BEYOND TOPIC RECOGNITION: SOCIAL-FUNCTION UNDERSTANDING IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) have made significant progress in social media mining, yet existing evaluations predominantly focus on surface-level topics and content categories, leaving the communicative functions of online interactions comparatively underexplored. To address this gap, we introduce a theory-informed framework for social-function understanding in social media. Guided by Granovetter's tie strength theory, Katz's uses-and-gratifications theory, and Habermas's public sphere theory, we construct a social-function taxonomy with three top-level categories (Relational, Instrumental, Public) and seven fine-grained subcategories. On this basis, we make three contributions: (1) **Reddit-Scene**, a multimodal dataset of 6,003 annotated samples from 22 Reddit communities, organized in a three-tier Post–Pillar Comment–Thread (P-R-T) structure with five micro-sociological features; (2) **SceneBench**, a benchmark evaluating eight frontier LLMs on social-function inference, where the best model reaches only 54.2% seven-class accuracy while a controlled community-identification probe reaches 90.1%, revealing a substantial topic–role gap that is not reliably resolved by additional conversational context; and (3) a controlled multi-agent **behavioral probe** showing that social-function information has an observable influence on downstream agent responses. A fine-tuned encoder reaches 76.2%, and a three-annotator human baseline reaches 78.3% on social-function inference but 84.7% on community identification, further suggesting that the main difficulty lies in social-role inference rather than surface community recognition.

1 INTRODUCTION

As LLMs are increasingly deployed as conversational agents in social platforms, a critical question arises: can they understand not just *what* people say, but *what they are socially doing*? Human communication often serves purposes beyond information transmission—many utterances perform social functions, such as seeking help, sharing an experience, or expressing a stance (Trillò et al., 2022)—and these are not reducible to surface topic alone. Consider two illustrative r/Parenting posts from our dataset (Figure 1). One asks how to avoid accidentally leaving a child in a car seat; the other shares a child's fascination with a refrigerator box. Both concern everyday parenting, yet they perform fundamentally different social actions: the first is an *instrumental request for practical guidance*, whereas the second is a *relational act of everyday sharing*. A model that fails to distinguish these functions risks offering entertainment to someone seeking help, or practical advice to someone sharing a personal moment (Welivita & Pu, 2020). This raises the central question of our work: *Can large language models go beyond what communication is about and infer what it is socially doing?*

However, this capability remains largely absent from existing evaluations. Although current benchmarks assess individual aspects of social reasoning—sentiment, hate speech, theory of mind, deception (Jin et al., 2024; Strachan et al., 2024; Chen et al., 2024; Kang et al., 2025)—they do not evaluate whether models can infer the social function of a communication. Two gaps remain unaddressed: **(i) social function is not an explicit evaluation target**, since related work on dialogue acts treats communicative function only at the utterance level (Joty & Mohiuddin, 2018; Hu et al., 2023); and **(ii) topic understanding does not imply social understanding**, as models may recognize what a post discusses while failing to determine whether it seeks advice, shares an experience, or expresses a stance. These limitations motivate the two-stage evaluation we adopt.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

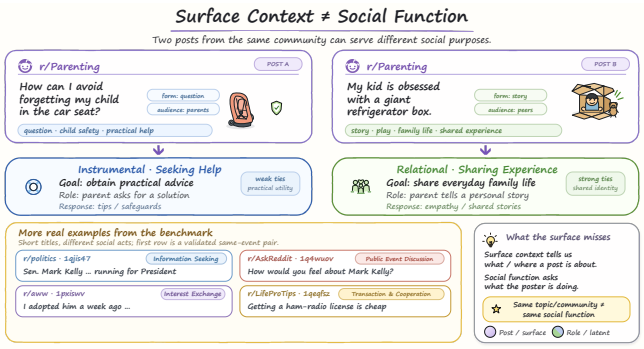


Figure 1: An illustrative pair of r/Parenting posts performing different social actions (instrumental request vs. relational sharing): surface context does not determine function.

To address these gaps, we introduce **SceneBench**, a benchmark for diagnosing social-function understanding in LLMs—how well models infer it, and whether the inferred social-function cues have observable consequences for downstream responses. We first operationalize social function as a theory-informed taxonomy of communicative purposes drawn from social and communication theory: three broad categories—**Relational**, **Instrumental**, and **Public**—refined into seven fine-grained subcategories (§2.1), which characterizes social functions beyond surface topics and community identities. Guided by this taxonomy, we construct **Reddit-Scene**, a multimodal dataset of 6,003 annotated posts from 22 Reddit communities; each instance preserves the original post, its multimodal content, and its Post–Pillar Comment–Thread (P-R-T) structure, allowing us to study how conversational context influences social-function inference. SceneBench evaluates LLMs through two complementary tasks: Task 1 measures whether models can infer social functions under different levels of context, while Task 2 probes whether such inference has measurable consequences for downstream agent responses. We therefore study social-function understanding along three dimensions: inference accuracy, the use of conversational context, and downstream behavioral relevance.

Our experiments reveal three findings. **First**, current LLMs recover social-function cues from the post alone, but do so substantially less reliably than they recognize surface topic or community information, exposing a clear **Topic–Role Gap** between what a post is about and what the poster is socially doing. **Second**, richer conversational context does not reliably resolve this gap: real reply trees degrade social-function inference substantially more than length-matched unrelated text, revealing a *Role-Relevant Context Misuse* rather than a generic long-context effect. **Third**, predicted scene information changes agent responses beyond a no-scene baseline, while additional controls show that instance-specific semantic matching provides incremental value beyond generic contextual and structured-label effects. Together, these results establish a consistent picture: **social perception is fragile under richer context, yet the inferred social-function cues remain behaviorally observable**. We make three contributions:

1. We introduce *social-function understanding* as a new evaluation dimension for LLM social intelligence and construct **Reddit-Scene**, a multimodal conversational dataset with a theory-informed taxonomy that separates communicative role from surface topic and community identity.
2. We introduce **SceneBench**, a benchmark for systematically diagnosing social-function inference across controlled context conditions. SceneBench reveals both a substantial **Topic–Role Gap** and a *Role-Relevant Context Misuse*, in which socially meaningful reply context can distract models from the communicative role of the original poster beyond what can be explained by input length alone.
3. We evaluate the *downstream behavioral relevance* of social-function inference through a controlled multi-agent probe. Ablations disentangle the effects of generic context, structured scene labels, and instance-specific semantic matching on simulated agent responses.

2 SceneBench: Social-Function Taxonomy and Reddit-Scene

We introduce **SceneBench**, a benchmark designed to separate two capabilities that are often conflated: recognizing what a post is about and inferring what the poster is socially doing. As illustrated

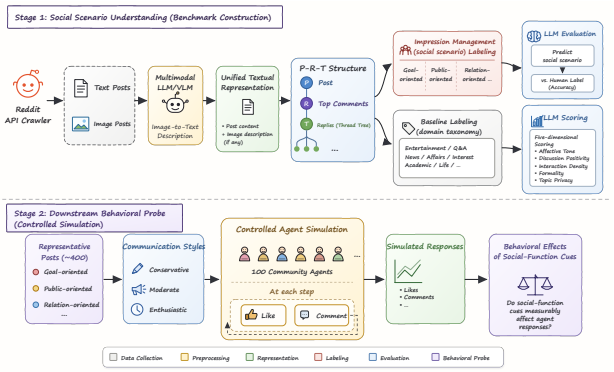


Figure 2: Overview of the SceneBench pipeline. Posts are organized into Post–Pillar Comment–Thread (P-R-T) structures and visual content is converted to text by a multimodal LLM. *Stage 1* builds and evaluates the benchmark; *Stage 2* probes whether social-function cues affect community-agent responses.

in Figure 2, SceneBench consists of three components: (1) **Reddit-Scene**, a multimodal conversational dataset with structured Post–Pillar Comment–Thread (P-R-T) contexts and social-function annotations; (2) a controlled evaluation framework that examines social-function inference under different levels of conversational context; and (3) a controlled multi-agent behavioral probe that tests whether inferred social functions have measurable effects on downstream responses.

2.1 Social-Function Taxonomy

We define the *social function* of a post as the communicative purpose it primarily serves within an interaction. Unlike topical categories, which describe what a post is about, social functions characterize what a post socially *does*, such as maintaining relationships, obtaining practical information, or engaging a broader audience. Our taxonomy is *theory-informed* rather than derived from a single theory, motivated by relational ties and information exchange Granovetter (1973), social and instrumental needs in media use Katz et al. (1973), public discourse Habermas (1962), and the dramaturgical view of social interaction as impression management Goffman (1959). These motivate three broad classes—**Relational**, **Instrumental**, and **Public**—refined into seven fine-grained categories.

Relational functions maintain interpersonal connections and shared experiences (*Family Daily Life*, *Workplace Collaboration*). **Instrumental** functions aim to obtain information, assistance, or practical outcomes (*Interest Exchange*, *Information Seeking*, *Transaction & Cooperation*). **Public** functions address broader audiences through expression or content diffusion (*Public Event Discussion*, *Content Creation & Diffusion*). Because a post may serve multiple purposes simultaneously, annotation identifies the *dominant* communicative function that best characterizes the post within its observed interaction context.

2.2 Dataset Construction

We construct **Reddit-Scene**, a conversational dataset for evaluating social-function understanding in LLMs. Reddit-Scene contains 6,003 annotated posts collected from 22 Reddit communities spanning six broad crawl domains, which are independent of the social-function labels defined in Section 2.1. The complete community coverage, domain mapping, and core statistics are provided in Appendix A.1.

To preserve the conversational environment surrounding each post, we organize every instance into a *Post–Pillar Comment–Thread (P-R-T)* structure: for an original post P, each root comment is a pillar comment R and its nested replies form the corresponding thread T. This separates the author's communication from subsequent audience responses, enabling controlled analysis of how conversational context affects social-function inference. For SceneBench evaluation we serialize each P-R-T structure into a textual sequence under a fixed input budget, with the number of pillar comments and reply depth constrained (Appendix A.2). Because SceneBench evaluates multimodal understand-

Model	Overall		Top-Level Categories			Fine-Grained Categories						
	7-cls	3-cls	Relational	Instrumental	Public	Family Daily	Workplace Collaboration	Interest Exchange	Information Seeking	Transaction & Cooperation	Public Event Discussion	Content Creation & Diffusion
Qwen3.6-Max	**54.2**	**61.9**	62.9	53.0	72.0	64.3	59.3	73.0	29.8	85.5	73.5	51.0
GPT-5.4	53.8	60.4	61.1	43.0	**80.2**	65.5	53.3	48.1	28.6	84.1	**77.4**	**60.4**
Qwen3.6-27B	51.4	58.6	59.7	48.9	69.4	59.5	60.0	65.1	29.5	79.7	71.4	46.6
MiniMax-M2.7	47.3	58.0	63.9	41.1	75.8	72.6	50.0	48.1	19.6	63.8	77.0	47.4
Gemini-3.5-flash	45.2	53.7	47.2	56.1	52.7	59.0	30.5	78.3	27.7	**88.4**	65.1	24.2
GLM-5.1	43.6	52.9	45.8	50.3	58.0	59.5	25.0	77.2	20.3	**88.4**	72.2	22.6
DeepSeek-V4-Flash	42.7	53.7	**75.0**	49.8	52.3	**77.4**	**70.0**	**80.6**	27.6	15.9	75.8	5.6
Qwen3-8B	29.8	43.1	22.2	**58.4**	31.3	26.2	15.0	40.2	**45.3**	8.7	34.7	6.2
Majority-class baseline	*31.6*	*47.3*	–	–	–	–	–	–	–	–	–	–
Dataset n	*6,003*	*6,003*	*686*	*2,840*	*2,477*	*400*	*286*	*614*	*1,897*	*329*	*1,200*	*1,277*

Table 1: SceneBench performance (%) under Full P-R-T. 7-cls/3-cls: seven- and collapsed three-class accuracy; top-level columns report three-class accuracy within that category, fine-grained columns one subcategory. Valid model outputs only; bold marks the best model.

ing through textual interfaces, visual content is converted into complementary descriptions using Qwen-VL-Max Bai et al. (2023) (Appendix A.3).

Each post carries one of the seven labels defined in Section 2.1. Three sociology-trained annotators labeled independently after a two-hour training session and a 50-post calibration set; final labels were decided by majority vote, and ambiguous cases were removed during quality control (Appendix A.4). Pairwise Cohen's κ across the three annotators was 0.78–0.84 (seven-class) and 0.81–0.87 (three-category), indicating substantial agreement. After quality control, Reddit-Scene contains 2,840 Instrumental (47.3%), 2,477 Public (41.3%), and 686 Relational (11.4%) samples. This imbalance reflects the prevalence of communicative functions in large-scale social-media interaction rather than an annotation constraint, and the Relational category still provides sufficient instances for reliable evaluation.

3 Task 1: SceneBench

3.1 Evaluation Setup

We evaluate eight LLMs spanning different model families, scales, and deployment settings (Table 1), covering four widely used families: Qwen3.6 Qwen Team (2026), GPT-5.4 OpenAI (2026), Gemini-3.5-flash Google DeepMind (2026), and DeepSeek-V4-Flash DeepSeek-AI et al. (2026). Qwen3.6-27B-local is a GGUF Q8_0 quantized, dual-GPU deployment. All models use the same seven-class prompt. In the main Full P-R-T setting, each model receives the original post, an image-derived description when available, and the retained P-R-T context; community identity and manually derived features are withheld unless stated in an ablation (Appendix B.1). To confirm that the Chinese interface does not itself depress accuracy on English posts, we repeat the evaluation with English prompts and label names and find a change of at most 2.3 points (Appendix B.9).

3.2 How Well Do LLMs Infer Social Function?

Table 1 reports seven-class social-function accuracy, the corresponding three broad categories, and per-category accuracy.

LLMs capture coarse social-function cues, but fine-grained inference remains difficult. Qwen3.6-Max-Preview achieves the highest seven-class accuracy (54.2%), followed by GPT-5.4 (53.8%) and Qwen3.6-27B-local (51.4%); all models but Qwen3-8B beat the majority-class baseline (31.6%). Collapsing the taxonomy into the three broad categories improves every model (Qwen3.6-Max: 61.9%), indicating that broad interaction types are recognized more readily than fine-grained functions.

The three broad categories differ markedly in difficulty. *Instrumental* is weakest for five of eight models (as low as 41.1%), driven by *Information Seeking* (29.8% even for the best model), whereas *Public* is strongest for five (up to 80.2%). *Relational* spans the widest range (22.2–75.0%) but is the smallest slice, so its figures should be read with caution. The consistent 7-class→3-class gains (6.6–13.3 points) come mainly from collapsing hard *within*-category distinctions, reflecting the removal of fine-grained confusions rather than a resolution of cross-category errors.

Input	Full benchmark				Stratified 500-post subset											
	Qwen3.6-Max				Qwen3.6-Max				DeepSeek-V4-Flash				Qwen3.6-27B			
	All	Pur.	Pub.	Rel.	All	Pur.	Pub.	Rel.	All	Pur.	Pub.	Rel.	All	Pur.	Pub.	Rel.
Full (P–R–T)	54.2	45.5	62.0	62.2	52.0	41.5	61.9	60.3	40.6	33.1	37.9	81.0	50.5	39.6	58.7	65.5
Post-Only	**63.8**	51.3	76.5	69.4	**63.0**	49.2	75.7	74.1	**53.4**	40.3	66.5	60.3	**60.8**	50.8	66.0	82.8
Root-Only			–		55.9	43.0	69.4	65.5	49.6	40.3	56.3	63.8	51.0	42.8	56.8	63.8
Full+Community			–		56.0	45.5	67.6	62.1	44.5	43.8	46.7	39.7	50.0	39.8	58.7	60.3
Random-Text			–		60.2	48.7	71.4	67.2	51.4	42.8	59.2	58.6	59.2	50.8	63.6	77.6

Table 2: Accuracy (%) under five input conditions. **Left block**: the full benchmark (6,003 posts, Qwen3.6-Max only). **Right block**: the stratified 500-post subset shared by all later analyses (hence 52.0% rather than 54.2%). **Pur./Pub./Rel.**: seven-class accuracy within that broad category. Bold marks each model's best condition; valid predictions only.

Errors are highly structured rather than uniform. Qwen3.6-Max-Preview reaches 85.5% on *Transaction & Cooperation* but only 29.8% on *Information Seeking*, and model profiles differ sharply (DeepSeek-V4-Flash is strong on relational categories but poor on *Content Creation & Diffusion*), so aggregate accuracy masks systematic category-specific biases. The most prominent confusion is *Information Seeking* $\rightarrow$ *Public Event Discussion*: 62% of Qwen3.6-Max-Preview's *Information Seeking* posts are predicted as *Public Event Discussion*—88% of its errors on that category (Appendix B.2). The two classes can share nearly identical topical content while differing in the poster's communicative role—requesting knowledge versus expressing a public stance—so the key question is whether models fail because they do not recognize the topic, or because topic recognition does not suffice for social-role inference?

3.3 Do Models Understand Role Beyond Topic?

We compare topical/community recognition with social-function inference on identical inputs, evaluating three models on the stratified 500-post subset. Only the prediction target changes: instead of a social-function label, the model selects the most likely of seven candidate communities (random guessing: 14.3%).

As shown in Appendix Table 8(a), Qwen3.6-Max-Preview, Qwen3.6-27B-local, and DeepSeek-V4-Flash reach 90.1%, 87.4%, and 86.6% on community recognition but only 52.0%, 50.5%, and 40.6% on social function—gaps of 36.9 to 46.0 points. Poor social-function performance therefore cannot be explained by weak topical or community recognition: models recover what a post is broadly about far more readily than what the poster is socially doing, a discrepancy we call the **Topic–Role Gap**. A supplementary analysis of 51 validated same-topic, different-role post pairs—in which models frequently collapse different social functions to the same prediction—is reported in Appendix B.3.

3.4 When Does Conversational Context Help or Hurt?

If replies provide additional social evidence, richer conversational context should improve social-function inference. Counterintuitively, it often does not: on the complete benchmark, Qwen3.6-Max-Preview achieves 54.2% accuracy with the Full P-R-T input but 63.8% under Post-Only, so removing the comment tree improves accuracy by 9.6 percentage points.

A natural alternative is that Full is harder simply because it contains far more text. We therefore repeat the analysis on the stratified 500-post subset across three models (Table 2), where the same ordering holds: relative to Post-Only, unrelated length-matched text costs only 2.8, 2.0, and 1.6 points (Qwen3.6-Max, DeepSeek-V4-Flash, Qwen3.6-27B), whereas the real conversation tree costs 11.0, 12.8, and 10.3 points. Sequence length alone therefore does not explain the degradation; reply semantics contribute to it. All conditions are scored against the same P–R–T-derived gold, so Post-Only is the handicapped condition: where a label depends on the thread (Appendix A.4.5) a post-only model cannot recover it. Its higher accuracy is therefore a conservative estimate. Root-Only also improves over Full P-R-T for all three models, while adding the community name yields only limited gains, suggesting that community identity provides at most a weak prior.

Context is nevertheless not uniformly harmful: DeepSeek-V4-Flash reaches 81.0% on Relational under Full P-R-T versus 58.6% under matched Random-Text, and Figure 5 shows the same flip on a single post. Real replies thus help when they expose relational structure, but current LLMs do not reliably separate role-relevant evidence from less useful context—a pattern we term **Role-Relevant Context Misuse**.

A separate concern is whether the low zero-shot accuracy simply reflects an ambiguous or weakly learnable label space. It does not: on the same 500-post subset and root-only input, supervised encoders reach far higher accuracy than the strongest zero-shot LLM, so the labels carry substantial learnable structure and the low zero-shot figures should not be attributed to intrinsic ambiguity or annotation noise (Section 3.5).

3.5 Human and Supervised Baselines: Where the Gap Actually Is

To determine how much of the remaining gap reflects intrinsic difficulty rather than modeling limitations, we collected human annotations on the same 500 posts and seven-way community options. Three annotators each labeled all 500 items under matched task variables: the same posts, the same seven candidate options, the same task instructions, and the same P–R–T structure. For readability, the human presentation retained up to five nested replies per pillar comment, compared with up to forty in the model input. Human accuracy (Table 3) is **78.3%** on social function (1174/1500; per-annotator 76.0–80.6%) and **84.7%** on community identification (1270/1500; 83.6–86.0%), with every annotator scoring higher on community ($+5.4$, $+10.0$, $+3.8$ points).

To test whether the label space is intrinsically ambiguous or genuinely learnable, we additionally fine-tune standard pre-trained encoders (DeBERTa-v3-base and RoBERTa-base) on the same root-only textual input, training on the remaining labeled posts with duplicate `post_ids` removed to avoid overlap with the test subset. RoBERTa-base reaches **76.2%** seven-class accuracy and 0.788 Macro-F1, essentially at the human level and far above the 55.9% of the strongest zero-shot LLM under the identical input (Appendix B.5). Nor is the gap an artifact of the prompting regime: five labelled in-context exemplars, drawn from labelled posts outside the evaluation subset, lift the three models only to 51.6–55.7% (gains of $+3.6$ to $+12.1$ points, largest for the weakest model), and an explicit chain-of-thought variant behaves the same way, leaving about 20 points to the fine-tuned encoder (Appendix B.10).

Task	Human	Best LLM	Supervised
Social function (7-class)	**78.3%**	52.0% (+26.3)	76.2%
Community (7-way)	84.7%	**90.1%** (-5.4)	–

Table 3: Baselines on the same 500 posts and community options. Human figures aggregate three annotators (1,500 judgments per task) and use the same posts, candidate options, task instructions, and P–R–T structure as the models, with a shallower reply depth for human readability; LLM figures are the matched 500-subset results from Table 1; the supervised column reports RoBERTa-base fine-tuned on the same root-only input, defined only for the social-function task.

Two findings follow. First, both humans and the supervised encoder far outperform the strongest zero-shot LLM on social function (78.3% and 76.2% vs. 52.0%), and every annotator individually exceeds the model, so the accuracy in Section 3.2 reflects a genuine deficiency of the zero-shot paradigm rather than an ambiguous or weakly specified label space. Second—and more diagnostic—the asymmetry *reverses* for community identification: the same annotators reach 84.7%, below all three models probed on this task (Qwen3.6-Max 90.1%, Qwen3.6-27B 87.4%, DeepSeek-V4-Flash 86.6%). With the encoder result, this localizes the failure precisely: *the gap between LLMs and humans lies in the social-role layer, not in the surface layer of topic or community identity.* LLMs match or surpass humans at recognizing what a post is *about* and *where* it was posted, yet do not recover what the poster is socially *doing*. Humans also exhibit the *same* confusion structure (their most frequent error is again Information Seeking→Public Event Discussion) but at a much lower rate, indicating that this confusable pair is intrinsic to the construct rather than a modeling artifact.

Summary of Task 1. Task 1 characterizes the *perception* side. LLMs extract social-function cues from a post alone, but track topical content far more reliably than communicative role (54.2% vs. 90.1% for the best model); richer context is consumed *selectively*; and the gap is not explained by task difficulty, since annotators reach 78.3% and a supervised encoder 76.2% on the same subset. The bottleneck therefore lies in inference without task-specific supervision rather than in the label space, raising the narrower question addressed by Task 2: does such fragile perception nevertheless leave an observable trace in downstream behavior?

4 Task 2: Downstream Behavioral Probe in Multi-Agent Simulation

Building on scene perception, we further investigate a downstream question: *do social-function cues measurably affect the behavior of simulated agents?* While Task 1 evaluates whether models can infer the social function of online interactions, Task 2 examines whether such scene-level information has observable consequences for downstream responses in a controlled multi-agent environment. **Our goal here is not to evaluate optimal strategy selection, nor to claim real-world engagement improvements**: we use the simulation as a behavioral probe.

4.1 Controlled Simulation Setup

We construct a simulation with two types of agents. **Poster agents** rewrite original Reddit-Scene posts under three communication styles—**Conservative**, **Moderate**, and **Enthusiastic**—that vary in positivity and information density while preserving the original topic and social function; three variants are generated for each of the 400 posts (1,200 in total). **Community agents** are heterogeneous LLM-based agents initialized with personality traits, user preferences, and interaction history; each observes a rewritten post and selects one action from $\{\text{like}, \text{comment}, \text{repost}, \text{ignore}\}$. For each post, 100 agents independently generate decisions, and engagement metrics are aggregated across agents.

To examine the effect of social-function conditioning, we compare several conditions while holding the posts, communication styles, community agents, and simulation procedure fixed. In the **Oracle** condition, the human-annotated social-function label is provided to the community agent; in **Predicted**, the social function inferred by the model in Section 3.2 is provided instead. The **No-Scene** condition removes social-function information entirely, while **Random-Text** appends unrelated length-matched text as an additional-context control. We further use a shuffled-label control to distinguish instance-specific semantic matching from effects induced by structured scene labels (Appendix C.2). All remaining components are unchanged across conditions. A rule-driven simulation baseline is reported in Appendix C.1.

4.2 Communication Styles and Simulated Responses

Table 4(a) reports simulated engagement under the three communication styles across the three broad social-function categories. Two patterns are consistent across scenes. First, communication style has a metric-dependent effect: the **Moderate** style receives the most likes in all three social scenes, whereas the **Enthusiastic** style consistently produces the most comments. Different styles therefore induce different forms of engagement rather than a single uniformly superior response pattern. Second, the relative ordering of the three styles remains largely stable across social scenes, although the magnitude of the responses varies, suggesting that social-function information mainly modulates the strength of style effects rather than changing the overall preference among styles. Figure 9 in Appendix C.1 shows the underlying inputs: the poster agent rewrites the *same* original post under the three styles, so the three columns differ only in communication style.

4.3 Behavioral Effects of Social-Function Cues

We next examine whether explicitly providing social-function information changes downstream agent behavior. Table 4(b) compares the Oracle, Predicted, No-Scene, and Random-Text conditions on the same posts, communication styles, and community agents.

Both Oracle and Predicted produce substantially higher aggregate behavioral scores than No-Scene. Predicted and Oracle are numerically similar (82.4 vs. 82.3, $p = 0.67$) despite the social-function

(a) Simulated responses by style and scene

	Relational		Instrumental		Public	
Style	**Likes**	**Comm.**	**Likes**	**Comm.**	**Likes**	**Comm.**
Conservative	70.0	40.3	65.8	37.7	60.9	35.1
Moderate	**77.4**	43.3	**77.3**	45.5	**76.0**	43.1
Enthusiastic	67.6	**46.8**	69.4	**53.3**	66.6	**48.0**

(b) Behavioral score under scene conditioning

Style	**Oracle**	**Pred.**	**No-Sc.**	**Rand.**
Conservative	81.2	80.8	64.2	72.3
Moderate	**87.1**	**87.3**	76.9	82.9
Enthusiastic	78.7	79.2	68.2	71.9
All	*82.3*	*82.4*	*69.8*	*75.7*

Table 4: Downstream behavioral-probe results. (a) Simulated engagement under three communication styles and three broad social-function categories (Comm. = comments); bold marks the highest response per scene and metric. (b) Aggregate behavioral scores under different scene-conditioning conditions, on the same 400 posts and their three rewrites; Predicted uses the outputs of Section 3.2, Random-Text unrelated length-matched context.

prediction accuracy being only 54.2%. Importantly, this numerical similarity should not by itself be interpreted as showing that imperfect predictions preserve all oracle-level semantic information: additional prompt content and structured social-function labels may themselves affect the behavior of LLM-based community agents. Indeed, Random-Text also increases the aggregate score relative to No-Scene (75.7 vs. 69.8), indicating that part of the observed effect is attributable to generic prompt conditioning rather than social-function semantics alone. We therefore introduce an additional control to distinguish these effects.

4.4 Disentangling Context, Label, and Semantic Effects

The Predicted–No-Scene difference can arise from three sources: added prompt content, a structured label, and whether that label is semantically matched to the post. **Random-Text** controls the first with unrelated length-matched text, and **Shuffled-Label** the second by reassigning valid labels across instances—preserving label format and distribution while breaking the post–label correspondence (Appendix C.2). Neither alone explains the effect: added context changes responses, and structured labels condition them further even when mismatched. Yet Predicted exceeds Shuffled-Label by $+2.5$ points, so instance-specific semantic matching contributes a further component—smaller than the Predicted–No-Scene difference, but genuine behavioral signal beyond generic prompt conditioning and label-format effects.

Overall, social-function information has an *observable* influence on downstream agent responses, though the magnitude is behavioral evidence rather than proof of an optimal strategy. The effect operates through **audience modulation**: agents adjust their responses to the provided social context, rather than because social scenes alter poster-agent communication styles.

5 Related Work

Social intelligence and communicative-function understanding. Recent benchmarks examine diverse aspects of social understanding in LLMs: MM-Soc Jin et al. (2024) covers multimodal social-media tasks, SOTOPIA Zhou et al. (2024) social competence through role-playing, ToMBench Chen et al. (2024) Theory of Mind, EgoSocialArena Hou et al. (2024) first-person social intelligence, and MIVA Kang et al. (2025) multimodal deception. These capture important components of social intelligence, but social function—what a post is socially *doing* rather than merely what it is *about*—is rarely a distinct evaluation target. Dialogue-act work does study communicative function Joty & Mohiuddin (2018); Hu et al. (2023), but assumes local, utterance-level settings; social media is instead asynchronous, community-embedded, and unfolds through extended reply structures, which is the setting SceneBench targets.

Theory-informed modeling of social interaction. Computational social science has long connected behavioral data with social theory Lazer et al. (2009). We draw on three perspectives to operationalize social function: Granovetter's relational embeddedness Granovetter (1973), uses-and-gratifications on social and instrumental needs Katz et al. (1973), and Habermas's public sphere Habermas (1962). Our contribution is not a new sociological theory but an operational taxonomy that makes these distinctions measurable in LLM evaluation.

LLM-based social simulation. LLM agents are increasingly used to study collective dynamics Mou et al. (2024): OASIS Yang et al. (2024) scales simulation to large agent populations, FDE-LLM Yao et al. (2025) couples opinion-dynamics models with LLM agents, and SITGNet Zhang et al. (2026) studies intervention in simulated networks, building on DeGroot-style influence dynamics DeGroot (1974); Dong et al. (2024). These ask how opinions evolve once interaction begins; we instead ask whether an agent's *perception of social function* informs its subsequent behavior.

6 Discussion

6.1 From Topic Recognition to Social Role Understanding

Our results indicate that current LLMs remain largely at the level of topical recognition. The dominant confusion—*Information Seeking* versus *Public Event Discussion*—is instructive: both may involve information exchange, yet they correspond to different social positions, one knowledge-seeking and one opinion-sharing, so separating them requires reasoning about the poster's role rather than about the surface topic. Social-function recognition is therefore not a finer-grained form of topic classification but a distinct capability.

6.2 Implications for Social-Function-Aware Systems

Adding conversational context does not by itself close this gap. Including the full comment tree lowers accuracy for most models and categories relative to the post alone—a *Role-Relevant Context Misuse*, since on *Relational* scenes the same reply tree is informative (DeepSeek-V4-Flash: 81.0%, above the length-matched control). The deficiency is thus not an inability to read conversational text, but an inability to identify which parts bear on the poster's role. Together with the encoder result (76.2%), this suggests that the missing signal is not more context but an explicit representation of social role, which may require role-reasoning objectives or additional signals such as user history and cross-post consistency. In simulation, the same information alters how community agents respond without changing the relative ordering of communication styles: it operates through audience modulation rather than style selection, so imperfect perception remains behaviorally observable.

6.3 Limitations and Future Work

Reddit-Scene is drawn from Reddit, so cross-platform generalization remains to be tested. Relational interactions are fewer than other categories, and although overall inter-annotator agreement is substantial, some fine-grained category boundaries remain context-sensitive and are not separately characterized by category-level reliability. The downstream results are derived from a closed LLM-agent simulation rather than real-user interactions, and should therefore be interpreted as simulation-internal behavioral evidence. Task 2 is also conducted in Chinese; although Task 1 shows only small Chinese–English differences, language invariance of the behavioral probe remains to be established. Finally, the simulation uses 100 agents, far fewer than real social networks; future work should examine broader platforms, larger agent societies, and external behavioral validation.

7 Conclusion

This paper examines what a post is *socially doing* rather than merely what it is about. We proposed a theory-informed taxonomy of social functions and constructed **Reddit-Scene**, a multimodal dataset of 6,003 annotated interactions organized as Post–Pillar Comment–Thread trees. **SceneBench** evaluates eight frontier LLMs: the best model reaches only 54.2% seven-class accuracy, whereas a controlled community-identification probe reaches 90.1%. Adding conversational context does not reliably close this gap: the full comment tree often reduces accuracy. In contrast, a fine-tuned encoder reaches 76.2%, and human annotators reach 78.3% on social-function inference but 84.7% on community identification, suggesting that the main bottleneck lies in social-role inference rather than surface community recognition. A controlled multi-agent probe further shows that social-function information has an observable influence on downstream agent responses. We hope this work encourages closer integration of sociological theory with the study of social intelligence in LLMs.

AI USE STATEMENT

In this work, we used generative AI tools for method implementation and data processing (writing and debugging parts of the evaluation harness, the statistical-analysis scripts, and the data-processing code), for result verification (re-computing the reported numbers from the raw per-post records), for drafting and revising portions of the manuscript text (in particular appendix material and the reporting of the few-shot and chain-of-thought controls), and for refining the design of those additional control experiments. We have not used generative AI tools for research ideation or hypothesis generation—the research questions, the social-function taxonomy, and the theoretical grounding were developed by the authors—and formal proofs and translation are not applicable to this work. Additionally, we used generative AI tools for figure creation (the example figures were produced with AI-written matplotlib code) and for readability and grammar editing. We have reviewed all AI-assisted work: every quantitative claim in this paper was re-computed from the raw data with independent verification scripts rather than copied from AI output, all figures were visually inspected, citation metadata was checked against public records, and the dataset labels were produced by trained human annotators rather than by AI. The large language models evaluated in this paper are objects of study, not tools used to generate the scientific content of the paper. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Reddit-Scene is derived from publicly available Reddit content. To limit privacy risks, we do not attempt to re-identify or profile individual users, and we intend to release only post identifiers together with derived labels and features rather than raw user text. Annotation was carried out by trained annotators following the protocol described in Appendix A. The multi-agent simulation is a closed sandbox and is not intended for deployment on real social platforms.

REPRODUCIBILITY STATEMENT

We report model names, deployment settings, decoding parameters, and full prompt templates in Appendix B.1; dataset construction, the annotation protocol, and inter-annotator agreement in Appendix A; and complete experimental configurations for all ablations and for the simulation study in Appendices B and C. The evaluation pipeline, the per-post predictions underlying every reported number, and the released labels and derived features are available at `https://anonymous.4open.science/r/scenebench-anon-6991/`.

REFERENCES

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, et al. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. ToMBench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15959–15983, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.847.

DeepSeek-AI, Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, et al. DeepSeek-V4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Morris H. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69 (345):118–121, 1974.

Yucheng Dong, Zhaogang Ding, and Gang Kou. *Social Network DeGroot Model: Supporting Consensus Reaching in Opinion Dynamics*. Springer Nature Singapore, Singapore, 2024. doi: 10.1007/978-981-97-0421-7.

Erving Goffman. *The Presentation of Self in Everyday Life*. Doubleday Anchor Books, New York, 1959.

Google DeepMind. Gemini 3.5 flash model card, 2026. URL https://deepmind.google/models/model-cards/gemini-3-5-flash/.

Mark S. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.

Jürgen Habermas. *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society*. MIT Press, Cambridge, MA, 1962. English translation published 1989.

Guiyang Hou, Wenqi Zhang, Yongliang Shen, Zeqi Tan, Sihao Shen, and Weiming Lu. EgoSocialArena: Benchmarking the social intelligence of large language models from a first-person perspective. *arXiv preprint arXiv:2410.06195*, 2024.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4194–4213. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.230.

Yiqiao Jin, Minje Choi, Gaurav Verma, Jindong Wang, and Srijan Kumar. MM-Soc: Benchmarking multimodal large language models in social media platforms. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6192–6210, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.370.

Shafiq Joty and Tasnim Mohiuddin. Modeling speech acts in asynchronous conversations: A neural-CRF approach. *Computational Linguistics*, 44(4):859–894, 2018. doi: 10.1162/coli_a_00339.

Caixin Kang, Yifei Huang, Liangyang Ouyang, Mingfang Zhang, and Yoichi Sato. Can MLLMs read the room? a multimodal benchmark for verifying truthfulness in multi-party social interactions. *arXiv preprint arXiv:2510.27195*, 2025.

Elihu Katz, Jay G. Blumler, and Michael Gurevitch. Uses and gratifications research. *Public Opinion Quarterly*, 37(4):509–523, 1973.

David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. Computational social science. *Science*, 323 (5915):721–723, 2009.

Xinyi Mou, Xuanwen Ding, Qi He, Liang Wang, Jingcong Liang, Xinnong Zhang, Libo Sun, Jiayu Lin, Jie Zhou, Xuanjing Huang, and Zhongyu Wei. From individual to society: A survey on social simulation driven by large language model-based agents. *arXiv preprint arXiv:2412.03563*, 2024.

OpenAI. GPT-5.4 thinking system card, 2026. URL https://deploymentsafety.openai.com/gpt-5-4-thinking/.

Qwen Team. Qwen3.6-plus: Towards real world agents, 2026. URL https://qwen.ai/blog?id=qwen3.6.

James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295, 2024. doi: 10.1038/s41562-024-01882-z.

Tommaso Trillò, Blake Hallinan, and Limor Shifman. A typology of social media rituals. *Journal of Computer-Mediated Communication*, 27(4):zmac011, 2022. doi: 10.1093/jcmc/zmac011.

Anuradha Welivita and Pearl Pu. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4886–4899. International Committee on Computational Linguistics, 2020. doi: 10.18653/v1/2020.coling-main.429.

Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, Prateek Gupta, Shuyue Hu, Zhenfei Yin, Guohao Li, Xu Jia, Lijun Wang, Bernard Ghanem, Huchuan Lu, Wanli Ouyang, Yu Qiao, Philip Torr, and Jing Shao. OASIS: Open agent social interaction simulations with one million agents. *arXiv preprint arXiv:2411.11581*, 2024.

Junchi Yao, Hongjie Zhang, Jie Ou, Dingyi Zuo, Zheng Yang, and Zhicheng Dong. Social opinions prediction utilizes fusing dynamics equation with LLM-based agents. *Scientific Reports*, 15: 15472, 2025. doi: 10.1038/s41598-025-99704-3.

Kun Zhang, Xiaoyan Yu, Hao Peng, Zhe Yang, Ye Tian, Hao Jin, Tuoyu Feng, and Hui Lin. Towards efficient optimization of multi-agent social simulation via large language models. *International Journal of Machine Learning and Cybernetics*, 17(1), 2026. doi: 10.1007/s13042-025-02816-5.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *International Conference on Learning Representations (ICLR)*, 2024.

A Dataset Construction Details

A.1 Data Collection

Reddit-Scene was constructed from 22 Reddit communities grouped into six broad crawl domains. The crawl domains were used only to diversify topical and community contexts during data collection and are not used as proxies for social-function labels. This separation prevents community identity from directly determining the target social-function categories. Table 5 summarizes the core statistics of the release, and Table 6 lists the communities in each crawl domain together with their share of the dataset.

A.1.1 Community selection

Communities were chosen to span three axes: *interaction type* (Q&A, news discussion, hobby interest, personal narrative), *community size*, and *moderation regime*, so that the corpus covers a range of social settings rather than a single discourse style. The six crawl domains are therefore collection strata, not analytical units, and no social-function label is defined in terms of them.

A.1.2 Crawling scope

For each community we collected the most recent posts at crawl time, moving backward in time only when the quota was not met: the window begins within the *current week* and is extended to the *current month* and then to the *current year* until enough posts are obtained. We retained English-language posts that had a retrievable comment thread, and kept each post's title, body text, any attached image, and its full comment tree.

A.1.3 Dataset scale

After quality control the release contains 6,003 annotated posts drawn from these 22 communities; Table 5 reports the class counts and structural limits, and Table 6 the per-domain composition.

The domains are unevenly represented: three of them (Entertainment & Humor, Q&A & Discussion, News & Current Affairs) account for 22.7% of the communities each, whereas Professional & Academic is represented by the single community r/science. Domain-level comparisons are therefore reported together with their per-domain sample counts in Appendix B.6.

A.2 Post–Pillar Comment–Thread Construction

To preserve the interaction structure of Reddit discussions, each sample is represented using a Post–Pillar Comment–Thread (P-R-T) format.

Statistic	Value
Communities	22
Crawl domains	6
Annotated posts	6,003
Relational / Instrumental / Public	686 / 2,840 / 2,477
Max. P-R-T reply depth	5
Replies retained per pillar comment	50
Pillar comments per evaluation input	5
Replies per pillar comment in evaluation input	40

Table 5: Core statistics of Reddit-Scene. Class counts are reported after quality control. The last three rows describe the stored structure and the serialized evaluation input, respectively; when a serialized context would exceed the input budget, the number of pillar comments and replies is reduced to three and thirty.

Crawl Domain	Communities	# Communities	Share (%)
Entertainment & Humor	funny, memes, dankmemes, wholesomememes, me_irl	5	22.7
Q&A & Discussion	AskReddit, IAmA, explainlikeimfive, todayilearned, LifeProTips	5	22.7
News & Current Affairs	worldnews, news, politics, upliftingnews, nottheonion	5	22.7
Interest & Hobby	gaming, movies, Music	3	13.6
Professional & Academic	science	1	4.5
Lifestyle & Community	aww, Parenting, maliciouscompliance	3	13.6

Table 6: Community distribution of Reddit-Scene across six crawl domains. Crawl domains were used only to diversify topical and community contexts during data collection and are not used as proxies for social-function labels.

Given an original post P, each independent root comment is represented as a pillar comment R_i, and all nested replies under this root comment form the corresponding thread T_i:

$$P \rightarrow R_i \rightarrow T_i.$$

The stored dataset retains up to 50 replies per pillar comment and supports a maximum reply depth of five levels. For SceneBench evaluation, conversational inputs are serialized with at most five pillar comments and at most forty nested replies per pillar comment; when a serialized context would exceed the input budget, these limits are reduced to three and thirty. The same serialization is used for every model and for all input conditions reported in Sections 3.4 and 3.3.

Therefore, the "Full P-R-T" condition refers to the complete conversational context available under this benchmark input constraint rather than an unrestricted collection of all Reddit replies.

A.3 Visual Content Processing

Since SceneBench models receive textual inputs, visual information in Reddit posts is converted into textual descriptions using Qwen-VL-Max Bai et al. (2023).

We design a complementary-description prompt that instructs the vision model to describe visual information not already provided in the post title or body. This strategy reduces redundant information and encourages the extraction of visually grounded signals relevant to social context.

Images exceeding API input constraints are adaptively resized before processing. For extremely high-resolution images, the longest dimension is limited to 2,048 pixels. The generated visual descriptions are appended to the original post as additional textual context.

The complete prompt template is provided below:

Top-level	Subcategory	Example post	Decisive criterion
Relational	Family Daily Life	"At what point does allowing natural consequences become bad parenting?" (r/Parenting)	Parenting reflection shared within a peer group; nothing is requested from strangers.
Relational	Workplace Collaboration	"Don't want me to do the work myself? Ok, then run and do it yourself." (r/MaliciousCompliance)	Narrates a workplace interaction and its interpersonal consequence; no practical outcome is sought.
Instrumental	Interest Exchange	"What is something older movies cared about that feels mostly lost now?" (r/movies)	Solicits peers' opinions on a shared hobby; asks for views, not a solution.
Instrumental	Information Seeking	"Three Members of Oglala Sioux Tribe Remanded to ICE Detention Center" (r/nottheonion)	Requests factual detail about a specific situation; the need is a concrete fact.
Instrumental	Transaction & Cooperation	"LPT: Load those Visa gift cards onto Amazon to use the full balance" (r/LifeProTips)	Shares advice with an actionable payoff; the exchange yields a practical benefit.
Public	Public Event Discussion	"I'm Senator Chris Murphy. AMA about why Republicans have shut down the government." (r/IAmA)	Invites public debate on a matter of public policy; the poster takes a public stance.
Public	Content Creation & Diffusion	"I'm the creator of 'Plague Inc.' — I just released my new game 'After Inc.' AMA" (r/IAmA)	Promotes a creative work to a broad audience; the aim is diffusion rather than discussion.

Table 7: One annotated example per fine-grained social-function category, with the criterion that decided its label. Posts are abbreviated; annotations were made on the full serialized P-R-T input. The last two rows share a community and a format (AMA) yet differ in social function, illustrating that surface context does not determine the label.

> **Context:** the Reddit post title and the first 300 characters of the post body.
>
> **Task:** describe the image in a natural, flowing English paragraph, as if explaining what you see to a friend who can only read the text but cannot see the image.
>
> **Guidelines:** (1) Be descriptive and visual—focus on the main subjects, their arrangement, colours, textures, and materials. (2) Integrate text naturally: if the image contains text, weave it into your sentences rather than listing it. (3) Avoid redundancy: do not repeat information already stated in the title or body; describe only what is visually present. (4) Use no structural tags: no OCR tags, no "visuals" markers, and no bullet points; write in plain, continuous text.
>
> **Output format:** a single, coherent English paragraph.

A.4 Annotation Protocol

Each Reddit post was assigned one of seven fine-grained social-function labels defined in Section 2.1.

A.4.1 Annotator qualification and training

Three annotators with graduate-level training in sociology independently performed annotation. Before full annotation, annotators completed a two-hour training session introducing the taxonomy and representative examples. They then annotated a 50-post calibration set to establish consistent labeling criteria.

A.4.2 Annotation input

Annotators were shown the same variables as the models (Section A.2): the post title, its body text, the image-derived description when present, and the retained pillar comments with their nested replies, in the same serialized form. The presentation depth was capped for readability—at most five nested replies per pillar comment for annotators, versus forty for the models. How much text a reader actually consults is treated as a property of the reader rather than as an experimental variable; all human–model comparisons therefore hold the items, the options, the format, and the instruction fixed. Each annotator worked independently and did not see the others' labels; labels were also assigned without access to model predictions. Table 7 gives one annotated example per fine-grained category together with the criterion that decided its label; the examples also illustrate the boundary cases discussed below.

A.4.3 CATEGORY BOUNDARY RULES

Four pairs account for most boundary cases. Annotators resolved them in the following order, which follows the taxonomy rather than the community or topic.

1. **Information Seeking vs. Interest Exchange.** Both solicit input from the community. The distinction is what the poster needs: a *fact or solution* about a specific situation (Information Seeking) versus *opinions or shared experience* on a broad topic (Interest Exchange).
2. **Information Seeking vs. Public Event Discussion.** Both may concern the same public event. The distinction is the poster's stance: *requesting* information about the event (Information Seeking) versus *expressing or soliciting views* on it (Public Event Discussion).
3. **Transaction & Cooperation vs. Interest Exchange.** Both offer or request something. Transaction & Cooperation requires an *actionable payoff*; Interest Exchange is discussion pursued for its own sake.
4. **Relational vs. Instrumental.** A post that primarily *maintains a relationship or shares a personal experience* is Relational even when it contains factual content. Instrumental requires the main purpose to be obtaining information, assistance, or a practical outcome.

A.4.4 DOMINANT-LABEL ADJUDICATION

Because a post can serve several purposes at once, annotation identified the *dominant* function under the observed interaction context—the purpose that the post's own framing and the thread's response pattern primarily serve. When two labels appeared equally supported, the boundary rules above were applied in order; posts that remained genuinely ambiguous were removed during quality control rather than forced into a label.

A.4.5 A BOUNDARY CASE

The examples in Table 7 are deliberately clear-cut; real posts are often less so. One calibration item illustrates the difficulty. The r/movies post "David Fincher's 'The Adventures of Cliff Booth', Starring Brad Pitt, Wraps Filming" reads as a plain factual news item when the post is seen alone, and was labeled *Information Seeking* in that setting. With the retained thread—where commenters compare the director's filmography and discuss his earlier work—the dominant function becomes *Interest Exchange*, which is the label the post carries. The case shows both that the boundary between *Information Seeking* and *Interest Exchange* is genuinely narrow, and that the surrounding thread carries information the post alone does not—one reason we annotate on the serialized P-R-T input rather than on the post alone.

A.4.6 QUALITY CONTROL

Posts that were ambiguous, incomplete, or insufficiently interpretable were removed during quality control. Final labels were determined through majority voting among the three annotators.

A.4.7 INTER-ANNOTATOR AGREEMENT

To evaluate annotation reliability, we randomly sampled 100 posts and had the annotators independently re-annotate them. With three annotators, pairwise comparison yields three Cohen's κ values. Agreement ranged from 0.78–0.84 for seven-class labels and from 0.81–0.87 for three-category labels, indicating substantial agreement.

B TASK1 ANALYSIS

B.1 MODEL SPECIFICATIONS AND EVALUATION PROMPTS

B.1.1 MODEL SPECIFICATIONS

We evaluated eight LLMs covering different model families and scales: Qwen3.6-Max-Preview Qwen Team (2026), GPT-5.4 OpenAI (2026), Qwen3.6-27B-local, MiniMax-M2.7, Gemini-3.5-flash Google DeepMind (2026), DeepSeek-V4-Flash DeepSeek-AI et al. (2026), GLM-5.1, and Qwen3-8B. The DeepSeek-V4-Flash results were obtained from its preview release.

The evaluated models include both API-based and locally deployed models. All models received identical inputs under each experimental condition, including the original post content, image-derived textual descriptions, and P-R-T conversational context. The same classification prompt and output format were used across all models to ensure fair comparison.

B.1.2 Evaluation Prompt Template

The evaluation prompt consists of three components: (1) a task instruction defining the objective as social-function inference for online interactions; (2) structured inputs containing the post title, post body, visual description, and P-R-T conversational context (at most five root comments and at most forty nested replies per root comment, with a maximum reply depth of five levels); and (3) a JSON-formatted output containing `predicted_scenario` (one of the seven social-function labels), `predicted_category` (the corresponding broad category), and a short justification (`reasoning`).

The prompt does not employ chain-of-thought prompting. Models are instructed to output the predicted label and a brief justification directly, allowing evaluation of classification performance without additional reasoning supervision. All models are queried at temperature 0.1 with an identical system prompt and identical serialized context. Chain-of-thought and few-shot variants are evaluated separately in Appendix B.10.

The complete prompt template is provided below, translated from Chinese; the evaluation itself was conducted in Chinese, and the label names given to the models are the seven Chinese category names defined in Section 2.1.

> "You are an analyst trained in sociology. In Goffman's theory, social behaviour is an incomplete-information game of impression management. Your task: read the interaction context of a Reddit post (post body → top comments → nested replies) and infer the deep social scene.
>
> *Taxonomy.* (1) **Instrumental sociality** (weak ties, oriented to a concrete goal): *Interest Exchange*, sharing among people with the same hobby or subculture, where the goal is resonance rather than problem solving; *Information Seeking*, providing or requesting factual information and knowledge, with the focus on "what it is" and "how to explain it"; *Transaction & Cooperation*, broadly instrumental knowledge exchange such as money-saving tips, life hacks, and product recommendations. (2) **Public sociality** (addressing a broad, unspecified audience): *Public Event Discussion*, expressing opinions or debating news, current affairs, and social issues, with stance expression as the goal; *Content Creation & Diffusion*, publicly showing original content such as art, memes, videos, or articles, with "being seen" as the goal. (3) **Relational sociality** (strong ties, where the interaction itself is the goal): *Family Daily Life*, personal sharing about marriage, parenting, and family anecdotes, with private topics and high emotional density; *Workplace Collaboration*, interpersonal interaction in a work setting, rooted in organisational relations and professional identity.
>
> *Analysis hints.* (1) Infer the poster's hidden intention—what do they want: resonance, answers, recognition, support, or to share experience? (2) Observe the comment section—information supplementation, opinion debate, empathic comfort, or skill sharing? (3) Judge the audience—are they speaking to insiders, to everyone, or to acquaintances who understand them?
>
> *Output format* (JSON only, no code fences): `{"predicted_scenario": "<one of the seven labels>", "predicted_category": "<instrumental/public/relational>", "reasoning": "<justification, within 60 characters>"}`"

The user message consists solely of the serialized P-R-T context described in Appendix A.2; no additional instructions are appended, so that the input conditions differ across ablations only in the conversational context supplied.

B.2 Fine-grained Error Analysis

This appendix provides detailed error analysis for Task 1, including the confusion matrix of Qwen3.6-Max-Preview and representative systematic error patterns.

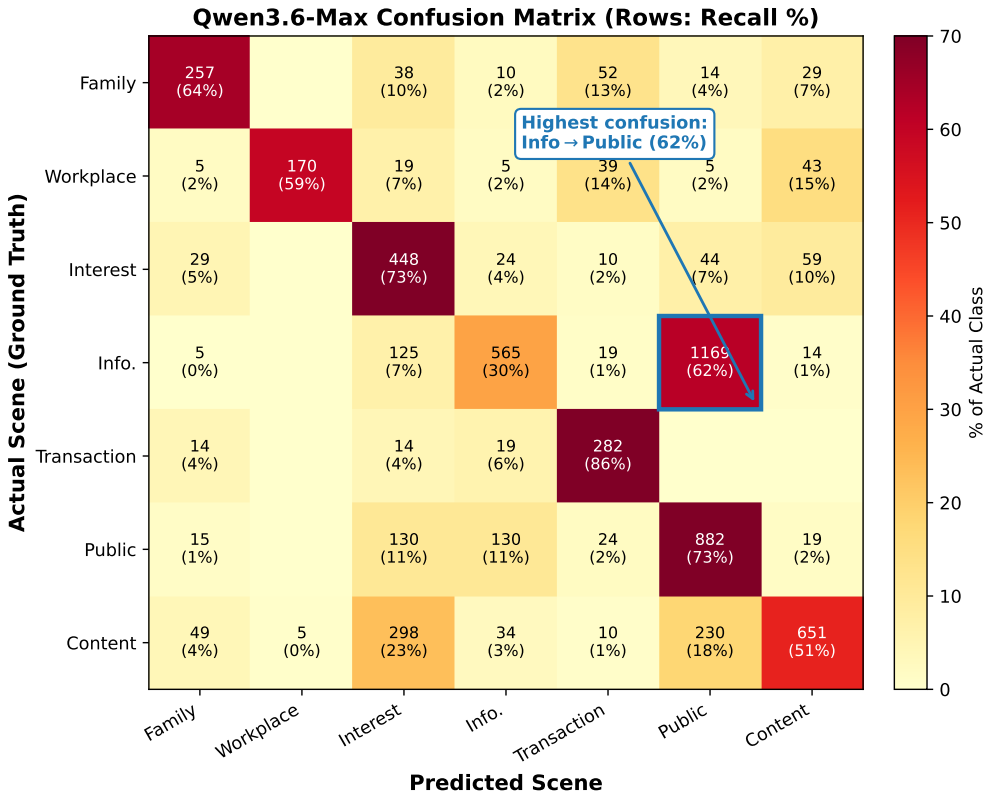


Figure 3: Confusion matrix of Qwen3.6-Max-Preview under the Full P-R-T setting. Cell entries report counts on the 6,003-sample scale used throughout the paper, followed by row-normalized recall percentages computed over valid predictions. The dominant error direction is Information Seeking → Public Event Discussion.

The confusion matrix reveals several systematic patterns.

First, the dominant error occurs between *Information Seeking* and *Public Event Discussion*. For Qwen3.6-Max-Preview, 244 of 396 Information Seeking posts (62%) are assigned to Public Event Discussion, so this single neighbouring class absorbs 88% of the model's errors on the category. Although both categories frequently involve similar topics, they differ in communicative role: the former positions the poster as seeking knowledge, whereas the latter positions the poster as expressing or interpreting a public stance.

Second, *Content Creation & Diffusion* is frequently confused with *Interest Exchange*. This confusion suggests that models may recognize shared topical interests but fail to distinguish between broadcasting content to a broader audience and participating in community-level exchange.

Additional category-level confusion patterns are provided in Figure 3.

B.3 Supplementary Analysis of the Topic–Role Gap

The 500-post diagnostic in Section 3.3 shows that models recover topical and community cues far more accurately than social function. As a complementary diagnostic, we construct 51 same-topic, different-role post pairs—retrieved by TF-IDF cosine similarity and then manually validated—and test whether the model can distinguish their social functions. Because this subset is retrieval-based rather than randomly sampled, we treat it as a diagnostic rather than an independent benchmark.

Qwen3.6-Max-Preview assigns *different* predictions to the two posts in only 3.9% of pairs under both Full and Post-Only inputs; in the remaining 96.1% the pair receives the same predicted label despite different ground-truth social functions, so removing conversational context does not change the collapse. The pattern is asymmetric across the dominant paired categories—*Information Seeking* posts are recovered in only 2.0% of cases versus 98.0% for *Public Event Discussion*—which

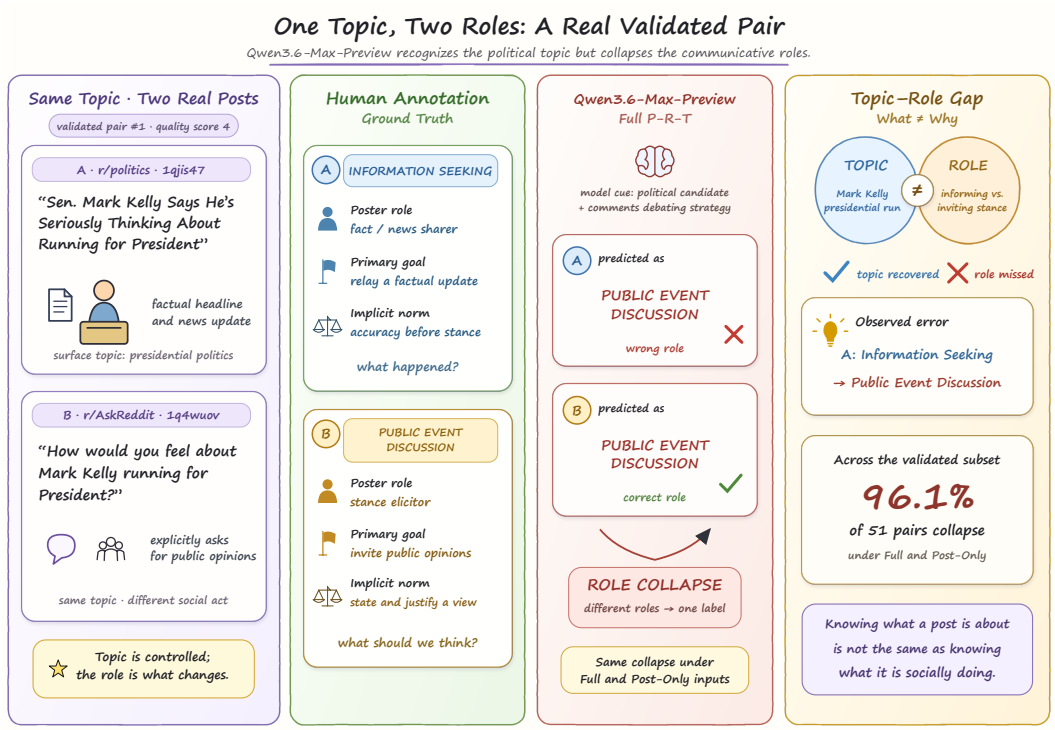


Figure 4: The Topic–Role Gap on a validated same-topic pair (r/politics `1qjis47` and r/AskReddit `1q4wuov`). Both posts concern the same event, but their annotated social functions differ: *Information Seeking* versus *Public Event Discussion*. Qwen3.6-Max-Preview predicts *Public Event Discussion* for both posts under Full and Post-Only inputs, recovering the shared topic while collapsing the distinct communicative roles—the pattern observed in 96.1% of the 51 validated pairs.

suggests the paired result partly reflects a category-specific preference toward the latter and should not be read as standalone evidence that topic alone determines the prediction.

Figure 4 illustrates a representative failure: both posts concern the same concrete event (Senator Mark Kelly considering a presidential run), yet differ in communicative role—the r/politics headline is annotated *Information Seeking*, whereas the r/AskReddit question is annotated *Public Event Discussion*. Qwen3.6-Max-Preview predicts *Public Event Discussion* for both under Full and Post-Only inputs, recovering the shared topic while collapsing the distinct roles. Taken together with the 500-post community-proxy analysis, this paired result provides supplementary evidence for the **Topic–Role Gap**, while the larger diagnostic remains the primary basis for prevalence claims.

B.4 Diagnostic Comparisons: Community Recognition and Supervised Encoders

Table 8 collects the two diagnostic comparisons discussed in Sections 3.3 and 3.4, both computed on the same stratified 500-post subset. Panel (a) contrasts recognition of the community of origin with recognition of social function under an identical Full P-R-T input; the gap between the two quantifies the Topic–Role Gap. Panel (b) contrasts the strongest zero-shot LLM with supervised encoders trained on the same inputs, indicating how much of the social-function label space is learnable. Both panels use valid model outputs only.

Two observations follow. First, in every model the community proxy is recovered far more accurately than social function, with gaps of 36.9–46.0 points, so the low social-function accuracy cannot be attributed to weak topical or community recognition. Second, supervised encoders trained on roughly one tenth of the data reach 70.0–76.2% accuracy, well above the 55.9% of the strongest zero-shot LLM, showing that the label space carries substantial learnable structure. Together, the panels indicate that the bottleneck lies in the zero-shot inference paradigm rather than in the labels themselves; per-category encoder results are reported in Appendix B.5.

(a) Community vs. social function

Model	Comm.	SF	Gap
Qwen3.6-Max	**90.1**	52.0	38.1
Qwen3.6-27B	87.4	50.5	36.9
DeepSeek-V4	86.6	40.6	**46.0**

(b) Zero-shot vs. supervised

Model	Sup.	Acc.	F1
Best zero-shot LLM	No	55.9	0.644
DeBERTa-v3-base	Yes	70.0	0.694
RoBERTa-base	Yes	**76.2**	**0.788**

Table 8: Diagnostic comparisons on the stratified 500-post subset. (a) Community-proxy versus social-function accuracy (%) under identical Full P-R-T inputs; community-proxy chance is 14.3%, and Gap is reported in percentage points. (b) Zero-shot LLM versus supervised encoders under matched root-only inputs, reporting accuracy (%) and Macro-F1.

B.5 Supervised Encoder Baselines

To examine whether the social-function labels contain learnable discriminative structure, we fine-tune standard pre-trained text encoders and evaluate them on the same stratified 500-post test subset used in the LLM diagnostic experiments. To ensure a fair input comparison, both the encoders and the zero-shot LLM baseline use the same root-only textual input. The supervised training set is constructed from the remaining labeled posts, with duplicate `post_id`s removed to prevent overlap with the test set. Because several classes are relatively sparse, we also evaluate class-weighted cross-entropy.

Model	Class Weight	Accuracy	Macro-F1	Weighted-F1
DeBERTa-v3-base (86M)	No	66.8%	0.478	0.637
DeBERTa-v3-base (86M)	Yes	70.0%	0.694	0.678
RoBERTa-base (125M)	Yes	**76.2%**	**0.788**	**0.760**
Qwen3.6-Max-Preview, zero-shot	*N/A*	*55.9%*	*0.644*	*0.554*

Table 9: Supervised encoder baselines and the strongest zero-shot LLM baseline on the same stratified 500-post test subset using the same root-only text input. Class Weight indicates whether class-weighted cross-entropy is used during supervised training.

Results. Table 9 shows that supervised encoders substantially outperform the strongest zero-shot LLM baseline under the same test subset and input setting. DeBERTa-v3-base reaches 66.8% accuracy without class weighting and 70.0% with class weighting, while RoBERTa-base achieves 76.2% accuracy and 0.788 Macro-F1. In comparison, Qwen3.6-Max-Preview reaches 55.9% accuracy and 0.644 Macro-F1 under the same root-only evaluation setting.

Class weighting is particularly important for the less frequent classes. Without weighting, DeBERTa-v3-base fails to recover some sparse categories, whereas class-weighted training improves both Macro-F1 and overall accuracy. This suggests that class imbalance contributes to the difficulty of the supervised setting but does not prevent the label structure from being learned.

The advantage of the supervised encoders is also visible at the broad category level. RoBERTa-base achieves 77.1% accuracy on Instrumental, 72.8% on Public, and 84.5% on Relational. These results show that the social-function taxonomy contains substantial learnable signal even with a relatively small supervised training set.

Overall, the supervised baselines indicate that the low zero-shot LLM performance should not be attributed solely to intrinsic task ambiguity or annotation noise. Instead, the results reveal a substantial gap between zero-shot social-function inference and supervised task-specific learning.

B.6 Performance Across Crawl Domains

To examine whether topical domains or community identities serve as shortcuts for social-function inference, we evaluate model performance across the six crawl domains used during dataset construction.

Table 10 reports seven-class accuracy across domains. Performance varies substantially across domains, and domain-level rankings do not consistently follow the overall ranking. For example, Qwen3-8B achieves relatively strong performance on *Professional & Academic* despite its lower

Model	Ent./ Humor	Q&A/ Disc.	News/ Curr.	Interest/ Hobby	Prof./ Acad.	Lifestyle/ Comm.
Qwen3.6-Max	50.2	65.0	**41.9**	34.8	58.3	72.8
GPT-5.4	**58.8**	**70.3**	41.1	28.7	50.0	62.2
Qwen3.6-27B	46.7	65.7	40.7	**35.4**	58.3	63.3
MiniMax-M2.7	48.1	58.7	39.5	30.9	35.0	58.3
Gemini-3.5-flash	24.9	59.3	39.1	34.3	61.7	65.6
GLM-5.1	24.6	57.7	41.1	34.3	38.3	60.6
DeepSeek-V4-Flash	11.4	53.7	40.7	34.3	55.0	**81.7**
Qwen3-8B	10.0	35.7	36.0	21.9	**68.3**	34.4
Dataset n	*1,377*	*1,429*	*1,205*	*848*	*286*	*858*

Table 10: Seven-class accuracy (%) across crawl domains. Ent./Humor = Entertainment & Humor; Q&A/Disc. = Q&A & Discussion; News/Curr. = News & Current Affairs; Prof./Acad. = Professional & Academic; Lifestyle/Comm. = Lifestyle & Community. The bottom row reports the number of samples in each domain.

overall accuracy, while DeepSeek-V4-Flash performs strongly on *Lifestyle & Community* but poorly on *Entertainment & Humor*.

In contrast, model performance on *News & Current Affairs* remains within a relatively narrow range (36–42%). These patterns suggest that domain-specific differences are more consistent with interactions between domain distributions and model-specific category preferences rather than uniformly stronger social-function understanding in particular domains.

Together, these results indicate that crawl domain provides contextual cues but is not a reliable proxy for social function. The same domain can contain diverse communicative purposes, while similar functions may appear across different communities.

B.7 Additional Qualitative Error Analysis

To better understand the failure modes behind social-function inference errors, we manually inspected 60 misclassified samples from the strongest-performing model (Qwen3.6-Max). Beyond the representative example discussed in the main text, we identified several recurring patterns. Overall, these errors reveal that models often rely on surface textual cues, including topic, vocabulary, and post format, while failing to recover the poster's underlying social role, interaction goal, or implicit communicative norm.

Case 1: Information Seeking misclassified as Interest Exchange. A representative example is a post from r/explainlikeimfive:

> "Why do some metals rust and others don't?"

The model predicted *Interest Exchange*, likely because the topic (chemistry) can appear in interest-oriented discussions. However, the human annotation identifies the post as *Information Seeking*. The author's role is a knowledge seeker who requests an explanation, as indicated by the "why" question and the explanatory nature of the discussion. This error illustrates a typical Topic–Role Gap: the model captures the topical domain but overlooks the communicative purpose of obtaining information.

Case 2: Family Daily Life misclassified as Interest Exchange. A post from r/Parenting asks:

> "Looking for advice on how to deal with my toddler's tantrums."

The model predicts *Interest Exchange* because the question format resembles general advice or experience-sharing discussions. However, the actual social function is *Family Daily Life*. The interaction is grounded in a parent–child relationship and seeks support within a family context rather than exchanging knowledge around a shared interest. This case shows that advice-seeking language alone is insufficient for identifying social function; models must also consider the relational context surrounding the interaction.

Case 3: Public Event Discussion misclassified as Content Creation & Diffusion. A post discussing a current event contains a meme-oriented framing, leading the model to predict *Content Creation & Diffusion*. However, the human annotation identifies the post as *Public Event Discussion*, because the primary purpose is expressing an interpretation or stance toward a public issue rather than distributing content. The error suggests that models may overweight surface format cues, such as the presence of memes, while underestimating the intended social purpose of the post.

Case 4: Workplace Collaboration misclassified as Transaction & Cooperation. A post from r/maliciouscompliance describes a workplace conflict in which an employee refuses to follow a supervisor's request to misrepresent information to a client. The model predicts *Transaction & Cooperation* due to the presence of client, delivery, and conflict-related terms. However, the actual function is *Workplace Collaboration*, where the central issue is the employee–employer relationship and the implicit norms governing workplace behavior rather than resource exchange or task coordination.

Common Error Pattern. Across these cases, the model's errors share a common structure. The model successfully detects surface-level properties of a post, such as topic, lexical cues, or conversational format, but often fails to infer the latent social role of the poster. For example, a question may indicate either information seeking or interest exchange; an advice request may reflect either relational support or general knowledge sharing. Distinguishing these cases requires understanding not only what information appears in a conversation, but also why the author initiates the interaction. These observations support our central finding that current LLMs exhibit a Topic–Role Gap in social-function understanding.

B.8 Qualitative Illustration of the Context-Control Ablations

This appendix illustrates how the input of each ablation condition is constructed on real posts. Post and comment text is quoted verbatim; long spans are abbreviated with ellipses.

Figure 5 walks through the context ladder on a single r/Parenting post (gold label: *Family Daily*). With the post alone (P) and with the root comments added (P–R), Qwen3.6-27B-local reads the request for lunch advice as *Information Seeking*: both inputs are equally compatible with a request for practical guidance. Only when each root comment is followed by its own nested replies (P–R–T) does the exchange reveal an everyday family-sharing situation, and the prediction flips to the gold label.

Figure 6 shows the length-matched control on an r/IAmA post. Replacing the real thread with equal-length text sampled from an unrelated post keeps the input size identical but removes exactly the information that identifies the poster's situation; the drop under Full P-R-T therefore cannot be attributed to sequence length.

Figure 7 contrasts the two auxiliary signals added on top of an otherwise identical P–R–T input: the five human-derived micro-sociological features, and the community name. The feature condition was evaluated for Qwen3.6-Max-Preview only, where it raises overall accuracy by $+0.2$ points on the full benchmark; the community condition adds $+4.0$ points on the 500-post subset, less than the $+8.2$ points obtained from unrelated length-matched text. Neither signal recovers the gold label for the post shown in the figure, indicating that auxiliary surface signals provide only weak priors for social-function prediction.

B.9 Cross-lingual Control

The dataset consists of English Reddit posts, whereas the evaluation prompt and the label names used throughout the paper are Chinese, and the Task 2 simulation also runs in Chinese. To test whether this language mismatch–rather than social-function inference itself–accounts for the low zero-shot accuracy, we repeat the Full P-R-T evaluation with an *English* system prompt and English label names, keeping every other factor fixed: the same posts, the same serialized P-R-T context, and the same decoding parameters. Only the language of the interface changes. We run this control on the stratified 500-post subset for the two models with the largest evaluation budget; Table 11 reports the result.

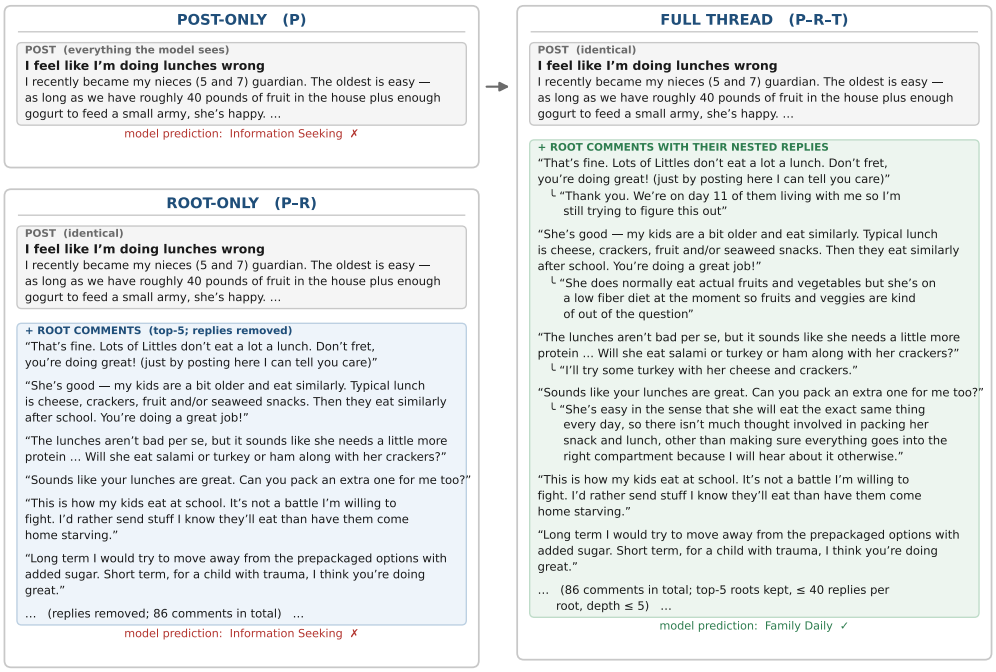


Figure 5: Input construction of the context ladder on a real r/Parenting post (gold label: *Family Daily*). Left: the post-only (P) and root-only (P–R) inputs. Right: the full P–R–T input, in which each retained root comment is followed by its own nested replies. Predictions are from Qwen3.6-27B-local; ellipses abbreviate long spans.

Model	Chinese	English	Δ
Qwen3.6-Max-Preview	52.0	**53.5**	+1.5
DeepSeek-V4-Flash	40.6	42.9	+2.3

Table 11: Cross-lingual control on the stratified 500-post subset. "Chinese" is the setup used throughout the paper (Chinese prompt and label names); "English" replaces both with their English counterparts and leaves everything else unchanged. Accuracy is seven-class (%) and follows the same convention as every other table—valid predictions only—so the "Chinese" column reproduces the corresponding baselines of Table 1 (52.0 and 40.6).

Changing the interface language moves seven-class accuracy by at most 2.3 points, and in the direction of a small *improvement* rather than a loss; the two versions also agree on 86.2% and 68.4% of posts respectively. The low zero-shot accuracy reported in the main text is therefore a property of the models' social-function inference rather than an artefact of evaluating English posts through a Chinese interface.

B.10 In-Context and Reasoning Variants

A natural concern is that the low zero-shot accuracy is an artifact of the prompting regime rather than of inference without task-specific supervision. We therefore repeat the evaluation on the same stratified 500-post subset, with the same Full P–R–T input, the same system prompt, and the same decoding parameters, under two additional regimes: (i) *five-shot*, which inserts five labelled exemplars as in-context turns before the query—each drawn from a different category, sampled from the labelled posts that lie outside the evaluation subset, with each exemplar's context truncated to 1,000 characters—and (ii) an explicit *chain-of-thought* variant, which asks the model to write out its analysis of the poster's intent, the comment behaviour, the audience, and the near-miss labels before emitting the final JSON. Everything else, including the taxonomy and the output format, is unchanged. Figure 8 contrasts the three regimes on one evaluated post. Table 12 reports the results.

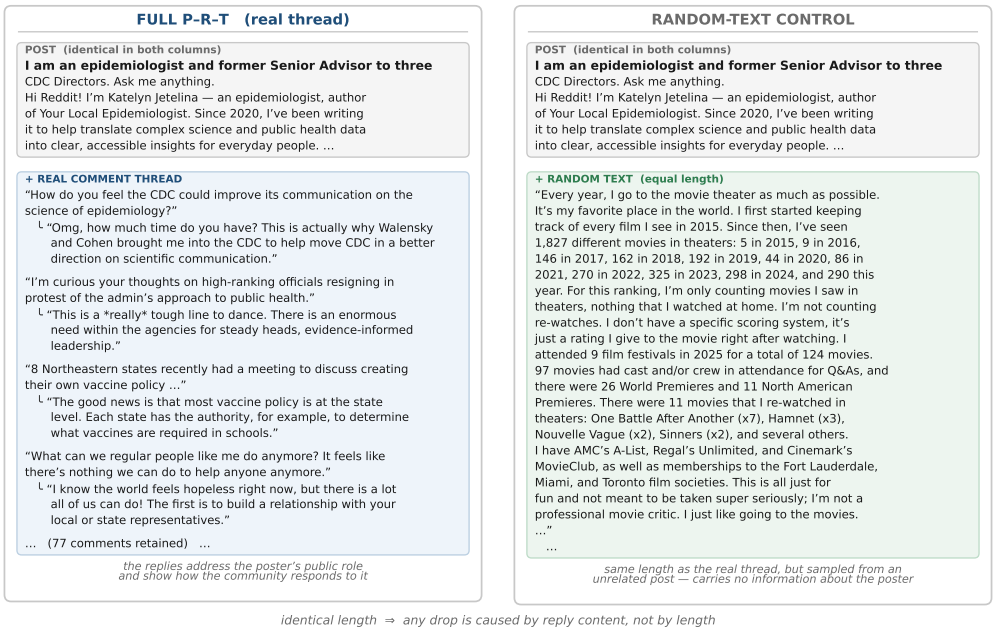


Figure 6: Length versus content. Replacing the real thread with equal-length random text removes exactly the information that identifies the poster's situation; the accuracy drop under Full P-R-T is therefore driven by reply *content*, not by sequence length.

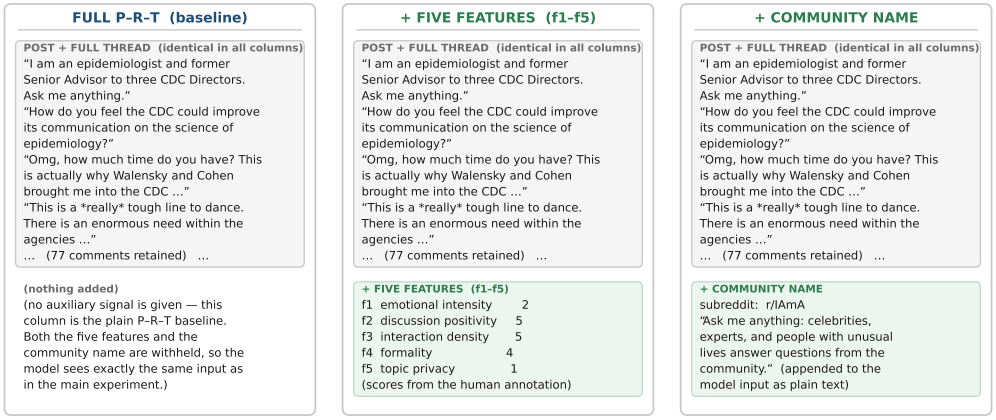


Figure 7: Auxiliary signals on the same r/IAmA post. The P–R–T input is identical in all three columns; only the extra signal differs. The feature values shown are the post's annotated values. Neither signal recovers the gold label here: with the community name supplied the model still predicts *Information Seeking*. In aggregate both signals add little—$+0.2$ points for the five features on the full benchmark, and $+4.0$ points for the community name on the 500-post subset (Table 2).

Two observations follow. First, the gain from in-context exemplars is proportional to how far a model was lagging: $+12.1$ points for DeepSeek-V4-Flash (zero-shot 40.6%), $+6.8$ for MiniMax-M2.7 (44.8%), and only $+3.6$ for Qwen3.6-Max (52.0%), so all three converge to a narrow 51.6–55.7% band. The exemplars therefore mostly supply task conventions and label boundaries that weaker models lack, rather than additional evidence about the posts. Second, even the best prompting-regime result (56.7%, chain-of-thought) remains 19.5 points below the fine-tuned RoBERTa-base (76.2%) and 21.6 points below the human annotators (78.3%) on the same subset. The low social-function accuracy is thus not explained by the prompting regime: neither in-context exemplars nor explicit reasoning recovers more than a fraction of the gap that supervised fine-tuning closes.

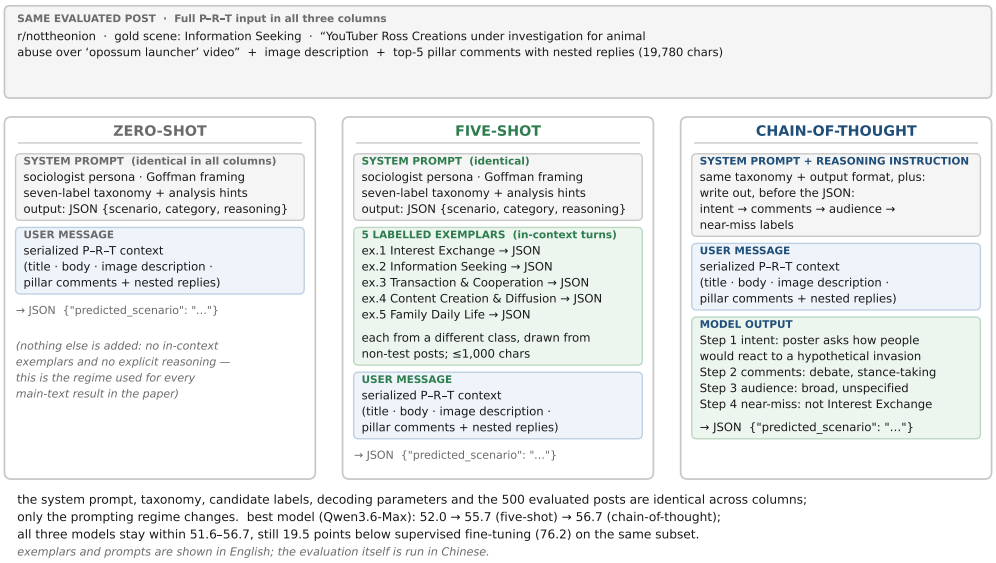


Figure 8: The three prompting regimes on the same evaluated post (Full P–R–T input). The system prompt, taxonomy, candidate labels, decoding parameters and the evaluated posts are identical across columns; only the prompting regime changes. Exemplars and prompts are shown in English; the evaluation itself is run in Chinese.

	Five-shot			Chain-of-thought		
Model	Acc.	Δ	p	Acc.	Δ	p
Qwen3.6-Max	**55.7**	+3.6	.025	**56.7**	+4.7	.001
MiniMax-M2.7	51.6	+6.8	.001	–	–	–
DeepSeek-V4-Flash	52.7	+12.1	$< .001$	51.1	+10.5	$< .001$

Table 12: Prompting-regime controls on the stratified 500-post subset (Full P–R–T input; seven-class accuracy, %). Accuracies are computed on the posts that are valid for both the variant and the zero-shot run of the same model (paired); Δ is the change with respect to the zero-shot accuracy, computed from unrounded accuracies; p is an exact McNemar test on the same pairs. Zero-shot baselines are 52.0 / 44.8 / 40.6 for the three models. Exemplars come from labelled posts outside the evaluation subset; no response reached the token limit, and the few posts blocked by the provider's content filter count as invalid predictions.

C Task2 Analysis

C.1 Simulation V1 vs. V2: Detailed Comparison

To provide a more detailed comparison between the rule-driven simulation (V1) and LLM-driven simulation (V2), Table 13 reports the complete metric breakdown averaged over the 400 posts and their three style rewrites.

Version	Style	Avg. Likes	Avg. Comments	Total
V1 (Rule)	Conservative	47.1	17.1	64.2
	Moderate	47.3	17.3	64.5
	Enthusiastic	46.6	18.0	64.7
V2 (LLM)	Conservative	63.3	36.4	99.7
	Moderate	**76.6**	44.0	**120.6**
	Enthusiastic	67.7	**50.4**	118.1

Table 13: Detailed comparison between the rule-driven (V1) and LLM-driven (V2) simulation. Averages are computed over the 400 posts and their three style rewrites (1,200 simulated posts per version); bold marks the best style in each column.

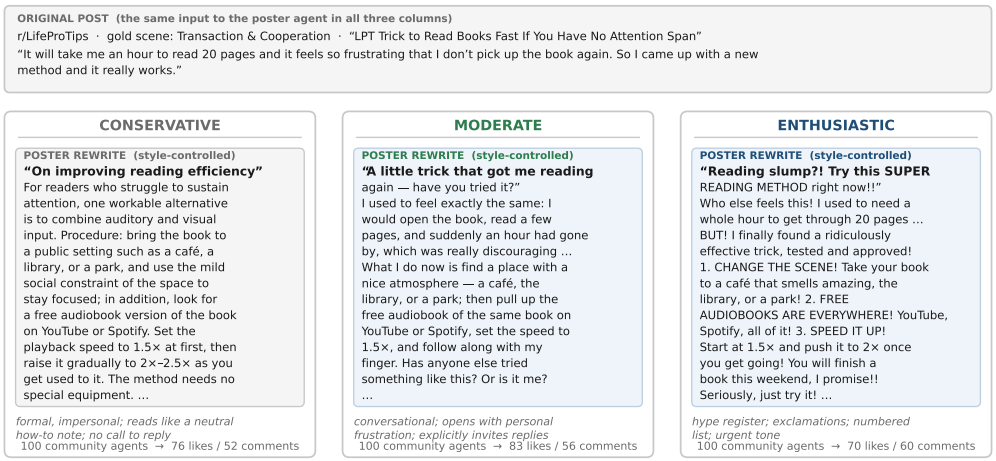


Figure 9: Input construction of the V2 behavioral probe on a single post (r/LifeProTips, gold scene: *Transaction & Cooperation*). The poster agent rewrites the same original post under three controlled communication styles, and 100 Big-Five-driven agents then read each rewrite and decide independently whether to like, comment, or ignore. The Moderate rewrite earns the most engagement (83 likes / 56 comments vs. 76/52 and 70/60), consistent with the aggregate results in Table 4(a). Rewrites are English translations of the simulation output, which is run in Chinese; ellipses abbreviate long spans.

In V1, different communication styles result in nearly identical engagement levels, with average likes ranging from 46.6 to 47.3 and comments from 17.1 to 18.0. This indicates that the rule-based simulator produces limited behavioral variation across styles. In contrast, V2 separates the styles sharply: the Moderate style attains the highest likes (76.6) and the highest total engagement (120.6), whereas the Enthusiastic style elicits the most comments (50.4) at a substantially lower like count (67.7).

The difference between V1 and V2 mainly arises from how agent behaviors are generated. In V1, actions are determined by predefined probabilistic rules, and therefore style effects are only weakly reflected in agent decisions. In V2, agents generate actions conditioned on social-function perception, personality traits, memory, and scene information, allowing different styles to produce differentiated interaction patterns.

We further analyze the action distribution in V2. The Moderate style produces a higher proportion of like actions, whereas the Enthusiastic style leads to a larger proportion of comment actions. Conservative posts generate relatively fewer active interactions and more ignored actions. These results suggest that different communication styles influence not only overall engagement volume but also the type of interaction generated.

C.2 End-to-End Ablation: Full Condition Set

The main text compares four ways of supplying scene information to the community agents. Table 14 reports the complete comparison, adding a fifth *Shuffled* condition in which each post is paired with a scene label drawn from the benchmark's label distribution but unrelated to the post itself. This control isolates how much of the gain reflects the presence of a label versus its correctness.

Three effects separate cleanly. First, *context presence*: appending unrelated length-matched text (Random) already raises likes by 5.9 over No-Scene, so part of the benefit comes from supplying any explicit social context at all. Second, a *label-distribution prior*: the Shuffled condition reaches 80.0 (+10.2), clearly above Random (+4.3, $p < 0.001$), even though none of its labels match the post—the marginal distribution of social functions is itself informative for community agents.

Condition	Avg. Likes	Avg. Comments	Δ Likes
Predicted	**82.4**	**50.1**	$+12.7$
Oracle	82.3	49.5	$+12.6$
Shuffled labels	80.0	47.5	$+10.2$
Random text	75.7	44.5	$+5.9$
No scene	69.8	43.9	—

Table 14: End-to-end ablation of scene-information injection. All conditions are compared on the same 400 posts and their three rewrites. Δ is relative to the No-Scene baseline; bold marks the best condition in each column.

Third, *semantic match*: Predicted (82.4) still exceeds Shuffled (80.0) by 2.5 points ($p < 0.001$) and is statistically indistinguishable from Oracle (82.3, $p = 0.67$).

The semantic component is thus the smallest of the three effects, yet it is precisely the component that Task 1 supplies: the Shuffled control confirms that this residual gain cannot be attributed to the label prior alone, and the Predicted–Oracle equivalence shows that imperfect social-function predictions recover essentially all of the downstream benefit.