.0000	0.0673	0.2177	0.2819	0.4323
GLM	0.7317	0.5165	0.5670	0.4571	0.4359	0.3729	0.6660	0.4684	0.6437
Qwen	0.7930	0.6250	0.5571	0.4252	0.4800	0.5882	0.6953	0.5205	0.7056
DeepSeek	0.7711	0.5976	0.5263	0.5000	0.4828	0.4068	0.6486	0.4916	0.6756
GLM_TOC	0.7528	0.5670	0.5399	0.4545	0.2581	0.3793	0.5922	0.4430	0.6341
Qwen_TOC	0.7813	0.5647	0.5124	0.4521	0.4565	0.5600	0.5575	0.4856	0.6584
DeepSeek_TOC	0.7735	0.6140	0.4455	0.4828	0.4444	0.4571	0.6279	0.4807	0.6724
GLM_THOR	0.7222	0.5385	0.4481	0.3802	0.3729	0.2807	0.5885	0.4164	0.6086
Qwen_THOR	0.7383	0.6066	0.5714	0.5068	0.3613	0.6000	0.5130	0.4872	0.6328
DeepSeek_THOR	0.7394	0.6047	0.4508	0.4000	0.3409	0.5000	0.5041	0.4425	0.622
GLM_DEBATE	0.7604	0.4574	0.5988	0.4375	0.3738	0.4571	0.5604	0.4557	0.6424
Qwen_DEBATE	0.7951	0.5256	0.6842	0.5342	0.3621	0.5429	0.5929	0.5046	0.6711
DeepSeek_DEBATE	0.6916	0.6218	0.5422	0.5072	0.4048	0.3793	0.6493	0.4745	0.6379

Table 9: Detailed results for Task A in Twitter dataset (w/ Profile)

Models	Happy	Sad	Anger	Surprise	Disgust	Fear	Neutral	M-F1	Acc
BERT	0.6880	0.0000	0.1505	0.0000	0.0000	0.0000	0.0599	0.1284	0.5182
GLM-3-6B	0.6118	0.1548	0.1544	0.1397	0.0000	0.0000	0.2183	0.1827	0.4132
GLM	0.8521	0.6710	0.3962	0.5455	0.4186	0.5373	0.5045	0.4907	0.7152
Qwen	0.8562	0.7273	0.5814	0.5126	0.4564	0.6818	0.5493	0.5456	0.7436
DeepSeek	0.8618	0.6871	0.6714	0.5215	0.3471	0.5455	0.5464	0.5226	0.7484
GLM_TOC	0.8384	0.6667	0.6782	0.5563	0.3918	0.5672	0.4023	0.5126	0.7219
Qwen_TOC	0.8327	0.6753	0.5737	0.5650	0.5342	0.5882	0.2990	0.5085	0.7046
DeepSeek_TOC	0.8469	0.6795	0.4502	0.5643	0.3822	0.5143	0.4897	0.4909	0.7183
GLM_THOR	0.8394	0.6452	0.3366	0.4478	0.3786	0.6479	0.4401	0.4669	0.6975
Qwen_THOR	0.8589	0.7176	0.5821	0.6286	0.5116	0.6667	0.5072	0.5591	0.7471
DeepSeek_THOR	0.8363	0.7066	0.5891	0.4742	0.3546	0.5625	0.3932	0.4896	0.7077
GLM_DEBATE	0.8485	0.6207	0.5159	0.5743	0.4216	0.7013	0.5782	0.5326	0.7360
Qwen_DEBATE	0.8657	0.6536	0.5965	0.5510	0.4706	0.6301	0.5872	0.6221	0.7484
DeepSeek_DEBATE	0.8374	0.6358	0.5939	0.4880	0.4027	0.5397	0.6129	0.5872	0.7245

Prompt for LLM_Based Pre_annotation

{"role": "system",
"content": """You are an expert in user communication intent analysis. Your task is to analyze the retweet user's intent based on the original text [retweet_text] and the retweet reply content [reply_text] provided by the user.
Intent categories are: ["Dissemination and Information", "Support and Agreement", "Opposition and Questioning", "Attack and Disparagement", "Emotional Venting"].
The communication intents are defined as follows:
Dissemination and Information: Conveying facts, supplementing details, or sharing information, aiming at information diffusion and sharing.
Support and Agreement: Expressing alignment, endorsement, or solidarity, aiming at reinforcing a stance or viewpoint.
Opposition and Questioning: Expressing disagreement, doubt, or challenge, aiming at correcting, arguing, or showing dissent.
Attack and Disparagement: Using insulting, mocking, or demeaning language, aiming at causing psychological harm or destroying reputation.
Emotional Venting: Releasing strong personal emotions without primarily seeking truth or attack, aiming at the direct outward expression of feelings.
Requirements are as follows:
Analyze each category separately, providing judgment basis and confidence level (between 0-1, indicating the possibility of that intent).
Finally, select the most appropriate category and provide an overall confidence level.
Strictly return the following JSON format, do not add any additional explanations.
Output format example:
{{
"Dissemination and Information": {{
"Analysis": {"analysis of belonging or not belonging"}.
"Possibility": confidence score (a number between 0-1)
}},
"Support and Agreement": {{
"Analysis": {"analysis of belonging or not belonging"}.
"Possibility": confidence score (a number between 0-1)
}},
"Opposition and Questioning": {{
"Analysis": {"analysis of belonging or not belonging"}.
"Possibility": confidence score (a number between 0-1)
}},
"Attack and Disparagement": {{
"Analysis": {"analysis of belonging or not belonging"}.
"Possibility": confidence score (a number between 0-1)
}},
"Emotional Venting": {{
"Analysis": {"analysis of belonging or not belonging"}.
"Possibility": confidence score (a number between 0-1)
}},
"label": "finally determined category",
"confidence": confidence score (a number between 0-1)
}}"""}

Figure 5: Prompt for LLM-based Propagation Intent pre-annotation.

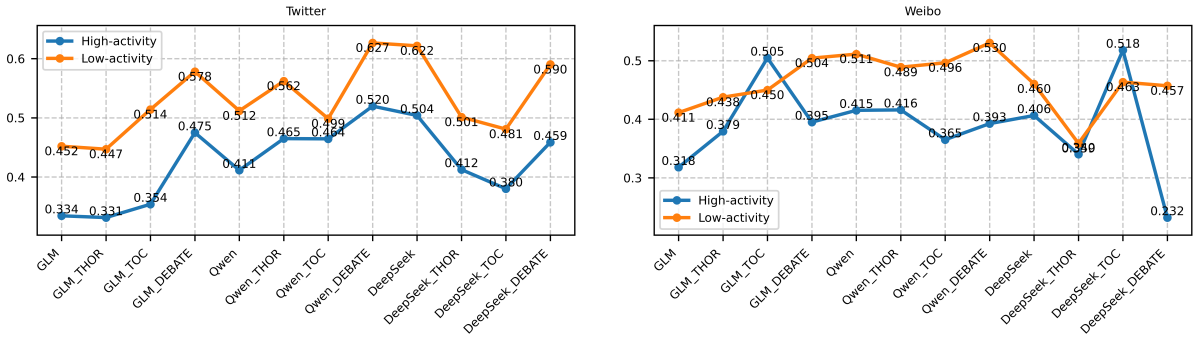


Figure 6: Comparison of high- and low- activity users for Task A.

Table 10: Detailed results for Task A in Weibo dataset (w/ Profile)

Models	**Happy**	**Sad**	**Anger**	**Surprise**	**Disgust**	**Fear**	**Neutral**	**M-F1**	**Acc**
BERT	0.4840	0.0325	0.2632	0.0408	0.0000	0.0417	0.3965	0.1798	0.3946
GLM-3-6B	0.5036	0.2010	0.5036	0.2405	0.1087	0.0656	0.4269	0.2928	0.4163
GLM	0.7111	0.5591	0.5161	0.4328	0.4928	0.2308	0.6342	0.4471	0.6073
Qwen	0.7643	0.6027	0.5455	0.3652	0.4490	0.5915	0.6434	0.4952	0.6628
DeepSeek	0.7464	0.5973	0.5179	0.4380	0.3929	0.3214	0.6062	0.4525	0.6424
GLM_TOC	0.7456	0.5385	0.4883	0.4094	0.3279	0.3390	0.6059	0.4318	0.6258
Qwen_TOC	0.7740	0.6128	0.5289	0.4203	0.4096	0.5556	0.5718	0.4841	0.6533
DeepSeek_TOC	0.7583	0.6161	0.4044	0.4615	0.4407	0.4839	0.6294	0.4743	0.6571
GLM_THOR	0.7342	0.5907	0.5484	0.4539	0.4202	0.3273	0.5781	0.4566	0.6175
Qwen_THOR	0.7531	0.6518	0.5346	0.4341	0.3761	0.5479	0.5953	0.4866	0.6501
DeepSeek_THOR	0.7248	0.5604	0.3542	0.4032	0.3939	0.2642	0.5356	0.4045	0.6098
GLM_DEBATE	0.7375	0.5773	0.6367	0.5429	0.3878	0.4478	0.5285	0.4823	0.6354
Qwen_DEBATE	0.7557	0.5052	0.6502	0.4490	0.3738	0.4516	0.6019	0.4734	0.6411
DeepSeek_DEBATE	0.6877	0.5414	0.4976	0.4286	0.3768	0.2642	0.6292	0.4282	0.6111

Table 11: Detailed results for Task B

	Weibo			Twitter		
Models	Subset Accuracy	Hamming Loss	Micro-F1	Subset Accuracy	Hamming Loss	Micro-F1
BERT	0.5236	0.1508	0.8019	0.3415	0.1956	0.6692
GLM-3-6B	0.4336	0.1766	0.7594	0.3211	0.2020	0.6713
GLM	0.4700	0.1478	0.7751	0.5221	0.1229	0.7684
Qwen	0.5089	0.1335	0.8023	0.5474	0.1135	0.7961
DeepSeek	0.5179	0.1342	0.8096	0.4677	0.1404	0.7375
GLM_TOC	0.5236	0.1281	0.8311	0.5000	0.1259	0.7786
Qwen_TOC	0.5715	0.1149	0.8502	0.5297	0.1136	0.8161
DeepSeek_TOC	0.5192	0.1341	0.8120	0.4553	0.1418	0.7461
GLM_THOR	0.3870	0.1668	0.7394	0.4694	0.1371	0.7758
Qwen_THOR	0.4885	0.1497	0.7847	0.4296	0.1519	0.7473
DeepSeek_THOR	0.4706	0.1572	0.7667	0.4739	0.1439	0.7314
GLM_DEBATE	0.3314	0.2080	0.6461	0.4411	0.1506	0.7517
Qwen_DEBATE	0.2797	0.2000	0.6667	0.5332	0.1220	0.7623
DeepSeek_DEBATE	0.3487	0.1770	0.7127	0.5394	0.1197	0.7717

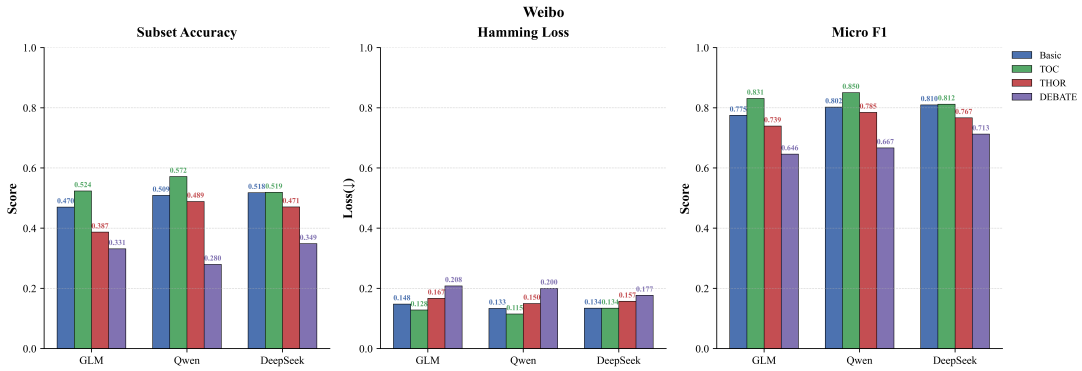


Figure 7: Task B results on Weibo dataset

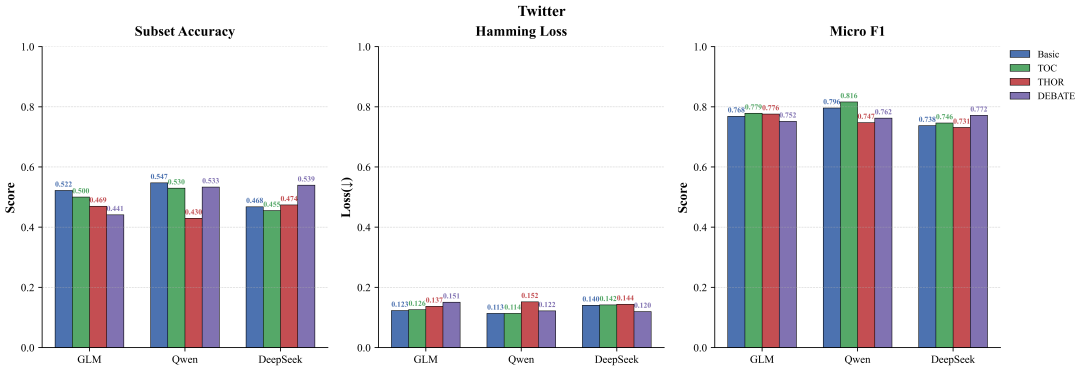


Figure 8: Task B results on Twitter dataset

Table 12: Detailed results for Task C

	Twitter					Weibo				
models	**R&C**	**R**	**L**	**M-F1**	**Acc**	**R&C**	**R**	**L**	**M-F1**	**Acc**
BERT	0.1510	0.0029	0.5172	0.2237	0.3570	0.0038	0.2331	0.5434	0.2601	0.3837
GLM-3-6B	0.4094	0.2355	0.3716	0.3388	0.3482	0.1123	0.3494	0.4906	0.3174	0.3761
GLM	0.3949	0.4303	0.3760	0.3003	0.4047	0.3411	0.4884	0.6511	0.3701	0.4896
Qwen	0.2801	0.5066	0.7084	0.3738	0.5320	0.4629	0.2788	0.3295	0.2678	0.3167
DeepSeek	0.4432	0.3474	0.6093	0.3500	0.4745	0.248	0.4055	0.4080	0.2654	0.3384
GLM_TOC	0.5340	0.284	0.4269	0.3112	0.4304	0.5129	0.3179	0.6332	0.3660	0.4879
Qwen_TOC	0.3923	0.4104	0.6463	0.3622	0.4874	0.4738	0.1468	0.5612	0.2955	0.4227
DeepSeek_TOC	0.4701	0.3411	0.6661	0.3693	0.4925	0.3881	0.3551	0.6232	0.3416	0.4666
GLM_THOR	0.3585	0.4436	0.2334	0.2589	0.3764	0.4106	0.4560	0.5449	0.3529	0.4542
Qwen_THOR	0.1841	0.4792	0.6303	0.3234	0.4576	0.4079	0.4143	0.5924	0.3536	0.4587
DeepSeek_THOR	0.4539	0.4118	0.6044	0.3675	0.4819	0.2791	0.4160	0.5857	0.3202	0.4417
GLM_DEBATE	0.4590	0.3087	0.2085	0.2440	0.3511	0.4194	0.3114	0.0481	0.1947	0.3135
Qwen_DEBATE	0.4532	0.3645	0.5438	0.3404	0.4497	0.4816	0.2407	0.6314	0.3384	0.4689
DeepSeek_DEBATE	0.4678	0.3555	0.5819	0.3513	0.4614	0.4181	0.3104	0.6173	0.3365	0.4464

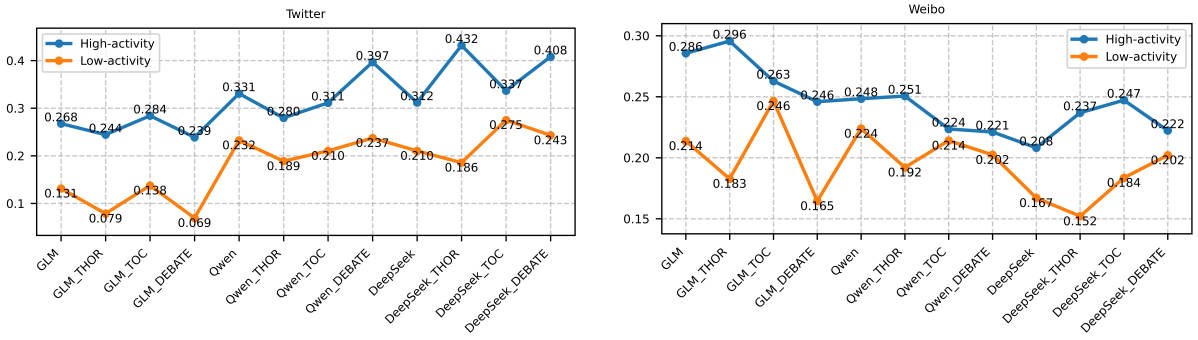
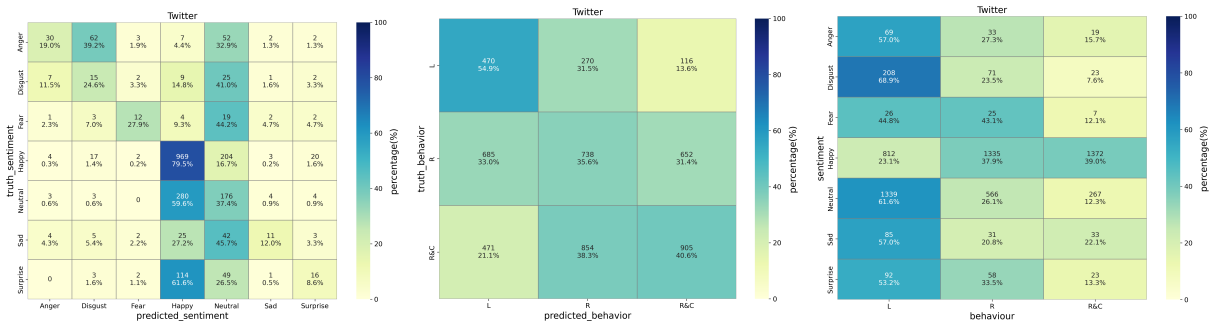
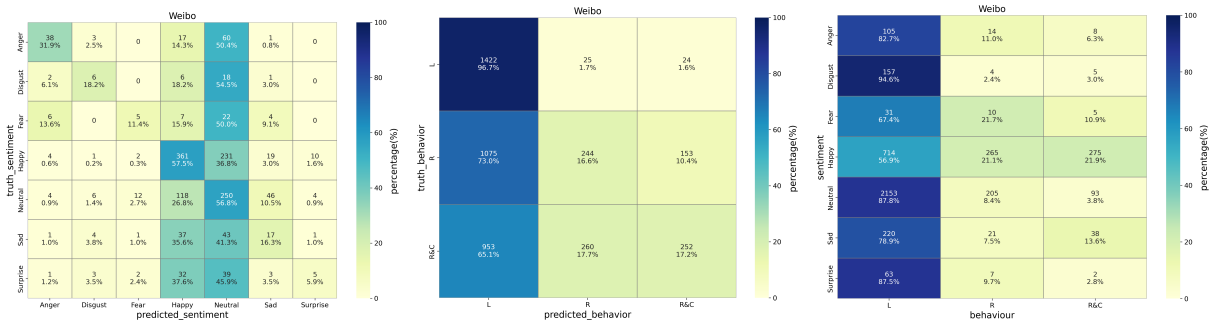


Figure 9: Comparison of high- and low- activity users for Task C.



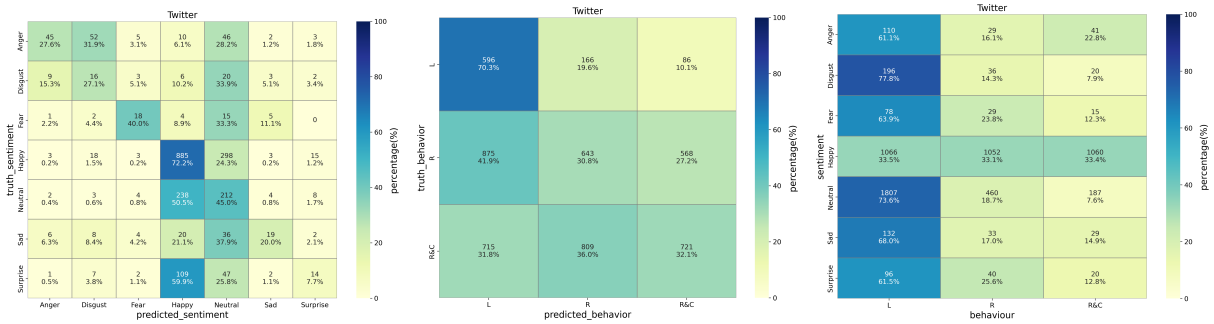
(a) EECM (b) BBCM (c) EBCM

Figure 10: Confusion matrices of emotion-behavior joint prediction using GLM in Twitter.



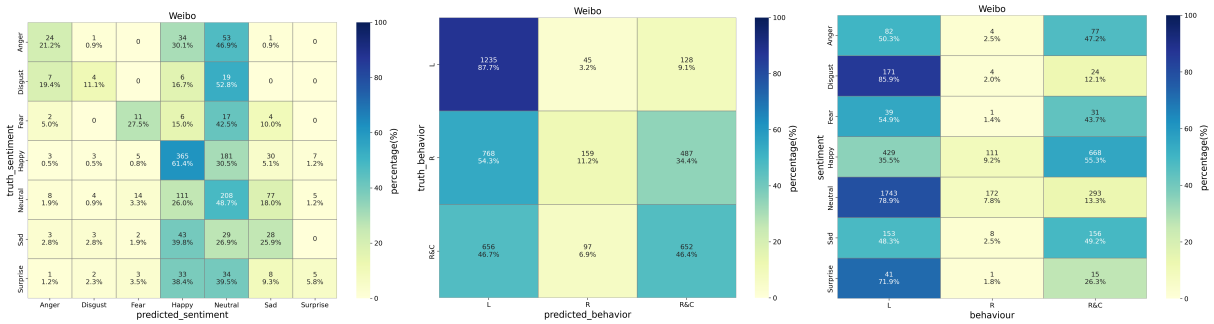
(a) EECM (b) BBCM (c) EBCM

Figure 11: Confusion matrices of emotion-behavior joint prediction using GLM in Weibo.



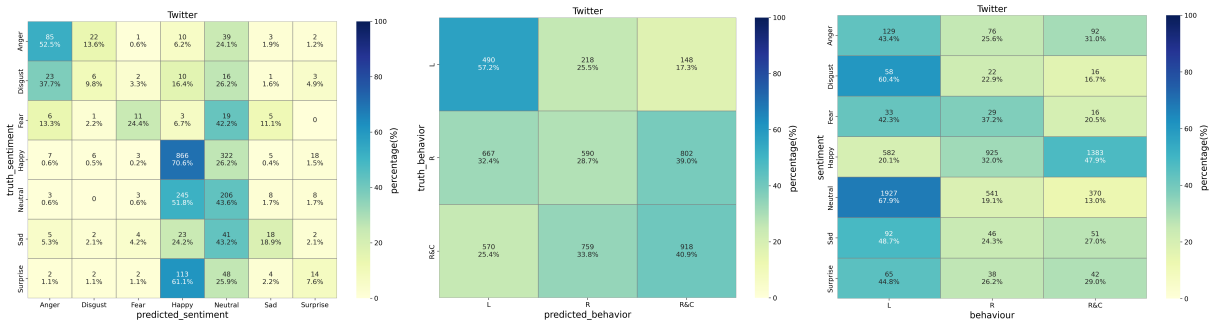
(a) EECM (b) BBCM (c) EBCM

Figure 12: Confusion matrices of emotion-behavior joint prediction using Qwen in Twitter.



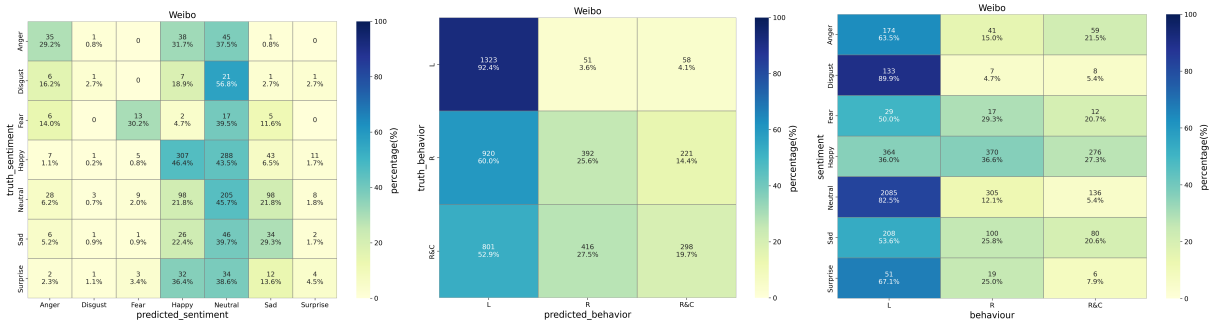
(a) EECM (b) BBCM (c) EBCM

Figure 13: Confusion matrices of emotion-behavior joint prediction using Qwen in Weibo.



(a) EECM (b) BBCM (c) EBCM

Figure 14: Confusion matrices of emotion-behavior joint prediction using DeepSeek in Twitter.



(a) EECM (b) BBCM (c) EBCM

Figure 15: Confusion matrices of emotion-behavior joint prediction using DeepSeek in Weibo.

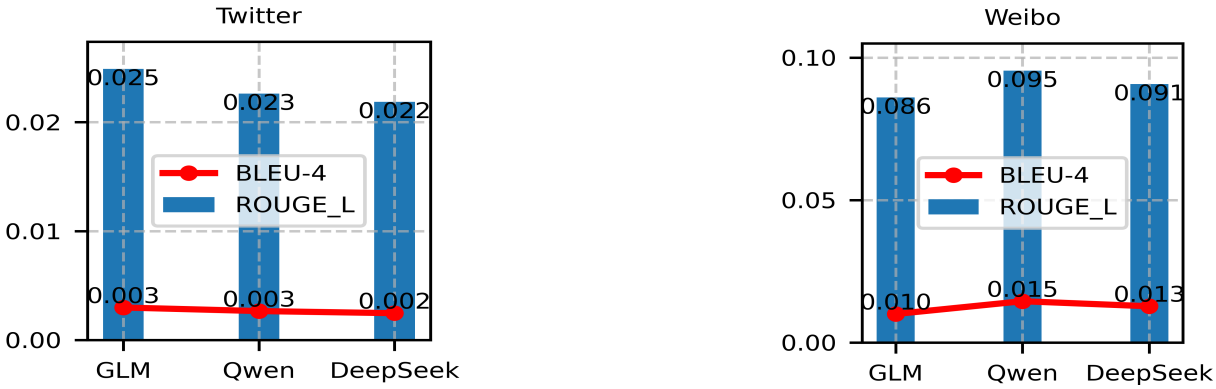


Figure 16: BLEU and Rouge-L of Task E based on different LLMs.

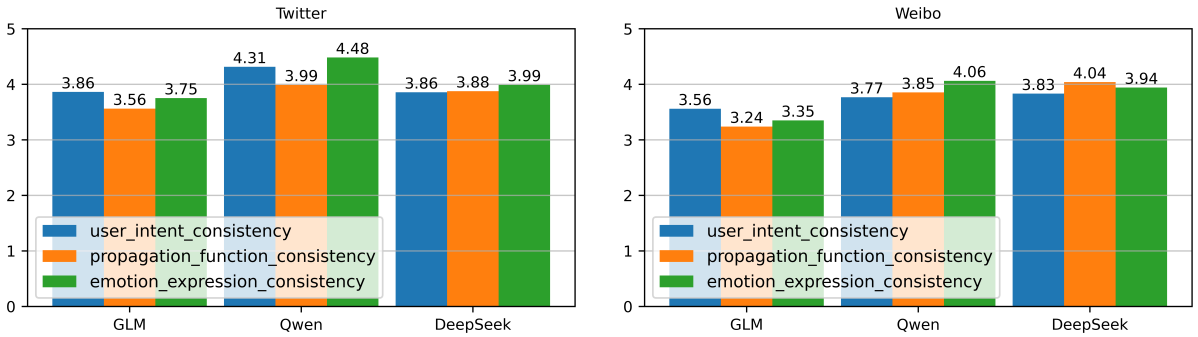


Figure 17: Comparison of UIC, PFC, and EEC based on different LLMs.

Task A: Prompt for Reader Emotion Classification (Basic LLM)

```
{"role": "system",
 "content": """You are an emotion analysis expert. Your task is to
              analyze the reply [reply_text] to the original tweet
              [retweet_text] provided by the user, and determine which of
              the following seven emotions it belongs to:
              ["Happy","Anger","Neutral","Sad","Surprise","Fear","Disgust"].
               Requirements:
               1. Output in pure JSON format.
               2. Strictly follow the JSON format: {"retweet_text":"original
text","reply_text":"reply","emotion":"emotion label"}.
               3. No additional explanations are allowed, including not using
Markdown code blocks, only return a simple JSON string."""
},
{"role": "user",
 "content": "retweet_text:{},reply_text:{}" }
```

Figure 18: Prompt for Task A (basic LLM).

Task A: Prompt for Reader Emotion Classification (LLM + THOR)

Prompt1:

{"role": "system",
"content": """You are an emotion analysis expert. Your task is to analyze the explicit target of the reply [reply_text] based on the original text [retweet_text] provided by the user.
Requirements are as follows:
1. Output in pure JSON format.
2. JSON format strictly follows: {"retweet_text":"original text","reply_text":"reply","target":"target"}.
3. No additional explanations are allowed, including not using Markdown code blocks, only return a simple JSON string."""},
{"role": "user",
"content": "retweet_text:{},reply_text:{}"}

Prompt2:

{"role": "system",
"content": """You are an emotion analysis expert. Your task is to analyze the core attitude expressed by the replier based on the original text [retweet_text], reply [reply_text], and the target [target] of the reply provided by the user.
Requirements are as follows:
1. Output in pure JSON format.
2. JSON format strictly follows: {"retweet_text":"original text","reply_text":"reply","attitude":"attitude description"}.
3. No additional explanations are allowed, including not using Markdown code blocks, only return a simple JSON string."""},
{"role": "user",
"content": "retweet_text:{},reply_text:{},target:{}"}

Prompt3:

{"role": "system",
"content": """You are an emotion analysis expert. Your task is to analyze which of the following seven emotions the reply belongs to based on the original text [retweet_text], reply [reply_text], target [target] of the reply, and the core attitude [attitude] expressed by the replier: ["Happy","Anger","Neutral","Sad","Surprise","Fear", "Disgust"].
Requirements are as follows:
1. Output in pure JSON format.
2. JSON format strictly follows: {"retweet_text":"original text","reply_text":"reply","emotion":"emotion category"}.
3. No additional explanations are allowed, including not using Markdown code blocks, only return a simple JSON string."""},
{"role": "user",
"content": "retweet_text:{},reply_text:{},target:{},attitude:{}"}

Figure 19: Prompt for Task A (LLM + THOR).

Task A: Prompt for Reader Emotion Classification (LLM + TOC)

Prompt1:

```
{"role": "system",
 "content": """You are an emotion analysis expert. Your task is to analyze
             the emotion of [reply_text] from the following seven
             emotions:["Happy","Anger","Neutral","Sad","Surprise","Fear",
             "Disgust"] in relation to the original text [retweet_text]
             provided by the user.
             Requirements are as follows:
             1. Output in pure JSON format.
             2. JSON format strictly follows: {"retweet_text":"original text",
             "reply_text":"reply","analysis":{"Happy":"relevant analysis",
             "Anger":"relevant analysis","Neutral":"relevant analysis",
             "Sad":"relevant analysis","Surprise":"relevant analysis",
             "Fear":"relevant analysis","Disgust":"relevant analysis"}}.
             3. No additional explanations are allowed, including do not use
             Markdown code blocks, only return a simple JSON string."""},
{"role": "user",
 "content": "retweet_text:{},reply_text:{}"}
```

Prompt2:

```
{"role": "system",
"content": """You are an emotion analysis expert. Your task is to combine the
             original text [retweet_text] and analysis [analysis] provided by the
             user, and determine which of the following seven emotions the Reply
             [reply_text] belongs to: ["Happy","Anger","Neutral","Sad",
             "Surprise","Fear","Disgust"].
             Requirements are as follows:
             1. Output in pure JSON format.
             2. JSON format strictly follows: {"retweet_text":"original text",
             "reply_text":"reply","emotion":"emotion category"}.
             3. No additional explanations are allowed, including do not use
             Markdown code blocks, only return a simple JSON string."""},
{"role": "user",
 "content": "retweet_text:{},reply_text:{},analysis:{}"}
```

Figure 20: Prompt for Task A (LLM + TOC).

Task A: Prompt for Reader Emotion Classification (LLM + DEBATE)

Prompt1:

{"role": "user",
"content": """Original retweet_text:{}; reply reply_text:{}; emotions: ["Happy","Anger","Neutral", "Sad","Surprise","Fear","Disgust"].
Please explain why the emotion of reply_text is {emotion}.
Output requirements:
1. Must return strict JSON format.
2. JSON format must exactly match: {{"answer":"The emotion of the reply text is {}", "reasoning":"Your reasoning process"}}.
3. Do not add any extra explanations, including not using Markdown code blocks, return only plain JSON string."""}

Prompt2:

{ "role": "user",
"content": """Original retweet_text:{}; reply reply_text:{}; emotions: ["Happy","Anger","Neutral", "Sad","Surprise","Fear","Disgust"].
Assistant:{Reply of Assistant}
Assistant's answer may be incorrect. Please convince the Assistant that his answer might be wrong.
Output requirements:
1. Must return strict JSON format.
2. JSON format must exactly match: {{"reasoning":"Your reasoning process"}}.
3. Do not add any extra explanations, including not using Markdown code blocks, return only plain JSON string."""}

Prompt3:

{"role": "user",
"content": """Original retweet_text:{}; reply reply_text:{}; emotions: ["Happy","Anger","Neutral", "Sad","Surprise","Fear","Disgust"].
Assistant:{Reply of Assistant}
Critic:{Reply of Critic}
As Assistant, please refute the Critic's answer and convince the critic that your answer is correct.
Output requirements:
1. Must return strict JSON format.
2. JSON format must exactly match: {{"reasoning":"Your reasoning process"}}.
3. Do not add any extra explanations, including not using Markdown code blocks, return only plain JSON string.""".}

Prompt4:

{"role": "user",
"content": """Original retweet_text:{}; reply reply_text:{}; emotions: ["Happy","Anger","Neutral", "Sad","Surprise","Fear","Disgust"].
Which of the seven categories is the emotion of reply_text? Analysis results for each possible category are as follows:
{Integrated Multi-Stance Debate Process Analysis}
After reviewing the above debate process, which option do you think is the most suitable answer? Please provide only the correct answer, no other response.
Output requirements:
1. Must return strict JSON format.
2. JSON format must exactly match: {{"Judgement":"Correct answer is [your answer]", "sentiment":"emotion category","reasoning":"Your reasoning process"}}.
3. Do not add any extra explanations, including not using Markdown code blocks, return only plain JSON string."""}

Figure 21: Prompt for Task A (LLM + Debate).

Task A: Prompt for Personalized Emotion Classification (Basic LLM)

{"role": "system",
"content": """You are an emotion analysis expert. Your task is to analyze the emotion expressed in the reply [reply_text] towards the original post [retweet_text] and categorize it into one of the following seven emotions: ["Happy","Anger","Neutral","Sad","Surprise","Fear","Disgust"].
Input data description:
retweet_text: Original post content.
reply_text: User reply content.
author: Simple information about the post author, including userid and username.
target_user: Simple information about the target user, including userid and username.
target_user_info: Personal information of the target user (viewer).
target_user_history: Historical behavior data of the target user.
author_info: Personal information of the original post author.
author_history: Historical data of the post author.
Requirements:
1. Output should be pure JSON format.
2. JSON format must strictly follow: {"retweet_text":"original text","reply_text":"reply", "emotion":"emotion label"}.
3. Do not include any additional explanations, including Markdown code blocks, return only a plain JSON string."""},
{"role": "user",
"content": "retweet_text:{},reply_text:{},author:{},target_user:{},author_info:{},target_user_info:{}, author_history:{},target_user_history:{}"}

Figure 22: Prompt for Task A_P (basic LLM).

Task A: Prompt for Personalized Emotion Classification (LLM + THOR)

Prompt1:

```
{"role": "system",
"content": """You are an emotion analysis expert. Your task is to analyze the specific target of the reply [reply_text]
           based on the information provided by the user.
           Input data description:
           retweet_text: Original tweet content.
           reply_text: User reply content.
           author: Simple information about the tweet author, including userid and username.
           target_user: Simple information about the target user, including userid and username.
           target_user_info: Personal information of the target user (viewer).
           target_user_history: Historical behavior data of the target user.
           author_info: Personal information of the tweet author.
           author_history: Historical data of the tweet author.
           Requirements:
           1. Output in pure JSON format.
           2. JSON format must strictly follow: {"retweet_text":"original text","reply_text":"reply","target":"target"}.
           3. No additional explanations are allowed, including not using Markdown code blocks, only return a plain
           JSON string."""},
{"role": "user",
 "content": "retweet_text:{},reply_text:{},author:{},target_user:{},author_info:{},target_user_info:{},author_history:{},
            target_user_history:{}"}
```

Prompt2:

```
{"role": "system",
"content": """You are an emotion analysis expert. Your task is to analyze the core attitude expressed by the replier based
           on the information provided by the user.
           Input data description:
           retweet_text: Original tweet content.
           author: Simple information about the tweet author, including userid and username.
           target_user: Simple information about the target user, including userid and username.
           target_user_info: Personal information of the target user (viewer).
           target_user_history: Historical behavior data of the target user.
           author_info: Personal information of the tweet author.
           author_history: Historical data of the tweet author.
           target: The specific object that the reply [reply_text] is addressing.
           Requirements:
           1. Output in pure JSON format.
           2. JSON format must strictly follow: {"retweet_text":"original text","reply_text":"reply",
           "attitude":"attitudedescription"}.
           3. No additional explanations are allowed, including not using Markdown code blocks, only return a plain
           JSON string."""},
{"role": "user",
 "content": "retweet_text:{},reply_text:{},author:{},target_user:{},author_info:{},target_user_info:{},
            author_history:{},target_user_history:{},target:{}"}
```

Prompt3:

```
{"role": "system",
"content": """You are an emotion analysis expert. Your task is to analyze which of the following seven emotions the reply
           belongs to, based on the information provided by the user: ["Happy","Anger","Neutral","Sad",
           "Surprise","Fear","Disgust"].
           Input data description:
           retweet_text: Original tweet content.
           reply_text: User reply content.
           author: Simple information about the tweet author, including userid and username.
           target_user: Simple information about the target user, including userid and username.
           target_user_info: Personal information of the target user (viewer).
           target_user_history: Historical behavior data of the target user.
           author_info: Personal information of the tweet author.
           author_history: Historical data of the tweet author.
           target: The specific object that the reply [reply_text] is addressing.
           attitude: The core attitude expressed by the replier.
           Requirements:
           1. Output in pure JSON format.
           2. JSON format must strictly follow: {"retweet_text":"original text","reply_text":"reply",
           "emotion":"emotion category"}.
           3. No additional explanations are allowed, including not using Markdown code blocks, only return a plain
           JSON string."""},
{"role": "user",
 "content": "retweet_text:{},reply_text:{},author:{},target_user:{},author_info:{},target_user_info:{},author_history:{},
            target_user_history:{},target:{},attitude:{}"}
```

Figure 23: Prompt for Task A_P (LLM + THOR).

Task A: Prompt for Personalized Emotion Classification (LLM + TOC)

Prompt1:

{"role": "system",
"content": """You are an emotion analysis expert. Your task is to analyze the emotion expressed in [reply_text] towards the original post [retweet_text] from the following seven emotional perspectives: ["Happy","Anger", "Neutral","Sad","Surprise","Fear","Disgust"].
Input data description:
retweet_text: Original post content.
reply_text: User reply content.
author: Simple information about the post author, including userid and username.
target_user: Simple information about the target user, including userid and username.
target_user_info: Personal information of the target user (viewer).
target_user_history: Historical behavior data of the target user.
author_info: Personal information of the original post author.
author_history: Historical data of the post author.
Requirements:
1. Output should be pure JSON format.
2. JSON format must strictly follow: {"retweet_text":"original text","reply_text":"reply", "analysis":{"Happy":"relevant analysis","Anger":"relevant analysis","Neutral":"relevant analysis", "Sad":"relevant analysis","Surprise":"relevant analysis","Fear":"relevant analysis","Disgust":"relevant analysis"}}.
3. Do not include any additional explanations, including Markdown code blocks, return only a plain JSON string."""},
{"role": "user",
"content": "retweet_text:{},reply_text:{},author:{},target_user:{},author_info:{},target_user_info:{}, author_history:{},target_user_history:{}"}

Prompt2:

{"role": "system",
"content": """You are an emotion analysis expert. Your task is to determine which of the following seven emotions the reply [reply_text] belongs to: ["Happy","Anger","Neutral","Sad","Surprise","Fear","Disgust"] based on the information provided by the user.
Input data description:
retweet_text: Original post content.
reply_text: User reply content.
author: Simple information about the post author, including userid and username.
target_user: Simple information about the target user, including userid and username.
target_user_info: Personal information of the target user (viewer).
target_user_history: Historical behavior data of the target user.
author_info: Personal information of the original post author.
author_history: Historical data of the post author.
analysis: Analysis of the seven emotions.
Requirements:
1. Output should be pure JSON format.
2. JSON format must strictly follow: {"retweet_text":"original text","reply_text":"reply", "emotion":"emotion category"}.
3. Do not include any additional explanations, including Markdown code blocks, return only a plain JSON string."""},
{"role": "user",
"content": "retweet_text:{},reply_text:{},author:{},target_user:{},author_info:{},target_user_info:{}, author_history:{},target_user_history:{},analysis:{}"}

Figure 24: Prompt for Task A_P (LLM + TOC).

Task A: Prompt for Personalized Emotion Classification (LLM + DEBATE)

Prompt1:
{"role": "user",
"content": """Original retweet_text:{} Reply reply_text:{}, basic information of original poster:{}, target user basic information:{}, original poster personal information:{}, target user personal information:{}, original poster historical records:{}, target user historical records:{};
emotions: ["Happy","Anger","Neutral","Sad","Surprise","Fear","Disgust"].
Please explain why the emotion of reply_text is {}.
Output requirements:
1. Must return in strict JSON format.
2. JSON format must perfectly match: {{"answer":"The emotion of the reply text is {}",
"reasoning":"Provideyour reasoning process as simply as possible"}}.
3. Do not add any additional explanation, including not using Markdown code blocks, return only plain JSON string."""}

Prompt2:
{"role": "user",
"content": """Original retweet_text:{}, Reply reply_text:{}, basic information of original poster:{}, target user basic information:{}, original poster personal information:{}, target user personal information:{}, original poster historical records:{}, target user historical records:{};
emotions: ["Happy","Anger","Neutral","Sad","Surprise","Fear","Disgust"].
Assistant:{}
The Assistant's answer may be incorrect. Please persuade the Assistant that his answer might be wrong.
Output requirements:
1. Must return in strict JSON format.
2. JSON format must perfectly match: {{"reasoning":"Provide your reasoning process as simply as possible"}}.
3. Do not add any additional explanation, including not using Markdown code blocks, return only plain JSON string."""}

Prompt3:
{"role": "user",
"content": """Original retweet_text:{}, Reply reply_text:{}, basic information of original poster:{}, target user basic information:{}, original poster personal information:{}, target user personal information:{}, original poster historical records:{}, target user historical records:{};
emotions: ["Happy","Anger","Neutral","Sad","Surprise","Fear","Disgust"].
Assistant:{}
Critic:{}
As Assistant, please refute the Critic's answer and convince the critic that your answer is correct.
Output requirements:
1. Must return in strict JSON format.
2. JSON format must perfectly match: {{"reasoning":"Provide your reasoning process as simply as possible"}}.
3. Do not add any additional explanation, including not using Markdown code blocks, return only plain JSON string."""}

Prompt4:
{"role": "user",
"content": """Original retweet_text:{}, Reply reply_text:{}, basic information of original poster:{}, target user basic information:{}, original poster personal information:{}, target user personal information:{}, original poster historical records:{}, target user historical records:{};
emotions: ["Happy","Anger","Neutral","Sad","Surprise","Fear","Disgust"].
Which of the seven categories is the emotion of reply_text? The analysis results for each possible category are as follows:
{Integrated Multi-Stance Debate Process Analysis}
After reviewing the above debate process, which option do you think is the most suitable answer? Please provide only the correct answer, no other response.
Output requirements:
1. Must return in strict JSON format.
2. JSON format must perfectly match: {{"Judgement":"Correct answer is [your answer]",
"sentiment":"emotion category","reasoning":"Provide your reasoning process as simply as possible"}}.
3. Do not add any additional explanation, including not using Markdown code blocks, return only plain JSON string."""}

Figure 25: Prompt for Task A_P (LLM + Debate).

Task C: Prompt for Reader Behavior Prediction (Basic LLM)

{"role": "system",
"content": """You are a user behavior analysis expert. Your task is to analyze the target user's behavior based on the original text [retweet_text] provided by the user.
Input data description:
retweet_text: Original tweet content.
author: Brief information about the tweet author, including userid and username.
target_user: Brief information about the target user, including userid and username.
target_user_info: Personal information of the target user (viewer).
target_user_history: Historical behavior data of the target user.
author_info: Personal information of the tweet author.
author_history: Historical data of the tweet author.
Behavior types to predict (choose one of three):
1. "Retweet and Reply";
2. "Retweet but no Reply",
3. "No Retweet and no Reply".
Output requirements:
1. Must strictly return in JSON format;
2. The JSON result must exactly match: {"author":{"userid":"author's userid",
"username":"author's username"},"target_user":{"userid":"target user's userid",
"username":"target user's username"},"retweet_text":"original tweet content",
"predicted_behavior":"behavior type","reasoning":"simple reasoning process"};
3. Do not add any extra explanation, including do not use Markdown code blocks, just return a plain JSON string."""},
{"role": "user",
"content": "retweet_text:{},author:{},target_user:{},author_info:{},target_user_info:{},
author_history:{},target_user_history:{}"}

Figure 26: Prompt for Task C (basic LLM).

Task C: Prompt for Reader Behavior Prediction (LLM + THOR)

Prompt1:

```
{"role": "system",
"content": """You are a user behavior analysis expert. Your task is to analyze the potential relationship between the target
           user and the original tweet [retweet_text] and the tweet author based on the information provided by the user.
           Input data description:
           retweet_text: Original tweet content.
           author: Brief information about the tweet author, including userid and username.
           target_user: Brief information about the target user, including userid and username.
           target_user_info: Personal information of the target user (viewer).
           target_user_history: Historical behavior data of the target user.
           author_info: Personal information of the tweet author.
           author_history: Historical data of the tweet author.
           Behavior types to predict (choose one of three):
           1. "Retweet and Reply";
           2. "Retweet but no Reply",
           3. "No Retweet and no Reply".
           Output requirements:
           1. Must strictly return in JSON format;
           2. The JSON result must exactly match: {"author":{"userid":"author's userid",
           "username":"author's username"},"target_user":{"userid":"target user's userid",
           "username":"target user's username"},"retweet_text":"original tweet content","relationship":"simple analysis of
           the potential relationship between the target user and the original tweet and tweet author"};
           3. Do not add any extra explanation, including not using Markdown code blocks, just return a plain
           JSON string."""},
{"role": "user",
"content":"retweet_text:{},author:{},target_user:{},author_info:{},target_user_info:{},author_history:{},
          target_user_history:{}"}
```

Prompt2:

```
{"role": "system",
 "content": """You are a user behavior analysis expert. Your task is to analyze the target user's participation motivation
            and tendency towards the original tweet [retweet_text] based on the information provided by the user.
            Input data description:
            retweet_text: Original tweet content.
            author: Brief information about the tweet author, including userid and username.
            target_user: Brief information about the target user, including userid and username.
            target_user_info: Personal information of the target user (viewer).
            target_user_history: Historical behavior data of the target user.
            author_info: Personal information of the tweet author.
            author_history: Historical data of the tweet author.
            relationship: Potential relationship between the target user and the original tweet and its author.
            Behavior types to predict (choose one of three):
            1. "Retweet and Reply";
            2. "Retweet but no Reply",
            3. "No Retweet and no Reply".
            Output requirements:
            1. Must strictly return in JSON format;
            2. The JSON result must exactly match: {"author":{"userid":"author's userid",
            "username":"author's username"},"target_user":{"userid":"target user's userid",
            "username":"target user's username"},"retweet_text":"original tweet content",
            "analysis":"simple analysis of the target user's participation motivation and tendency towards the original
            tweet"}};
            3. Do not add any extra explanation, including not using Markdown code blocks, just return a plain
            JSON string."""},
{"role": "user",
"content": "retweet_text:{},author:{},target_user:{},author_info:{},target_user_info:{},author_history:{},
           target_user_history:{},relationship:{}"}
```

Prompt3:

```
{"role": "system",
 "content": """You are a user behavior analysis expert. Your task is to analyze the behavior that the target user will
            exhibit towards the original tweet [retweet_text] based on the information provided by the user.
            Input data description:
            retweet_text: Original tweet content.
            author: Brief information about the tweet author, including userid and username.
            target_user: Brief information about the target user, including userid and username.
            target_user_info: Personal information of the target user (viewer).
            target_user_history: Historical behavior data of the target user.
            author_info: Personal information of the tweet author.
            author_history: Historical data of the tweet author.
            relationship: Potential relationship between the target user and the original tweet and its author.
            analysis: Target user's participation motivation and tendency towards the original tweet.
            Behavior types to predict (choose one of three):
            1. "Retweet and Reply";
            2. "Retweet but no Reply",
            3. "No Retweet and no Reply".
            Output requirements:
            1. Must strictly return in JSON format;
            2. The JSON result must exactly match: {"author":{"userid":"author's userid",
            "username":"author's username"},"target_user":{"userid":"target user's userid",
            "username":"target user's username"},"retweet_text":"original tweet content",
            "predicted_behavior":"behavior type"}};
            3. Do not add any extra explanation, including not using Markdown code blocks, just return a plain
            JSON string."""},
{"role": "user",
 "content": "retweet_text:{},author:{},target_user:{},author_info:{},target_user_info:{},author_history:{},
            target_user_history:{},relationship:{},analysis:{}"}
```

Figure 27: Prompt for Task C (LLM + THOR).

Task C: Prompt for Reader Behavior Prediction (LLM + TOC)

Prompt1:

{"role": "system",
"content": """You are a user behavior analysis expert. Your task is to analyze the potential behavior of the target user towards the original text [retweet_text] based on the information provided by the user from the following three behaviors: ["Retweet and reply", "Retweet but no reply", "No retweet and no reply"].
Input data description:
retweet_text: Original tweet content.
author: Brief information about the tweet author, including userid and username.
target_user: Brief information about the target user, including userid and username.
target_user_info: Personal information of the target user (viewer).
target_user_history: Historical behavior data of the target user.
author_info: Personal information of the tweet author.
author_history: Historical data of the tweet author.
Behavior types to predict (choose one of three):
1. "Retweet and Reply";
2. "Retweet but no Reply",
3. "No Retweet and no Reply".
Output requirements:
1. Must strictly return in JSON format;
2. The JSON result must exactly match: {"author":{"userid":"author's userid", "username":"author's username"},"target_user":{"userid":"target user's userid", "username":"target user's username"},"retweet_text":"original tweet content", "analysis":{"Retweet and Reply":"simple reasoning process","Retweet but no Reply":"simple reasoning process","No Retweet and no Reply":"simple reasoning process"}};
3. Do not add any extra explanation, including do not use Markdown code blocks, just return a plain JSON string."""},
{"role": "user",
"content": "retweet_text:{},author:{},target_user:{},author_info:{},target_user_info:{},author_history:{}, target_user_history:{}"}

Prompt2:

{"role": "system",
"content": """You are a user behavior analysis expert. Your task is to analyze the potential behavior of the target user towards the original tweet [retweet_text] based on the information provided by the user.
Input data description:
retweet_text: Original tweet content.
author: Brief information about the tweet author, including userid and username.
target_user: Brief information about the target user, including userid and username.
target_user_info: Personal information of the target user (viewer).
target_user_history: Historical behavior data of the target user.
author_info: Personal information of the tweet author.
author_history: Historical data of the tweet author.
analysis: Analysis of three behaviors.
Behavior types to predict (choose one of three):
1. "Retweet and Reply";
2. "Retweet but no Reply",
3. "No Retweet and no Reply".
Output requirements:
1. Must strictly return in JSON format;
2. The JSON result must exactly match: {"author":{"userid":"author's userid", "username":"author's username"},"target_user":{"userid":"target user's userid", "username":"target user's username"},"retweet_text":"original tweet content", "predicted_behavior":"behavior type"};
3. Do not add any extra explanation, including not using Markdown code blocks, just return a plain JSON string."""},
{"role": "user",
"content": "retweet_text:{},author:{},target_user:{},author_info:{},target_user_info:{},author_history:{}, target_user_history:{},analysis:{}"}

Figure 28: Prompt for Task C (LLM + TOC).

Task C: Prompt for Reader Behavior Prediction (LLM + DEBATE)

Prompt1:

{"role": "user",
"content": """Original retweet_text:{}, Original author basic information:{}, Target user basic information:{},
Original author personal information:{}, Target user personal information:{}, Original author's historical records:{},
Target user's historical records:{}, Behavior type: ["Retweet and Reply","Retweet but no Reply",
"No Retweet and no Reply"].
Please explain why the target user's behavior upon seeing the retweet_text is {}.
Output requirements:
1. Must return in strict JSON format.
2. JSON format must exactly match: {{"answer":"Target user's behavior is {}",
"reasoning":"your reasoning process"}}
3. Do not add any extra explanation, including not using Markdown code blocks, return plain JSON string only."""}

Prompt2:

{"role": "user",
"content": """Original retweet_text:{}, Original author basic information:{}, Target user basic information:{},
Original author personal information:{}, Target user personal information:{}, Original author's historical records:{},
Target user's historical records:{}, Behavior type: ["Retweet and Reply","Retweet but no Reply",
"No Retweet and no Reply"].
Assistant:{}
Assistant's answer may be incorrect. Please convince the Assistant that his answer may be wrong.
Output requirements:
1. Must return strict JSON format.
2. JSON format must exactly match: {{"reasoning":"your reasoning process"}}
3. Do not add any additional explanation, including not using Markdown code blocks, return only
plain JSON string."""}

Prompt3:

{"role": "user",
"content": """Original retweet_text:{}, Original author basic information:{}, Target user basic information:{},
Original author personal information:{}, Target user personal information:{}, Original author's historical records:{},
Target user's historical records:{}, Behavior type: ["Retweet and Reply","Retweet but no Reply",
"No Retweet and no Reply"].
Assistant:{}
Critic:{}
As Assistant, please refute Critic's answer and convince the critic that your answer is correct.
Output requirements:
1. Must return in strict JSON format.
2. JSON format must exactly match: {{"reasoning":"your reasoning process"}}
3. Do not add any extra explanation, including not using Markdown code blocks, return plain JSON string only."""}

Prompt4:

{"role": "user",
"content": """Original retweet_text:{}, Original author basic information:{}, Target user basic information:{},
Original author personal information:{}, Target user personal information:{}, Original author's historical records:{},
Target user's historical records:{}, Behavior type: ["Retweet and Reply","Retweet but no Reply",
"No Retweet and no Reply"].
Which of the three behavior types might the target user exhibit towards the retweet_text? Analysis results for each
possible type are as follows:
{Integrated Multi-Stance Debate Process Analysis}
After reviewing the above debate, which option do you think is the most appropriate answer? Please provide only the
correct answer without any other response.
Output requirements:
1. Must return strict JSON format.
2. JSON format must exactly match: {{"Judgement":"The correct answer is [your answer]",
"behavior":"behavior type","reasoning":"your reasoning process"}}
3. Do not add any additional explanation, including not using Markdown code blocks, return only plain
JSON string."""}

Figure 29: Prompt for Task C (LLM + Debate).

Task D: Prompt for Joint Emotion-Behavior Modeling

Prompt1:

```
{"role": "system",
"content": """You are a social behavior prediction expert. Your task is to predict the possible behavior of a target user towards a
           tweet based on the provided social data.
           Input data description:
           retweet_text: Original tweet content.
           author: Basic information of the tweet author, including id and screen_name.
           target_user: Basic information of the target user, including id and screen_name.
           target_user_info: Personal information of the target user (viewer).
           target_user_history: Historical behavioral data of the target user.
           author_info: Personal information of the tweet author.
           author_history: Historical data of the tweet author.
           Behavior types to predict (choose one of three):
           1. "Retweet and Reply";
           2. "Retweet but no Reply",
           3. "No Retweet and no Reply".
           Output requirements:
           1. Must strictly return in JSON format;
           2. JSON result must exactly match: {"author":{"id":"author_id","screen_name":"author_screen_name"},
           "target_user":{"id":"target_user_id","screen_name":"target_user_screen_name"},
           "retweet_text":"original_tweet_content","predicted_behavior":"behavior_type",
           "reasoning":"simple_reasoning_process"};
           3. Do not add any extra explanation, including not using Markdown code blocks, only return plain JSON string."""},
{"role": "user",
"content": "retweet_text:{},author:{},target_user:{},author_info:{},target_user_info:{},author_history:{},
           target_user_history:{}"}
```

Prompt2:

```
{"role": "system",
 "content": """You are an emotion analysis expert. Your task is to predict the emotion of a target user towards a tweet based on
           the provided social data.
           Input data description:
           retweet_text: Original tweet content.
           author: Basic information of the tweet author, including id and screen_name.
           target_user: Basic information of the target user, including id and screen_name.
           target_user_info: Personal information of the target user (viewer).
           target_user_history: Historical behavioral data of the target user.
           author_info: Personal information of the tweet author.
           author_history: Historical data of the tweet author.
           Emotion types to analyze (choose one of seven): ["Happy","Sad","Neutral","Anger","Fear","Disgust","Surprise"].
           Output requirements:
           1. Must strictly return in JSON format;
           2. JSON result must exactly match: {"author":{"id":"author_id","screen_name":"author_screen_name"},
           "target_user":{"id":"target_user_id","screen_name":"target_user_screen_name"},
           "retweet_text":"original_tweet_content","predicted_sentiment":"emotion_label",
           "reasoning":"simple_reasoning_process"};
           3. Do not add any extra explanation, including not using Markdown code blocks, only return plain JSON string."""},
{"role": "user",
  "content": "retweet_text:{},author:{},target_user:{},author_info:{},target_user_info:{},author_history:{},
           target_user_history:{}"}
```

Prompt3:

```
{"role": "system",
 "content": """You are a comprehensive social behavior analysis expert. Your task is to provide a final comprehensive judgment
           based on the provided social data and the existing behavior and emotion prediction results of the target user.
           Input data description:
           retweet_text: Original tweet content.
           author: Basic information of the tweet author, including id and screen_name.
           target_user: Basic information of the target user, including id and screen_name.
           target_user_info: Personal information of the target user (viewer).
           target_user_history: Historical behavioral data of the target user.
           author_info: Personal information of the tweet author.
           author_history: Historical data of the tweet author.
           predict_behavior: Existing predicted behavior, including behavior type and simple reasoning process.
           predict_sentiment: Existing predicted emotion, including emotion category and simple reasoning process.
           Behavior types to predict (choose one of three): ["Retweet and Reply", "Retweet but no Reply",
           "No Retweet and no Reply"].
           Emotion types to analyze (choose one of seven): ["Happy","Sad","Neutral","Anger","Fear","Disgust","Surprise"].
           You need to consider comprehensively:
           1. Relevance of the original social data;
           2. Reasonableness of the existing prediction results;
           3. Provide a final comprehensive judgment.
           Output requirements:
           1. Must strictly return in JSON format;
           2. JSON result must exactly match: {"author":{"id":"author_id","screen_name":"author_screen_name"},
           "target_user":{"id":"target_user_id","screen_name":"target_user_screen_name"},
           "retweet_text":"original_tweet_content","final_behavior":{"behavior":"predicted_behavior_type",
           "reasoning":"simple_reasoning_process"},"final_sentiment":{"sentiment":"predicted_emotion_category",
           "reasoning":"simple_reasoning_process"}};
           3. Do not add any extra explanation, including not using Markdown code blocks, only return plain JSON string."""},
{"role": "user",
 "content": "retweet_text:{},author:{},target_user:{},author_info:{},target_user_info:{},author_history:{},
           target_user_history:{},predict_behavior:{},predict_sentiment:{}"}
```

Figure 30: Prompt for Task D.

Task E: Prompt for In-context Comment Generation

{"role": "system",
"content": """You are a user behavior analysis expert. Your task is to simulate the replies that the target user might generate after seeing the original text [retweet_text] based on the user-provided information.
Input data description:
retweet_text: Original tweet content.
author: Brief information about the tweet author, including userid and username.
target_user: Brief information about the target user, including userid and username.
target_user_info: Personal information of the target user (viewer).
target_user_history: Historical behavioral data of the target user.
author_info: Personal information of the tweet author.
author_history: Historical data of the tweet author.
Output requirements:
1. Must strictly return in JSON format;
2. The JSON result must exactly match: {"author":{"userid":"author userid", "username":"author username"},"target_user":{"userid":"target user userid", "username":"target user username"},"retweet_text":"original tweet content", "generate_reply":"generated reply"};
3. Do not add any extra explanations, including not using Markdown code blocks, just return a plain JSON string."""},
{"role": "user",
"content": "retweet_text:{},author:{},target_user:{},author_info:{},target_user_info:{}, author_history:{},target_user_history:{}"}

Figure 31: Prompt of in-context comment generation for Task E.

Prompt of UIC, PFC, EEC measurements for Task E

{"role": "user",
"content": """
You are a social media reply quality evaluation expert. Your task is to assess the quality of generated replies based on the following three metrics, providing integer scores between [1-5] based on the information provided by the user:
1. user_intent_consistency: Measures whether the generated reply aligns with the user intent and core information of the real reply [reply_text].
2. propagation_function_consistency: Measures whether the generated reply exhibits similar propagation functions (such as triggering interactions, reinforcing viewpoints, promoting sharing) as the real reply [reply_text].
3. emotion_expression_consistency: Measures whether the generated reply maintains consistency with the real reply [reply_text] in terms of emotional tendency, intensity, and expression style.
Input Data Description:
retweet_text: Original tweet content.
reply_text: Real reply content.
generate_reply_text: Generated reply content.
Output Requirements:
1. Must strictly return JSON format;
2. The JSON result must exactly match:
{"reply_text":"reply_text","retweet_text":"retweet_text",
"metric":{"user_intent_consistency":"score",
"propagation_function_consistency":"score",
"emotion_expression_consistency":"score"}}.
3. Do not add any additional explanations, including not using Markdown code blocks, only return plain JSON string."""}
{"role": "user",
"content": "retweet_text:{},reply_text:{},generate_reply_text:{}"}

Figure 32: Prompt of UIC, PFC, EEC measurements for Task E.