Theory-Guided Psychological Prompting for Health Rumor Reposting Prediction with Large Language Models

Anonymous ACL submission

Abstract

Predicting whether a user will repost the health rumor on social media is important for understanding rumor diffusion and supporting early intervention. Existing approaches primarily rely on post content or shallow user attributes, while recent LLM-based methods often produce fluent behavioral predictions but without explicitly grounding them in validated psychological theory. In this study, we propose a theory-guided psychological prompting framework **PsyRumorAgent** for health rumor reposting prediction. The approach integrate psychological scales and behavioral hypotheses derived from social-psychological theories into the structured prompts, inserting scale completion as an explicit intermediate step between user context understanding and final behavioral prediction. Experimental results on the real dataset show an improvement of AUC by above 14% over the baseline COT prompting, validating the effectiveness of introducing theory-grounded psychological measurement into the LLM reasoning chain. However, theory-driven hypotheses would not improve the performance consistently, which provides crucial insights into how to guide social theory in prompt engineering. These findings suggest that LLMs can be transformed from generic predictors into theory-constrained cognitive simulators for online behavior prediction under uncertainty.

1 Introduction

Health rumor on social media has become a prominent challenge for public health governance and online platform regulation. During public health emergencies, users are frequently exposed to large volumes of uncertain, emotionally charged, and rapidly spreading messages, many of which are reposted before verification. Predicting whether a user will repost the rumor is therefore an critical issue as it can support early rumor intervention and reveal the micro-level mechanisms of rumor diffusion. Compared with rumor detection, rumor reposting prediction is intrinsically more difficult, since reposting behavior is influenced not only by message veracity, but by users' cognitive states, emotional responses, social environments, prior experiences and some uncertain factors.

Existing studies on social media reposting prediction approached this problem from multiple directions, where early methods mainly relied on diffusion modeling (Zhang et al., 2015), feature-driven prediction (Cheng et al., 2014), or deep neural architectures (Hu et al., 2018) to infer the overall number of reposts or whether a user would repost it. Recent studies further emphasize temporal dynamics and structural dependencies via graph neural networks (Sun et al., 2022), while (Aravamudan et al., 2023) directly predicts whether users will participate in future cascades under arbitrary observation windows. More recently, large language models (LLMs) have enabled cognitively richer rumor simulation frameworks, using LLM Agents to model online social media users as autonomous agents that make reasoning-based reposting decisions based on rules or contexts(Qiao et al., 2025; Zhang et al., 2024; Hu et al., 2025).

However, existing studies still suffer from several important limitations. First, most rumor diffusion prediction methods remain *behavior-descriptive* rather than *psychology-explanatory*: they infer reposting outcomes directly from observable Information such as message content, propagation traces, and user profile, but rarely model the latent psychological states that mediate between exposure and behavioral response. Second, user representations are often overly simplified, focusing on isolated demographics or shallow behavioral metadata while under-utilizing richer combinations of structured account information, interaction history, and long-form textual behavior. Third, although recent LLM-based agents and prompting frameworks have demonstrated strong reasoning and simulation

capabilities, their decision processes are typically guided by heuristic personas or generic behavior records rather than validated psychological theories and psychometric measurements. As a result, their outputs may appear fluent but remain weakly grounded in interpretable cognitive mechanisms.

Meanwhile, social science studies on health rumor diffusion monachism have shown that rumor sharing behavior is strongly associated with psychological and social factors such as anxiety, perceived credibility, information overload, trust in institutions, and social reinforcement (Wang et al., 2019; Pennycook et al., 2020; Liang et al., 2025). They also suggest that individuals often share rumor not because they deliberately endorse falsehoods, but because emotionally salient and socially engaging content reduces deliberative reasoning during online interactions (Vosoughi et al., 2018; Pennycook and Rand, 2019). Despite the gap between the empirical findings from subjective survey-based data and the real behavior mechanism, they provided us inspirations that rumor reposting prediction should move beyond shallow input-output mapping and instead incorporate psychologically grounded mechanisms that explain how users transform information exposure into behavioral decisions.

Thus, we propose **PsyRumorAgent** (**Psy**chology-driven **Rumor** Reposting **Agent**), a theory-guided LLM framework for rumor reposting prediction that explicitly integrates psychological measurement into the decision-making pipeline. Rather than treating LLMs as purely statistical classifiers or unrestricted role-playing agents, our framework operationalizes them as psychologically grounded decision makers whose outputs are linked to measurable latent cognitive states. We believe this perspective advances rumor reposting behavior prediction from direct label inference toward theory-constrained cognitive simulation, and may generalize to broader tasks involving interpretable modeling of human decision-making. Furthermore, this study builds a link between real data-driven computational prediction model and traditional survey-driven social psychological behavioral mechanism.

2 Related Works

2.1 Prompt Engineering for Behavioral Simulation

Prompt engineering has evolved from task instruction to simulating human cognition. Early work demonstrated that LLMs could perform various tasks via natural language prompts and few-shot demonstrations without task-specific fine-tuning (Brown et al., 2020), and that zero-shot reasoning could be elicited simply by appending "Let's think step by step" (Kojima et al., 2022). These findings motivated a wave of structured reasoning techniques like Chain-of-Thought (CoT) (Wei et al., 2022), Tree-of-Thoughts (ToT) (Yao et al., 2023), and Graph-of-Thoughts (GoT) (Besta et al., 2024). While these methods significantly improve task performance, they treat the model as a rational problem solver and largely ignore the psychological variability that would shape human decision-making in social contexts.

When applying LLMs into human behavior simulation, (Park et al., 2023) first introduced generative agents which can store experiences, reflect on memories, and plan actions, producing believable daily routines. However, simply assigning demographic personas is insufficient for faithful social simulation, and aligning LLM agents with human belief networks, rather than raw demographics, yields substantially better role-playing fidelity (Chuang et al., 2024). (Wang et al., 2025) went a step further by proposing the PEBA framework, explicitly using a psychological equation to align agent behaviors in crowd simulations.

2.2 Psychological Determinants of Health Rumor Sharing

The propensity to share unverified health information is a multifaceted behavioral outcome driven by an interplay of individual psychological traits, cognitive states, and linguistic markers. Conventional studies usually proposed a behavioral model based on related social psychological theories, such as MOA (Motivation Opportunity Ability) (Liang et al., 2025), ELM (Elaboration Likelihood Model) (Wang et al., 2023), TPB (Theory of Planned Behavior) (Zhao et al., 2016), etc., and use the subjective survey data to validate the influencing mechanism of rumor diffusion. Recent research also has focused on user-centric behavioral profiling, highlighting that social media users' post histories are an underused yet valuable resource for uncovering hidden sharing motivations (Schoenmueller et al., 2025). They have identified specific markers, such as the frequent use of words related to *anger*, *power*, and *existential concerns* (*e.g.*, religion and mortality), that significantly distinguish misinformation sharers from general users.

A central cognitive variable in this domain is *news discernment*, the ability to correctly distinguish between accurate and fabricated content (Sultan et al., 2024). The belief in and sharing of fake news is often linked to a lack of analytical reasoning and a higher reliance on intuitive heuristics (Pennycook and Rand, 2021; Sultan et al., 2024). Furthermore, motivations for sharing extend beyond cognitive errors to include social factors such as status-seeking, self-expression, and even a misguided sense of altruism (Melchior and Oliveira, 2024). While existing literature provides a rich descriptive understanding of these psychological determinants through empirical analysis (Sultan et al., 2024; Melchior and Oliveira, 2024), they often fail to capture the dynamic, generative nature of individual decision-making. This gap motivates the development of **PsyRumorAgent**, which transitions from analyzing "who shares" to a simulation framework that synthesizes these multifaceted variables into fine-grained individual predictions.

2.3 Integrating Psychometric Scales into LLM Agents

A recent line of work has begun to bring standardized psychometric instruments into the evaluation of LLMs. It was demonstrated that well-established psychological questionnaires for anxiety, depression, and sense of coherence can be reformulated as natural language inference prompts, and that the resulting latent constructs in pre-trained language models align with human psychological theory (Reuben et al., 2025) . This finding confirms that LLMs can internalize and exhibit psychometrically interpretable states, opening the door for their use as psychology-informed agents. However, existing efforts remain at the level of *assessment*: they examine whether an LLM can reproduce a stable psychological profile in isolation. They ignored operationalizing these scales as an active component of an LLM agent's decision-making chain, that is, instructing LLMs to complete a validated psychological items and then using condition a subsequent behavioral prediction. This is the unique contribution of **PsyRumorAgent**.

3 Methodology

3.1 Problem Formulation

We investigate the rumor reposting prediction problem in social media context. Given an online rumor instance r and a user u on social media, the traditional rumor reposting prediction task can be formally defined as:

$$f(r, u) \rightarrow z \tag{1}$$

where r denotes the rumor content; u denotes the user information consisting of profile, social activity and historical records, etc; z denotes the behavioral prediction result, including the reposting decision and corresponding reposting probability. This function characterizes the direct mapping from external information and user features to behavioral decisions.

However, this modeling approach neglects the crucial psychological mechanisms in the user decision-making process. We argue that whether a user reposts a rumor is not directly determined by content and attributes, but is mediated by latent psychological variables grounded in social-psychological theory. To make the framework theory-driven yet extensible, we firstly model the derivation from theory to measurement and hypotheses as:

$$(Q, h) = \Phi(\mathcal{T}) \tag{2}$$

where $\mathcal{T}$ denotes an integrated psychological theory framework, Q denotes the resulting scale system, and h denotes the corresponding theoretical hypotheses. Thus, the overall prediction process by integrating psychological theories can be unified as:

$$z = f\big(r,\ u,\ Q,\ h\big) \tag{3}$$

where $f(\cdot)$ denotes the overall prediction function that directly maps rumor information, user attributes, theory-derived scales with theoretical hypotheses to the behavioral prediction result z. This expression explicitly characterizes the unified modeling relationship between psychological theory and LLM-based behavioral prediction.

3.2 Three-Layer Modeling Framework

Based on the above problem formulation and the key idea, we propose **PsyRumorAgent**, a three-layer psychologically-driven rumor reposting prediction framework. It constructs an LLM-driven cognitive decision system by integrating physical information, user attributes, and psychological measurement results. Figure 1 presents the overall architecture of **PsyRumorAgent**, including the information layer, user layer, and psychological layer, together with the end-to-end prediction flow from rumor exposure to reposting decision. Specifically, it consists of three layers:

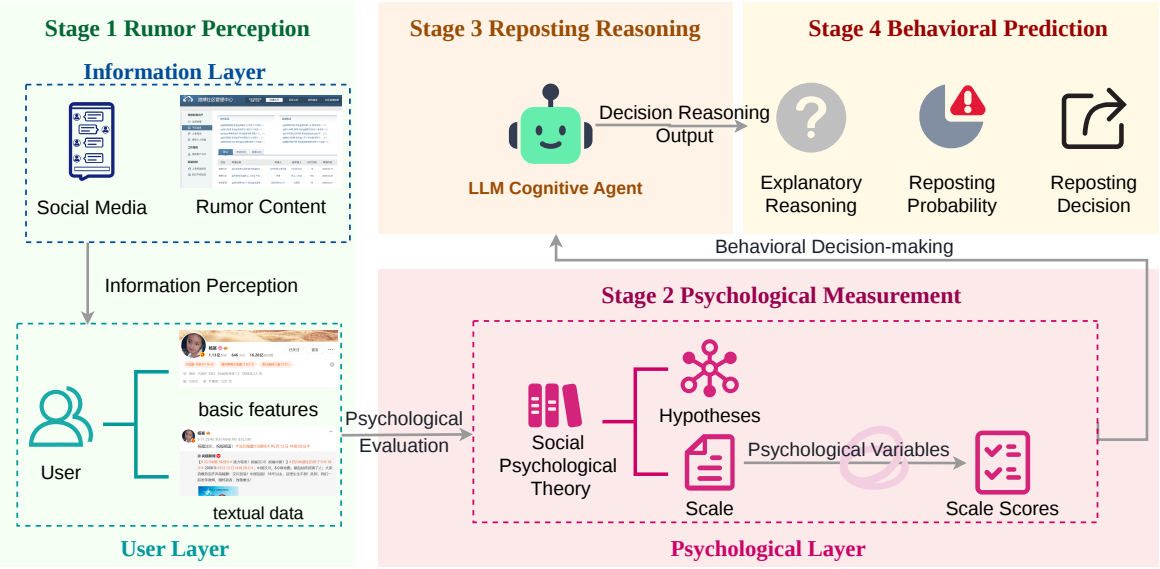

Figure 1: Overview of the proposed health rumor reposting prediction framework: **PsyRumorAgent**.

- **Information Layer**: Models the rumor content and its propagation context, serving as the external stimulus.

- **User Layer**: Characterizes the structural attributes and behavioral features of users to provide individual differences. Specifically, it includes three types of user information: the *structural profile data*, which contains structured information such as account authentication status, membership level, and account type; the *numerical activity data*, which contains numerical information such as the number of followers, the number of followings, and the number of posts; and the *textual data*, which contains user biographies and historical post/repost records. Among them, the first two together form the basic features u^{base}, and formally, the user representation is written as $u = \left(u^{\text{base}},\ u^{\text{text}}\right)$, where u^{text} summarizes the textual data.

- **Psychological Layer**: The Psychological Layer is the core component of this model as it provides a principled bridge between observable user inputs and latent decision mechanisms. Rather than prompting the agent only with surface-level information, we ground psychological reasoning with the help of the integrated theoretical framework which can derive the scale system Q and the theoretical hypotheses h. In this way, theory serves as the source of structured psychological constructs, scale items, and expected behavioral relationships, allowing the prompt to encode not merely raw information but also interpretable constraints on how the agent should simulate user cognition. Injecting theory-derived scales and hypotheses into the prompt helps the LLM produce psychologically coherent responses, reduces arbitrary reasoning drift, and makes the simulated mental state consistent with established behavioral explanations.

In this three-layer structure, the Information Layer provides stimulus input, the User Layer provides individual background, and the Psychological Layer plays a crucial mediating role, enabling the model to simulate the complete cognitive path of "*Information Perception–Psychological Evaluation–Behavioral Decision Making*".

3.3 LLM-based Cognitive Agent

We formally propose the psychologically-driven LLM agent that simulates the user's cognitive and behavioral processes to make final decision through four stages:

- **Rumor Perception Stage (Information Domain):** We first construct an external stimulus representation by integrating rumor text, contextual information, and basic user features, providing the agent with initial contextual input to form the "stimulus memory" that triggers subsequent cognitive processes.

- **Psychological Measurement Stage (Cognitive Domain):** We input structured psycholog-

ical scales constructed based on the selected social theory (such as MOA-SOR, TPB, CLT, and SOR-ELM, detailed in the experiment section) into the LLM, requiring the model to simulate the user's responses to the scale items, thereby generating psychological measurement results m to characterize the user's latent psychological state.

- **Reposting Reasoning Stage (Cognitive Decision Domain):** Under the joint condition of rumor information and psychological measurement results m, the LLM performs structured reasoning on whether the user will repost the rumor and produces an intermediate behavioral prediction result, thereby simulating the causal cognitive path consistent with the logic of Structural Equation Modeling.
- **Behavioral Prediction Stage (Decision Output Domain):** The model finally outputs a unified prediction signal including the binary reposting decision, the reposting probability, and optional explanatory reasoning.

$\langle 1 \rangle$ **Rumor Perception Stage** This stage integrates rumor information and basic user features to form an initial stimulus representation:

$$\mathrm{LLM}_U = \mathrm{LLM}_{\mathrm{RP}}(r,\ u,\ p_u) \tag{4}$$

where $\mathrm{LLM}_{\mathrm{RP}}$ denotes the context construction agent based on role-playing, p_u represents the structured user prompt information, and the output LLM_U represents the perception state in the user's context. This stage provides a contextual memory foundation for subsequent psychological measurement.

$\langle 2 \rangle$ **Psychological Measurement Stage** Given the complete scale system Q derived from the integrated theory framework $\mathcal{T}$, we construct the scale items into structured prompts and input them into the LLM for item-by-item measurement:

$$m = \mathrm{LLM}_M\big(Q \mid r,\ u,\ p_m\big) \tag{5}$$

where LLM_M denotes the psychological measurement agent, p_m represents the structured scale prompt, and the output m is the simulated scale answering result, reflecting the user's response state in the theory-grounded psychological space.

$\langle 3 \rangle$ **Reposting Reasoning Stage** After obtaining the psychological variable m, with the guidance of theoretical hypotheses h, the model enters the decision reasoning stage:

$$z = \mathrm{LLM}_D\big(r,\ u,\ m,\ h,\ p_{\mathrm{decision}}\big) \tag{6}$$

where LLM_D denotes the decision agent, and p_{decision} represents the structured decision prompt used to constrain the output format and reasoning method. The output z denotes the behavioral prediction result and contains the reposting decision, reposting probability, and necessary explanation.

This process corresponds to the following psychological chain:

$$r \longrightarrow m \longrightarrow z \tag{7}$$

This path characterizes how rumor stimuli influence the unified behavioral prediction result through the mediating role of psychological variables, thereby achieving an approximate modeling of the causal paths.

$\langle 4 \rangle$ **Behavioral Prediction Stage** In the final stage, the model outputs a unified behavioral prediction result z:

$$z = (y,\ p,\ e). \tag{8}$$

Here, y denotes the binary reposting decision, p denotes the reposting probability, and e denotes the necessary natural-language explanation. In this way, the prediction unifies classification output, probabilistic confidence, and explanatory reasoning within a single theory-guided decision interface.

4 Experiments

The replicable data and code are shared via https://anonymous.4open.science/r/PsyRumorAgent-357C.

4.1 Experimental Settings

4.1.1 Dataset

We construct the experimental dataset from Sina Weibo. Positive samples are sampled from the MisBot dataset, a large-scale newly published dataset from Sina Weibo that contains the diffusion path of misinformation and annotations for social bots (Wan et al., 2025). We first filter health-related rumors from it and then treat users' reposting behaviors of these rumors as positive samples. Then we collected a set of non-robot users from commenting users under non-rumor public event-related posts also on Sina Weibo, and randomly assigned a rumor post in positive samples to each of these users,

Theory	Measurement Dimensions
MOA-SOR	health consciousness; time cost; health literacy; anxiety;
TPB-NAM	awareness of adverse consequences; ascribed responsibility; attitudes; subjective norms; perceived behavioral control; personal norms; behavioral intention to combat rumors; actual rumor combating behavior
A-CLT	deficient self-regulation; entertainment; exploration; social media fatigue; religiosity; self-promotion; unverified news/information sharing
SOR-ELM	personal involvement; source credibility; health information anxiety; perceived trust in health rumors; willingness to share health rumors; electronic health literacy

Table 1: Theoretical sources and measurement dimensions used to construct the psychological scale system.

forming negative samples. To avoid introducing inactive or bot accounts, we retain only users with observable activity records and remove accounts that show clear bot-like or inactive behavior. In this way, users in negative samples represents active non-bot users who participate in public information diffusion rather than health-rumor propagation. In total, we keep a dataset of 3,000 user-post pairs with the same number of positive and negative ones.

4.1.2 Social-psychological Theories

Our psychological scale system is grounded by selecting four popular theoretical frameworks, namely MOA-SOR (Liang et al., 2025), TPB-NAM (Zhao et al., 2016), A-CLT (Islam et al., 2020), and SOR-ELM (Wang et al., 2023), described in Table 1, where the full names can be found in Appendix A. These theories serve as the conceptual foundation for constructing measurement items and behavioral hypotheses. MOA-SOR captures reposting as emerging from motivational readiness, ability, and opportunity; TPB-NAM explains intention via rational choice and pro-social norm activation; A-CLT addresses how information overload depletes cognitive resources; SOR-ELM characterizes how users rely on analytical or peripheral cues in reposting decisions.

4.1.3 Compared Prompting Strategies

To systematically evaluate how the proposed **PsyRumorAgent** can improve the performance of the health-rumor reposting prediction over the general baselines, we construct six progressive prompting settings around three core dimensions: Chain-of-Thought (CoT) reasoning, psychological scale injection, and theory-based hypothesis guidance. All settings share the same data input, task objective, and model configuration; the only controlled difference lies in the prompting structure and reasoning workflow, provided in Appendix B.

- **GP** (General Prompt) directly outputs the behavioral prediction without additional cognitive guidance given only the rumor content r and user representation u.
- **CoT** introduces a Chain-of-Thought reasoning step before outputting prediction, using the same input as GP.
- **Scale** injects structured psychological scale items Q, requiring the model to first assess the user's psychological tendency before making the behavioral prediction.
- **H-Scale** (Hypothesis-Guided Scale Prompt) further incorporates theoretical hypotheses h, providing prior knowledge to guide reasoning.
- **Scale-A** (Scale with Analysis Prompt) adopts a three-stage workflow, *i.e.*, scale assessment, motivation analysis, and behavioral prediction, allowing a behavioral motivation explanation before the final prediction.
- **H-Scale-A** (Hypothesis-Scale with Analysis Prompt) constitutes psychological scales, theoretical hypotheses, and behavioral motivation analysis before final prediction.

We evaluate these prompts on four representative LLMs: GLM (glm-4.6), DS (DeepSeek-V3.2), GPT (gpt-4.1-mini), and GROK (grok-4-1-fast-non-reasoning), where the temperature is set 0.5 for all. These models cover different model families and providers, allowing to examine whether theory-guided psychological prompting is effective beyond a single LLM backend. For each model, we use the same dataset, user-post input format, prompt definitions, and output schema.

4.1.4 Evaluation Metrics

We evaluate reposting prediction as a binary classification task. We report two groups of classification metrics. $\mathbf{Acc_{no}}$ and $\mathbf{F1_{no}}$ are computed from the model's original output binary decision y. $\mathbf{Acc_{med}}$ and $\mathbf{F1_{med}}$ are computed after applying a median-based threshold to the predicted reposting probabilities, matching the balanced proportion of positive and negative samples in the dataset. We also report AUC (Area Under the ROC Curve) to evaluate the ranking discriminatity of output reposting probabilities independent of a fixed threshold.

Finally, RepoRate denotes the proportion of samples predicted as reposting in the LLM outputs, which helps reveal whether a prompt strategy tends to produce overly conservative or overly aggressive reposting predictions.

4.2 Main Results

Table 2 presents the main comparison across vaiours prompt strategies on four LLMs. Several key observations are shown. First, scale-based strategies (Scale, H-Scale, Scale-A, H-Scale-A) generally improve over the baseline GP and CoT across most models, though with occasional exceptions (e.g., the $\mathbf{Acc_{no}}$ and $\mathbf{F1_{no}}$ of GPT under Scale/H-Scale-A, Grok under H-Scale/H-Scale-A). Second, among scale-based strategies, the best-performing setting varies by model and metric: GLM achieves its highest Acc and F1 under H-Scale, while DS and GPT perform the best under Scale-A on certain metrics. Third, the performance gap between models is notable, i.e., GLM and DS generally achieve higher scores than GPT and Grok, indicating that larger-scale models with stronger instruction-following capabilities are better suited for theory-grounded prediction tasks. It is noting that the analysis-based COT can improve GP much, but the analysis in the scale strategies do not significantly enhance the counterpart without analysis (Scale-A v.s. Scale, H-Scale-A v.s. H-Scale-A), and sometimes it would degrade. Also, the hypotheses in the prompt can not improve the performance consistently when comparing H-Scale with Scale, H-Scale-A with Scale-A, except for H-scale using GLM. It indicates that the scale is more critical form the theoretical-enhanced reasoning. However, explicit hypotheses on the absolute relationships between constructs can interfere with the LLMs' inference based on implicitly subtle knowledge between constructs automatically learned from large-scale diverse pre-training data.

Table 2: Performance of different models and prompting strategies. The results based on theoretical scale prompts (Scale, H-Scale, Scale-A, H-Scale-A) are the ensemble results of the reposting decision voting and average probability from the four theoretical outcomes.

Model	Prompt	Acc_{no}	$F1_{no}$	Acc_{med}	$F1_{med}$	AUC	RepoRate
GLM	GP	0.605	0.574	0.544	0.510	0.586	0.231
	CoT	0.563	0.551	0.595	0.630	0.667	**0.656**
	Scale	0.652	0.652	0.652	0.656	0.703	0.468
	H-Scale	**0.665**	**0.665**	**0.668**	**0.669**	**0.722**	0.502
	Scale-A	0.631	0.622	0.653	0.662	0.714	0.655
	H-Scale-A	0.638	0.634	0.645	0.649	0.709	0.588
DS	GP	0.615	0.612	0.602	0.571	0.619	0.411
	CoT	0.625	0.618	0.591	0.587	0.654	0.366
	Scale	0.637	0.637	0.643	0.642	0.690	0.498
	H-Scale	0.620	0.620	0.619	0.619	0.671	0.505
	Scale-A	0.641	**0.639**	**0.649**	**0.650**	**0.700**	**0.578**
	H-Scale-A	**0.643**	0.635	0.642	0.640	0.689	0.351
GPT	GP	0.575	0.515	0.510	0.510	0.532	0.149
	CoT	0.579	0.525	0.574	0.537	0.580	**0.162**
	Scale	0.562	0.477	**0.592**	**0.587**	**0.619**	0.097
	H-Scale	0.586	0.528	0.589	0.568	0.612	0.151
	Scale-A	**0.589**	**0.531**	0.571	0.553	0.584	0.149
	H-Scale-A	0.571	0.512	0.567	0.561	0.579	0.154
GROK	GP	0.555	0.521	0.505	0.502	0.532	0.234
	CoT	**0.575**	**0.564**	0.536	0.532	0.577	0.337
	Scale	0.566	0.557	**0.582**	**0.588**	**0.609**	**0.646**
	H-Scale	0.548	0.548	0.530	0.533	0.573	0.508
	Scale-A	0.552	0.547	0.560	0.566	0.593	0.601
	H-Scale-A	0.562	0.554	0.554	0.550	0.574	0.367

Table 3 further breaks down the results by theoretical framework, reporting the 4-model average for each prompt-theory combination, where the full results are shown in Appendix C. On average, A-CLT achieves the highest scores in most settings, while TPB-NAM surpasses it on several metrics under H-Scale-A. TPB-NAM is generally the second best one, where its structured behavioral intention constructs align well with the motivation analysis stage. However, SOR-ELM tends to yield lower scores, possibly due to the more abstract nature of its information-processing constructs. These findings highlight that the choice of different psychological theories would make differences, and theory-grounded prompting can substantially improve LLM-based behavioral prediction.

Table 3: Average performance of different prompt strategies across theories (4-model average)

Prompt	Theory	Acc_{no}	$F1_{no}$	Acc_{med}	$F1_{med}$	AUC	RepoRate
GP	—	0.587	0.556	0.540	0.523	0.568	0.256
CoT	—	0.586	0.565	0.574	0.571	0.620	0.380
Scale	MOA-SOR	0.595	0.571	0.602	0.581	0.632	0.422
	TPB-NAM	0.592	0.557	0.618	0.609	0.652	**0.476**
	A-CLT	**0.625**	**0.610**	**0.632**	**0.628**	**0.658**	0.445
	SOR-ELM	0.580	0.556	0.580	0.537	0.604	0.351
H-Scale	MOA-SOR	0.583	0.575	0.572	0.566	0.597	0.418
	TPB-NAM	0.597	0.571	0.612	**0.625**	0.642	**0.485**
	A-CLT	**0.622**	**0.614**	**0.622**	0.605	**0.646**	0.466
	SOR-ELM	0.592	0.568	0.578	0.557	0.605	0.332
Scale-A	MOA-SOR	0.569	0.548	0.588	0.581	0.609	0.527
	TPB-NAM	0.595	0.573	0.623	0.626	0.654	**0.554**
	A-CLT	**0.625**	**0.614**	**0.631**	**0.629**	**0.670**	0.487
	SOR-ELM	0.582	0.560	0.580	0.548	0.604	0.398
H-Scale-A	MOA-SOR	0.556	0.544	0.540	0.530	0.560	0.379
	TPB-NAM	0.602	0.580	**0.619**	**0.612**	**0.652**	**0.439**
	A-CLT	**0.622**	**0.612**	0.615	0.600	0.642	0.381
	SOR-ELM	0.586	0.556	0.582	0.544	0.601	0.320

4.3 Extensive Experiments

Beyond the main comparison of prompt strategies, We also conduct extensive experiments to further examine the robustness, interpretability, efficiency, and psychological validity of **PsyRumorAgent**.

Table 4: Performance gains from psychology-guided prompting by user information richness

Model	Sparse Users		Rich Users	
	ΔF1	ΔAUC	ΔF1	ΔAUC
DS	+0.101	+0.072	+0.003	+0.018
GLM	+0.140	+0.102	+0.104	+0.063
GPT	-0.010	+0.052	-0.032	+0.028
GROK	+0.164	+0.029	+0.077	+0.032
Avg.	**+0.099**	**+0.063**	+0.038	+0.035

Table 5: Aggregated topic distribution by prompt setting

Topic Group	CoT	Scale	Hypo	All
CoT-oriented	0.369	0.296	0.244	0.282
Scale-oriented	0.387	0.438	0.285	0.365
Hypothesis-oriented	0.244	0.266	0.471	0.354

4.3.1 User Heterogeneity Analysis

First, we perform a user heterogeneity analysis by dividing users into information-rich and information-sparse groups according to the amount of available profile information, with additional details provided in Appendix D.1. The results in Table 4 show that, for both groups, the average performance of the four prompts with social-psychological guidance (Scale, H-Scale, Scale-A, and H-Scale-A) is higher than that the two prompts without such guidance (GP and CoT) in terms of AUC and F1. More importantly, the average improvement is larger for the information-sparse user group. This suggests that theory-guided psychological prompting is especially helpful when user information is limited, as explicit psychological scale answering can assist the LLM in simulating latent user states and behavior motivations from sparse observations. In contrast, for information-rich users, the LLM can already perform relatively effective simulation based on the available profile information and behavioral records, leaving less room for additional improvement from theory-based guidance.

4.3.2 Topic Analysis of the Analysis Text

Second, we conduct a topic analysis on the explanations generated by the LLM under different prompt strategies, with details shown in Appendix D.2. The topics of explanation text vary across different prompting settings. Scale-oriented and Hypo-oriented topics accounted for a larger proportion in Scale-based prompts (Scale and Scale-A) and Hypothesis-based version (H-Scale and H-Scale-A) respectively, validating that the designed prompts could influence the LLM reasoning as excepted.

4.3.3 The Robustness of Scale filtering

Third, we perform a scale filtering experiment to evaluate prompting efficiency. For each psychological dimension of the scales, we respectively select the most relevant one-third and two-thirds scale items according to the examples of user information using prompting on GLM-4.6, and then run the theory-based prediction under these selected scales. Compared to the performance on the original full scale systems, the differences are relatively small, where the prompt templates and the results are shown in Appendix D.3, indicating that profile-matched scale filtering with shorter prompt and thus less token cost will maintain competitive prediction performance compared to the full scales.

4.3.4 Comparing LLM-generated Scale Data to Real Survey Data

Fourth, we feed the LLM-generated scale scores into the same statistical models as the corresponding literature. The resulting paths are largely consistent with those obtained from human survey responses in the original studies, although the path hit rates vary across different theories. Details of this results are provided in Appendix D.4.

5 Conclusion

This study shows that LLMs not only serve as prediction models, but as scalable proxy tools for theory-guided psychological measurement. By estimating latent psychological states from users' social media records with social psychological theory-based guidance, the proposed framework provides an alternative way to study health rumor reposting behavior when direct questionnaire collection from social media users is costly or infeasible in some contexts. Also,it builds a link between real data-driven computational prediction and traditional survey-driven social psychological behavioral mechanism. Our experiments lead to several main findings. First, mapping social-psychological theories into prompts can improve prediction performance, especially for information-sparse users, suggesting that theory-grounded psychological prompts help LLMs better infer users' reposting tendencies. Second, further mapping theories into explicit hypotheses does not necessarily bring additional accuracy gains, although it may result in more theory-oriented explanations. Third, all scale items are not necessary for the prediction, but the key factor is whether the selected scales cover the important psychological dimensions.

Limitations

This study has several limitations. First, the experiments are conducted on a health-rumor reposting dataset constructed from Sina Weibo, and the negative samples are collected from users who participate in non-rumor public-event discussions. Although this design controls for inactive and bot-like accounts, the setting may not fully capture all forms of non-reposting behavior, and the findings still need to be validated on other platforms, languages, and rumor domains. Second, the psychological scale scores are simulated by LLMs from users' social media records rather than collected from human questionnaire responses. Therefore, these scores should be interpreted as model-estimated psychological constructs, not as direct psychometric measurements. Third, our framework focuses on a selected set of social-psychological theories and prompt strategies. Other theories, scale dimensions, model families, and prompt designs may lead to different results, especially for the effect of hypothesis-guided prompting; future work should therefore explore more systematic ways to integrate theories into LLM reasoning, such as theory-aware adaptation or structured theory representations. Finally, this work mainly evaluates prediction accuracy and explanation topics, while the causal validity of the inferred psychological pathways remains to be further examined through human studies or longitudinal behavioral data.

AI Assistant Statement

AI assistance was used in this work for code writing and for checking the grammar and language of the manuscript. The authors reviewed, edited, and take full responsibility for all generated content, analyses, and final writing.

Ethical Considerations

Data Privacy. The user profile data used in this study was collected from publicly accessible Weibo posts. During persona construction, we anonymized user names and user IDs to ensure that no identity information could be exposed. Given the public nature of the data, individual informed consent was not obtained; however, we minimized privacy risks by retaining only anonymized textual content and interaction features for the experiment. **API Compliance.** This study utilized the API services provided by OpenAI, DeepSeek, Zhipu AI, and xAI in full compliance with their respective terms of service and usage policies. All API calls were made within the permitted rate limits and exclusively for academic research purposes.

Potential Misuse and Bias. The primary goal of this research is to explore how psychological theories and structured prompting strategies can enhance LLM-based behavioral simulation. However, we acknowledge several ethical risks. First, the framework could be misappropriated to manipulate user behavior or design targeted misinformation campaigns; we explicitly prohibit such applications and release our methodology solely for academic purposes. Second, although our scale-based approach incorporates validated psychological instruments, the underlying LLMs may inherit biases from their training data, potentially producing stereotypical or skewed predictions—for example, over-attributing health anxiety to older demographics or underrepresented groups. Furthermore, we emphasize that the scale scores generated by LLMs during persona role-playing are *simulated* outputs and should not be interpreted as actual psychological measurements or clinical diagnoses of real individuals. They serve as structured reasoning scaffolds for behavioral prediction, not as judgments of real users' psychological states. Accordingly, this framework should be viewed as a research tool for exploring behavioral mechanisms rather than a definitive solution for real-world deployment. Researchers and practitioners are expected to adhere to ethical guidelines and exercise caution when interpreting model outputs.

References

Akshay Aravamudan, Xi Zhang, and Georgios C. Anagnostopoulos. 2023. Anytime user engagement prediction in information cascades for arbitrary observation periods. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4):4999–5009.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and 1 others. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Justin Cheng, Lada Adamic, P Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. 2014. Can cascades be predicted? In *Proceedings of the 23rd international conference on World wide web*, pages 925–936.

Yun-Shiuan Chuang, Krirk Nirunwiroj, Zach Studdiford, Agam Goyal, Vincent V Frigo, Sijia Yang, Dhavan V Shah, Junjie Hu, and Timothy T Rogers. 2024. Beyond demographics: Aligning role-playing LLM-based agents using human belief networks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14010–14026.

Tianrui Hu, Dimitrios Liakopoulos, Xiwen Wei, Radu Marculescu, and Neeraja J. Yadwadkar. 2025. Simulating rumor spreading in social networks using LLM agents. *Preprint*, arXiv:2502.01450.

Wenjian Hu, Krishna Kumar Singh, Fanyi Xiao, Jinyoung Han, Chen-Nee Chuah, and Yong Jae Lee. 2018. Who will share my image? predicting the content diffusion path in online social networks. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 252–260.

A.K.M. Najmul Islam, Samuli Laato, Shamim Talukder, and Erkki Sutinen. 2020. Misinformation sharing and social media fatigue during covid-19: An affordance and cognitive load perspective. *Technological Forecasting and Social Change*, 159:120201.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213.

Xuedong Liang, Yanda Huo, and Peng Luo. 2025. The role of health consciousness in online health rumour sharing: Insights from the integrated MOA–SOR framework. *Journal of Information Science*, pages 1–16.

Cristiane Melchior and Mírian Oliveira. 2024. A systematic literature review of the motivations to share fake news on social media platforms and how to fight them. *New Media & Society*, 26(2):1127–1150.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G. Lu, and David G. Rand. 2020. Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, 31(7):770–780. Epub 2020 Jun 30.

Gordon Pennycook and David G. Rand. 2019. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188:39–50. The Cognitive Science of Political Thought.

Gordon Pennycook and David G Rand. 2021. The psychology of fake news. *Trends in cognitive sciences*, 25(5):388–402.

Boyu Qiao, Kun Li, Wei Zhou, and Songlin Hu. 2025. Dynamic simulation framework for disinformation dissemination and correction with social bots. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4688–4710, Suzhou, China. Association for Computational Linguistics.

Maor Reuben, Ortal Slobodin, Idan-Chaim Cohen, Aviad Elyashar, Orna Braun-Lewensohn, Odeya Cohen, and Rami Puzis. 2025. Assessment and manipulation of latent constructs in pre-trained language models using psychometric scales. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2433–2444.

Verena Schoenmueller, Simon J Blanchard, and Gita Venkataramani Johar. 2025. Who shares fake news? uncovering insights from social media users' post histories. *Journal of Marketing Research*, 62(2):316–341.

Mubashir Sultan, Alan N Tump, Nina Ehmann, Philipp Lorenz-Spreen, Ralph Hertwig, Anton Gollwitzer, and Ralf HJM Kurvers. 2024. Susceptibility to online misinformation: A systematic meta-analysis of demographic and psychological factors. *Proceedings of the National Academy of Sciences*, 121(47):e2409329121.

Ling Sun, Yuan Rao, Xiangbo Zhang, Yuqian Lan, and Shuanghe Yu. 2022. Ms-hgat: memory-enhanced sequential hypergraph attention network for information diffusion prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 4156–4164.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Herun Wan, Minnan Luo, Zihan Ma, Guang Dai, and Xiang Zhao. 2025. How do social bots participate in misinformation spread? a comprehensive dataset and analysis. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31493–31516.

Ping Wang, Qing Wen, Yanbo Chu, and Jingxin Song. 2023. Factors affecting users' intention to share online health rumor in the context of infodemic: Based on the integrated model of S-O-R and ELM. *Library and Information Service*, 67(22):114–127. (In Chinese).

Yunzhe Wang, Gale Lucas, Burcin Becerik-Gerber, and Volkan Ustun. 2025. 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.

Yuxi Wang, Martin McKee, Aleksandra Torbica, and David Stuckler. 2019. Systematic literature review on the spread of health-related misinformation on social media. *Social Science & Medicine*, 240:112552.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36:11809–11822.

Qi Zhang, Yeyun Gong, Ya Guo, and Xuanjing Huang. 2015. Retweet behavior prediction using hierarchical dirichlet process. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 403–409.

Yifan Zhang and 1 others. 2024. From a tiny slip to a giant leap: An LLM-based simulation for fake news evolution. *arXiv preprint arXiv:2410.19064*.

Liming Zhao, Jianli Yin, and Yao Song. 2016. An exploration of rumor combating behavior on social media in the context of social crises. *Computers in Human Behavior*, 58:25–36.

A Theoretical Scales

This appendix presents the four theoretical frameworks employed in our study. Each framework includes a brief description, its research hypotheses, and the corresponding measurement scales used for data collection. Note that all scale items presented below are translated from the original Chinese versions, as Chinese scales were used throughout the experiments.

A.1 MOA-SOR

Description: The MOA-SOR framework integrates the MOA (Motivation-Opportunity-Ability) theory with the SOR (Stimulus-Organism-Response) model to explain health rumor sharing behavior. This framework posits that individuals' health consciousness (HC) serves as a key motivational factor that influences anxiety (AN), which in turn drives online health rumor sharing behavior.

Research Hypotheses:

- H1: Anxiety (AN) positively predicts individuals' online health rumor sharing behavior.
- H2: Health consciousness (HC) positively predicts online health rumor sharing behavior through increased anxiety (AN).

Scale Items are shown in Table 6, where we using 5-point Likert scale: 1 = Strongly Disagree, 5=Strongly Agree.

A.2 TPB-NAM

Description: The TPB (Theory of Planned Behavior) combined with the NAM (Norm Activation Model) provides a comprehensive framework for understanding information verification behavior during crises. This theory suggests that individuals' intention to verify information before sharing is influenced by their attitude toward verification, subjective norms, and perceived behavioral control.

Research Hypotheses:

- H1: Attitude toward verification (AT) positively influences intention during crises.
- H2: Subjective norm (SN) positively influences verification intention.
- H3: Perceived behavioral control (PC) positively influences intention.
- H4: PC positively influences actual verification behavior.
- H5: Behavioral intention (BI) positively influences actual behavior.

Scale Items are shown in Table 7, where we using 5-point Likert scale: 1=Strongly Disagree, 5=Strongly Agree.

A.3 A-CLT

Description: The Affordance-Construal Level Theory (A-CLT) framework examines how social media affordances and psychological factors influence unverified information sharing during health crises. This theory explores the dual role of social media features and individual differences in shaping information sharing behaviors.

Research Hypotheses:

- H1: Self-promotion (SP) reduces unverified information sharing.
- H2: Entertainment (EN) increases unverified information sharing.
- H3: Exploration (EX) reduces unverified information sharing.
- H4: Religiosity (CR) increases unverified information sharing.
- H5: Self-regulation deficit (SRD) increases unverified sharing.
- H6: Social media fatigue (SMF) increases unverified sharing.

Scale Items are shown in Table 8, where we using 5-point Likert scale: 1=Strongly Disagree,

5=Strongly Agree.

A.4 SOR-ELM

Description: The SOR-ELM framework combines the SOR (Stimulus-Organism-Response) model with the ELM (Elaboration Likelihood Model) to explain health information sharing willingness during infodemics. This theory proposes that information characteristics serve as stimuli that trigger cognitive and emotional responses, ultimately influencing sharing behavior.

Research Hypotheses:

- H1: Health information anxiety (HIA) positively affects unverified health information sharing willingness.
- H2: Health information perceived trust (HRB) positively affects unverified health information sharing willingness.
- H3: Health information anxiety (HIA) mediates the relationship between information involvement (PI) and unverified health information sharing willingness.
- H4: Health information perceived trust (HRB) mediates the relationship between information involvement (PI) and unverified health information sharing willingness.
- H5: Health information anxiety (HIA) mediates the relationship between source credibility (SC) and unverified health information sharing willingness.
- H6: Health information perceived trust (HRB) mediates the relationship between source credibility (SC) and unverified health information sharing willingness.
- H7: Health information anxiety (HIA) and health information perceived trust (HRB) have a chain mediating effect between information involvement (PI) and unverified health information sharing willingness.
- H8: Health information anxiety (HIA) and health information perceived trust (HRB) have a chain mediating effect between source credibility (SC) and unverified health information sharing willingness.

Scale Items are shown in Table 9, where we using 5-point Likert scale: 1=Strongly Disagree, 5=Strongly Agree.

Table 6: MOA-SOR Scale Items

Construct	Item
HC	1. I frequently reflect on my own health status.
	2. I am relatively sensitive to my own health condition.
	3. I proactively pay attention to bodily change signals.
TC	1. Searching for health info on social media requires time.
	2. I do not have much time to search and read health info.
	3. I consider whether time investment is worthwhile.
HL	1. I know which sources are reliable for disease info.
	2. I know how to obtain materials for health actions.
	3. I can distinguish information source of a message.
	4. I check authority of health information sources.
	5. I verify authenticity of health information.
AN	1. I feel nervous about health status of myself/others.
	2. I feel anxious about health status of myself/others.
	3. I feel uneasy about health status of myself/others.

Table 7: TPB-NAM Scale Items

Construct	Item
AC	1. Insufficiently verified info is common during crises.
	2. Unverified info may undermine public trust.
	3. Inaccurate info may affect rescue operations.
	4. Inaccurate info may exacerbate crisis situations.
AR	1. I have responsibility to alleviate crisis situations.
	2. I exercise caution to prevent unverified info spread.
AT	1. Verifying info before publishing is good for me.
	2. Verifying info before posting is pleasant.
	3. Verifying info before publishing is wise.
	4. Verifying info before publishing is beneficial.
SN	1. My friends expect me to verify info on social media.
	2. My family expects me to verify info on social media.
	3. Important people expect me to verify info.
PC	1. I am capable of verifying info on social media.
	2. I can investigate the source of information.
	3. I have time and resources to verify info.
PN	1. I feel moral obligation to value info accuracy.
	2. I personally feel obligated to value info accuracy.
	3. I take responsibility for authenticity of info.
BI	1. I will verify authenticity when facing uncertain info.
	2. I am willing to refute false info.
	3. I will make efforts to refute false info.
AB	1. I prioritize sharing info from authoritative sources.
	2. I prefer sharing info with clear source links.
	3. I confirm facts before sharing info.

Table 8: A-CLT Scale Items

Construct	Item
CR	1. Do you participate in communal prayers?
	2. Do you engage in personal religious activities?
	3. How important is your religious belief?
	4. Does faith help you during difficult times?
	5. How would you rate your level of faith?
EX	1. I actively seek as much info as possible.
	2. I look for new opportunities to grow.
	3. I look for new things or experiences.
SRD	1. I find it difficult to control social media usage.
	2. I need constant stimulation from social media.
	3. I tried reducing social media time without success.
	4. I hide time spent on social media.
	5. My social media usage is out of control.
EN	1. When news is interesting, I share on social media.
	2. When news is compelling, I share on social media.
	3. When I see exciting news, I share it.
SP	1. Sharing helps me impress others.
	2. Sharing helps me express myself.
	3. Sharing makes me feel important.
	4. Sharing conveys my ideal image.
SMF	1. After using social media, I find it hard to relax.
	2. After using social media, I feel exhausted.
	3. After using social media, I find it hard to concentrate.
	4. During use, I feel too tired for other tasks.
UIS	1. I share COVID-19 info without checking authenticity.
	2. I share info without verifying facts.
	3. I share info without verifying authenticity.
	4. I share info even when I feel it may be incorrect.

Table 9: SOR-ELM Scale Items

Construct	Item
PI	1. This information is very important to me.
	2. This information is highly relevant to me.
	3. This info is important to my family and friends.
	4. This info is relevant to my family and friends.
SC	1. Health info publishers on social media are trustworthy.
	2. Publishers are knowledgeable.
	3. Publishers are professional.
	4. Publisher accounts are attractive.
HIA	1. I feel overwhelmed by amount of health info.
	2. I fail to notice needed health info emergence.
	3. Hard-to-distinguish info makes me uneasy.
	4. I frequently exit browsing health info.
	5. Vast health info makes me feel distracted.

Construct	Item
HRB	1. When first exposed to this message, how much do you trust it? "A 26-year-old anchor suddenly had a heart attack on stage, with no warning signs at all. Life is but a moment. So, don't dwell on the past, don't hope for the future, live in the present. Enjoy life, health first, grateful to be together." 2. When first exposed to this message, how much do you trust it? "Society is sick! Yu Qiuyu was hospitalized with liver disease, feeling: I am sick, and society is also sick." 3. When first exposed to this message, how much do you trust it? "An 80-cent disease, but the hospital wants the patient's family to pay 100,000 yuan for surgery... Hospitals are really black-hearted!" 4. When first exposed to this message, how much do you trust it? "A patient's family held a doctor hostage with a knife at Wuhan First People's Hospital. A female SWAT officer disguised as a nurse delivered water and suddenly shot and killed the family member." 5. When first exposed to this message, how much do you trust it? "Suffocating, the AIDS man in Moments: I am proud! I have successfully infected AIDS to 40 female college students...!"
SI	1. I consider this info worth sharing. 2. I am willing to share via social media. 3. I plan to share via social media.
EHL	1. I know how to find useful health info online. 2. I know how to use internet to solve health problems. 3. I know what health resources are available online. 4. I know where to find useful health info online. 5. I know how to use health info to help myself. 6. I can evaluate quality of online health info. 7. I can distinguish high-quality from low-quality info. 8. I am confident in health decisions using online info.

B Prompt Templates

This study employs six prompt strategies to investigate how theory-guided psychological prompting affect LLM performance in predicting users' rumor reposting behavior. The strategies range from simple general prompts to complex frameworks that incorporate theoretical scales, research hypotheses, and chain-of-thought reasoning. Note that all prompt templates presented below are translated from the original Chinese versions, as Chinese prompts were used throughout the experiments.

Overview of Prompt Strategies

GP	Baseline prompt without scales or theory.
CoT	Extends the baseline with step-by-step reasoning.
Scale	Integrates theoretical scale questions into the prompt.
H-Scale	Extends Scale with research hypotheses.
Scale-A	Places analysis before prediction.
H-Scale-A	Combines hypotheses and analysis steps.

B.1 GP: General Prompt

GP Prompt

You now need to fully immerse yourself in and assume the identity of the following user, simulating their cognition and behavior in a real social media environment.
[Given Event] The following is a piece of health-related information spreading on social media: {rumor_text}
[User Profile] The following information is derived from this user's long-term behavioral patterns: {user_profile}
[Task Description] Core Prediction Task: Given this specific general health description, combined with the user's personal information, analyze the user's individual state, behavioral habits, attitudes toward unverified information, and thereby determine whether the user has a forwarding motivation and an actual forwarding tendency.
[Output Requirements] Please output the result only in strict JSON format: { "uid": "{uid}", "rumor_forwarding_probability": "0.0-1.0", "will_forward": true or false, "analysis": " 50 words explaining the Prediction" }

B.2 CoT: Chain-of-Thought Prompt

CoT Prompt

You now need to fully immerse yourself in and assume the identity of the following user, simulating their cognition and behavior in a real social media environment.
[Given Event] The following is a piece of health-related information: {rumor_text}
[User Profile] The following information is derived from this user's long-term behavioral patterns: {user_profile}
[Task Description] Analysis Task: Comprehensively consider the user's individual traits and the current specific health information context, to deduce whether they would disseminate the information. Core Prediction Task: Given this specific general health description, first analyze and reason the user's individual state, and thereby determine whether the user has a forwarding motivation and an actual forwarding tendency.
[Output Requirements] { "uid": "{uid}", "analysis": " 50 words analytical summary", "rumor_forwarding_probability": "0.0-1.0", "will_forward": true or false }

B.3 Scale: Scale Prompt

Scale Prompt

[a] You now need to fully immerse yourself in and assume the identity of the following user, simulating their cognition and behavior in a real social media environment.
[Given Event] The following is a piece of health-related information: {rumor_text}
[User Profile] The following information is derived from this user's long-term behavioral patterns: {user_profile}
[Task Description] 1. Scale Response Task: Assume the identity of the user and answer each question in the survey questionnaire. 2. Core Prediction Task: In the specific context of this health description, combined with the scale response results, determine whether the user has a forwarding motivation and an actual forwarding tendency.3. Prediction Explanation: After making the Prediction, generate a 50-word explanation of the reasoning process and basis.

{survey_questions}
[Output Requirements] { "uid": "{uid}", "survey_responses": {"Q1": score, ...}, "rumor_forwarding_probability": "0.0-1.0", "will_forward": true or false, "analysis": " 50 words explaining the Prediction" }

[a]It was noted that though we prompt LLMs to out put analysis in **Scale** and **H-scale**, this analysis is only for investigating how LLMs make predictions and it can not influence the decision output that appears at the front. It is different from "analysis-then-decision" mode (**Scale-A** and **H-Scale-A**), where the analysis would determine how the LLMs make perdition.

B.4 H-Scale: Scale with Hypotheses Prompt

H-Scale Prompt

You now need to fully immerse yourself in and assume the identity of the following user, simulating their cognition and behavior in a real social media environment.
[Given Event] The following is a piece of health-related information: {rumor_text}
[User Profile] The following information is derived from this user's long-term behavioral patterns: {user_profile}
[Task Description] 1. Scale Response Task: Assume the identity of the user and answer each question in the survey questionnaire. 2. Core Prediction Task: Leverage the research hypotheses regarding the scale, analyze the user's individual state, and thereby determine whether the user has a forwarding motivation.3. Prediction Explanation: After making the Prediction, generate a 50-word explanation of the reasoning process and basis.
{research_hypotheses}
{survey_questions}
[Output Requirements] { "uid": "{uid}", "survey_responses": {"Q1": score, ...}, "rumor_forwarding_probability": "0.0-1.0", "will_forward": true or false, "analysis": " 50 words explaining the Prediction" }

B.5 Scale-A: Scale with Analysis Prompt

Scale-A Prompt

You now need to fully immerse yourself in and assume the identity of the following user, simulating their cognition and behavior in a real social media environment.
[Given Event] The following is a piece of health-related information: {rumor_text}
[User Profile] The following information is derived from this user's long-term behavioral patterns: {user_profile}
[Task Description] 1. Scale Response Task: Assume the identity of the user and answer each question in the survey questionnaire. 2. Analysis Task: Combining the user's individual traits and the scale response results, comprehensively consider the current specific health information context and extract an analytical summary of 50 words. 3. Core Prediction Task: Ultimately deduce and determine whether the user has a forwarding motivation and an actual forwarding tendency.
{survey_questions}
[Output Requirements] { "uid": "{uid}", "survey_responses": {"Q1": score, ...}, "analysis": " 50 words analytical summary", "rumor_forwarding_probability": "0.0-1.0", "will_forward": true or false }

B.6 H-Scale-A: Hypothesis-Scale wtih Analysis Prompt

H-Scale-A Prompt

You now need to fully immerse yourself in and assume the identity of the following user, simulating their cognition and behavior in a real social media environment.
[Given Event] The following is a piece of health-related information: {rumor_text}
[User Profile] The following information is derived from this user's long-term behavioral patterns: {user_profile}
[Task Description] 1. Scale Response Task: Assume the identity of the user and answer each question in the survey questionnaire. 2. Behavioral Reasoning and Analysis Task: Explicitly leveraging the research hypotheses regarding the scale, analyze the user's individual state and generate 50 words explaining the logical reasoning process. 3. Core Prediction Task: Strictly based on the logical reasoning results from Step 2, ultimately infer the user's forwarding probability and their final behavior.
{research_hypotheses}
{survey_questions}

[Output Requirements] { "uid": "{uid}", "survey_responses": {"Q1": score, ...}, "analysis": " 50 words reasoning process", "rumor_forwarding_probability": "0.0-1.0", "will_forward": true or false }

C Complete Experimental Results

This appendix presents the complete experimental results for all model–prompt–theory combinations, where Table 10 shows the results using raw output categories by LLMs and Table 11 shows the results using median threshold-based output categories.

D Extensive Experimental Results

This appendix reports the detailed results of the four extended experiments: user heterogeneity analysis, topic analysis, scale filtering analysis, and SEM consistency analysis. These results complement the main prompt-strategy comparison by examining the robustness, interpretability, and efficiency of the proposed framework.

D.1 User Heterogeneity Analysis Results

To examine whether theory-guided prompting benefits different users to the same extent, we divide users into two groups according to the amount of textual information available in their profiles. Specifically, we calculate the text length of each user's available information and use the median text length as the cutoff: users below the median are assigned to the *sparse* group, while users at or above the median are assigned to the *rich* group. Table 12 reports the prediction results for each group after this median-based split. For each group, we first compute the average AUC and F1 scores of prompt settings without explicit psychological theory guidance, including GP and CoT. We then compute the corresponding average scores of prompt settings with social-psychological theory guidance, including Scale, H-Scale, Scale-A, and H-Scale-A. Finally, we report the difference between the two averages in the main text, where positive values indicate that theory-guided prompting improves prediction performance over non-theory-guided prompting.

The detailed prompt-level results are reported in Table 12. As summarized in Table 4 in the main text, both sparse and rich user groups obtain positive gains after introducing social-psychological theory, indicating that theory-guided prompting consistently improves LLM-based reposting prediction. More importantly, the improvement is larger for sparse users than for rich users. This pattern suggests that psychological theories are especially useful when user information is limited: explicit psychological constructs and behavioral hypotheses provide additional guidance for the LLM to infer latent user states and simulate possible reposting motivations from sparse observations. In contrast, for users with richer profiles, the model can already make relatively strong predictions based on the available behavioral evidence, so the marginal benefit of additional theory-guided prompting becomes smaller.

Table 10: Results Using Raw Output Categories by LLMs

Model	Prompt:Theory	Acc	F1	TN	FP	FN	TP	AUC	RepoRate
GLM	GP	0.6048	0.5743	1299	186	991	502	0.5865	0.2310
	CoT	0.5626	0.5515	600	876	420	1067	0.6675	0.6558
	Scale: MOA-SOR	0.6453	0.6428	1082	399	654	834	0.6527	0.4153
	Scale: TPB-NAM	0.6630	0.6625	930	554	448	1041	0.7042	0.5365
	Scale: A-CLT	0.6761	0.6761	1015	469	496	999	0.7059	0.4928
	Scale: SOR-ELM	0.5843	0.5834	939	546	692	801	0.6249	0.4523
	H-Scale: MOA-SOR	0.6207	0.6201	979	502	625	865	0.6504	0.4601
	H-Scale: TPB-NAM	0.6796	0.6796	1014	467	484	1003	0.7083	0.4953
	H-Scale: A-CLT	0.6872	0.6871	999	479	450	1042	0.7281	0.5121
	H-Scale: SOR-ELM	0.6180	0.6171	847	636	499	989	0.6792	0.5470
	Scale-A: MOA-SOR	0.5792	0.5635	579	904	347	1143	0.6655	0.6885
	Scale-A: TPB-NAM	0.6094	0.5915	594	884	275	1214	0.7198	0.7071
	Scale-A: A-CLT	0.6670	0.6637	844	636	354	1139	0.7396	0.5970
	Scale-A: SOR-ELM	0.5991	0.5971	787	696	496	994	0.6552	0.5684
	H-Scale-A: MOA-SOR	0.5853	0.5850	832	651	584	911	0.6170	0.5245
	H-Scale-A: TPB-NAM	0.6266	0.6098	621	854	252	1235	0.7113	0.7053
	H-Scale-A: A-CLT	0.6696	0.6693	951	532	452	1043	0.7143	0.5289
	H-Scale-A: SOR-ELM	0.6110	0.6103	849	632	525	968	0.6643	0.5380
DS	GP	0.6152	0.6121	1056	444	710	789	0.6195	0.4111
	CoT	0.6252	0.6183	1139	361	763	736	0.6536	0.3658
	Scale: MOA-SOR	0.6209	0.6208	913	587	550	949	0.6594	0.5122
	Scale: TPB-NAM	0.6121	0.6099	806	693	470	1029	0.6590	0.5744
	Scale: A-CLT	0.6499	0.6492	1040	460	590	909	0.6612	0.4565
	Scale: SOR-ELM	0.6582	0.6579	1036	464	561	938	0.6708	0.4675
	H-Scale: MOA-SOR	0.5829	0.5816	791	709	542	957	0.6360	0.5555
	H-Scale: TPB-NAM	0.5925	0.5899	769	731	491	1008	0.6344	0.5799
	H-Scale: A-CLT	0.6275	0.6273	903	597	520	979	0.6399	0.5255
	H-Scale: SOR-ELM	0.6533	0.6423	1243	257	783	717	0.6784	0.3247
	Scale-A: MOA-SOR	0.5990	0.5873	646	854	349	1151	0.6637	0.6683
	Scale-A: TPB-NAM	0.6117	0.6058	735	765	400	1100	0.6723	0.6217
	Scale-A: A-CLT	0.6547	0.6542	925	575	461	1039	0.7085	0.5380
	Scale-A: SOR-ELM	0.6383	0.6379	1010	490	595	905	0.6756	0.4650
	H-Scale-A: MOA-SOR	0.5613	0.5571	989	511	805	695	0.5967	0.4020

Model	Prompt:Theory	Acc	F1	TN	FP	FN	TP	AUC	RepoRate
	H-Scale-A: TPB-NAM	0.6600	0.6549	1173	327	693	807	0.6965	0.3780
	H-Scale-A: A-CLT	0.6347	0.6283	1148	339	757	756	0.6683	0.3650
	H-Scale-A: SOR-ELM	0.6430	0.6335	1206	294	777	723	0.6735	0.3390
GPT	GP	0.5747	0.5148	1389	111	1165	335	0.5318	0.1487
	CoT	0.5790	0.5246	1376	124	1139	361	0.5804	0.1617
	Scale: MOA-SOR	0.5517	0.4708	1414	86	1259	241	0.5842	0.1090
	Scale: TPB-NAM	0.5270	0.4153	1446	45	1374	135	0.6001	0.0600
	Scale: A-CLT	0.6097	0.5612	1413	87	1084	416	0.6319	0.1677
	Scale: SOR-ELM	0.5590	0.4672	1461	39	1284	216	0.5912	0.0850
	H-Scale: MOA-SOR	0.5967	0.5726	1251	249	961	539	0.5868	0.2627
	H-Scale: TPB-NAM	0.5520	0.4785	1391	109	1235	265	0.5825	0.1247
	H-Scale: A-CLT	0.6040	0.5751	1297	203	985	515	0.6061	0.2393
	H-Scale: SOR-ELM	0.5753	0.5031	1435	65	1209	291	0.5854	0.1187
	Scale-A: MOA-SOR	0.5747	0.5210	1364	136	1140	360	0.5468	0.1653
	Scale-A: TPB-NAM	0.6017	0.5548	1389	111	1084	416	0.5864	0.1757
	Scale-A: A-CLT	0.6020	0.5671	1329	171	1023	477	0.6090	0.2160
	Scale-A: SOR-ELM	0.5537	0.4678	1433	67	1272	228	0.5349	0.0983
	H-Scale-A: MOA-SOR	0.5553	0.5306	1177	323	1011	489	0.5135	0.2707
	H-Scale-A: TPB-NAM	0.5647	0.4982	1393	107	1199	301	0.6018	0.1360
	H-Scale-A: A-CLT	0.5943	0.5658	1276	224	993	507	0.5810	0.2437
	H-Scale-A: SOR-ELM	0.5457	0.4519	1439	61	1302	198	0.5398	0.0863
GROK	GP	0.5547	0.5208	1231	269	1067	433	0.5324	0.2340
	CoT	0.5752	0.5636	1107	393	881	618	0.5770	0.3371
	Scale: MOA-SOR	0.5607	0.5506	616	884	434	1066	0.6320	0.6500
	Scale: TPB-NAM	0.5650	0.5403	500	1000	305	1195	0.6451	0.7317
	Scale: A-CLT	0.5663	0.5542	602	898	403	1097	0.6327	0.6650
	Scale: SOR-ELM	0.5198	0.5151	928	571	869	631	0.5302	0.4008
	H-Scale: MOA-SOR	0.5312	0.5257	957	542	864	636	0.5139	0.3928
	H-Scale: TPB-NAM	0.5639	0.5368	483	1016	292	1208	0.6423	0.7416
	H-Scale: A-CLT	0.5697	0.5664	724	776	515	985	0.6105	0.5870
	H-Scale: SOR-ELM	0.5213	0.5086	1023	477	959	541	0.4762	0.3393
	Scale-A: MOA-SOR	0.5222	0.5188	656	841	591	909	0.5581	0.5839
	Scale-A: TPB-NAM	0.5593	0.5384	513	969	336	1143	0.6372	0.7133
	Scale-A: A-CLT	0.5769	0.5729	720	779	489	1009	0.6233	0.5966

Model	Prompt:Theory	Acc	F1	TN	FP	FN	TP	AUC	RepoRate
	Scale-A: SOR-ELM	0.5362	0.5354	864	636	755	744	0.5519	0.4602
	H-Scale-A: MOA-SOR	0.5203	0.5043	1050	448	990	510	0.5113	0.3195
	H-Scale-A: TPB-NAM	0.5576	0.5571	782	716	608	887	0.5974	0.5356
	H-Scale-A: A-CLT	0.5911	0.5857	1057	440	785	714	0.6042	0.3852
	H-Scale-A: SOR-ELM	0.5431	0.5275	1085	413	955	541	0.5254	0.3186

Table 11: Results using on Median Threshold-based Output Categories

Model	Prompt:Theory	Acc	F1	TN	FP	FN	TP	AUC
GLM	GP	0.5440	0.5101	913	572	786	707	0.5865
	CoT	0.5953	0.6303	742	734	465	1022	0.6675
	Scale: MOA-SOR	0.6255	0.6073	997	484	628	860	0.6527
	Scale: TPB-NAM	0.6630	0.6751	930	554	448	1041	0.7042
	Scale: A-CLT	0.6757	0.6745	1012	472	494	1001	0.7059
	Scale: SOR-ELM	0.5745	0.5639	892	593	674	819	0.6249
	H-Scale: MOA-SOR	0.6210	0.6109	961	520	606	884	0.6504
	H-Scale: TPB-NAM	0.6799	0.6788	1014	467	483	1004	0.7083
	H-Scale: A-CLT	0.6872	0.6917	999	479	450	1042	0.7281
	H-Scale: SOR-ELM	0.6230	0.6371	868	615	505	983	0.6792
	Scale-A: MOA-SOR	0.6290	0.6349	911	572	531	959	0.6655
	Scale-A: TPB-NAM	0.6805	0.6920	954	524	424	1065	0.7198
	Scale-A: A-CLT	0.6667	0.6967	844	636	355	1138	0.7396
	Scale-A: SOR-ELM	0.6014	0.6161	837	646	539	951	0.6552
	H-Scale-A: MOA-SOR	0.5850	0.5955	832	651	585	910	0.6170
	H-Scale-A: TPB-NAM	0.6442	0.6950	707	768	286	1201	0.7113
	H-Scale-A: A-CLT	0.6696	0.6795	951	532	452	1043	0.7143
	H-Scale-A: SOR-ELM	0.6224	0.6300	895	586	537	956	0.6643
DS	GP	0.6019	0.5708	1011	489	705	794	0.6195
	CoT	0.5905	0.5865	900	600	628	871	0.6536
	Scale: MOA-SOR	0.6222	0.6259	918	582	551	948	0.6594
	Scale: TPB-NAM	0.6187	0.6389	844	655	488	1011	0.6590
	Scale: A-CLT	0.6355	0.6271	987	513	580	919	0.6612
	Scale: SOR-ELM	0.6492	0.6458	988	512	540	959	0.6708
	H-Scale: MOA-SOR	0.6102	0.6148	897	603	566	933	0.6360
	H-Scale: TPB-NAM	0.6152	0.6221	895	605	549	950	0.6344
	H-Scale: A-CLT	0.6272	0.6363	903	597	521	978	0.6399
	H-Scale: SOR-ELM	0.6493	0.6340	1037	463	589	911	0.6784
	Scale-A: MOA-SOR	0.6197	0.6433	830	670	471	1029	0.6637
	Scale-A: TPB-NAM	0.6273	0.6534	828	672	446	1054	0.6723
	Scale-A: A-CLT	0.6560	0.6667	936	564	468	1032	0.7085
	Scale-A: SOR-ELM	0.6353	0.6329	963	537	557	943	0.6756
	H-Scale-A: MOA-SOR	0.5640	0.5461	905	595	713	787	0.5967
	H-Scale-A: TPB-NAM	0.6690	0.6536	1070	430	563	937	0.6965
	H-Scale-A: A-CLT	0.6313	0.6047	1048	439	667	846	0.6683
	H-Scale-A: SOR-ELM	0.6470	0.6309	1036	464	595	905	0.6735
GPT	GP	0.5097	0.5095	765	735	736	764	0.5318

Model	Prompt:Theory	Acc	F1	TN	FP	FN	TP	AUC
	CoT	0.5740	0.5366	982	518	760	740	0.5804
	Scale: MOA-SOR	0.5863	0.4681	1213	287	954	546	0.5842
	Scale: TPB-NAM	0.5997	0.4698	1267	224	977	532	0.6001
	Scale: A-CLT	0.6230	0.5837	1076	424	707	793	0.6319
	Scale: SOR-ELM	0.5937	0.4411	1300	200	1019	481	0.5912
	H-Scale: MOA-SOR	0.5580	0.5529	854	646	680	820	0.5868
	H-Scale: TPB-NAM	0.5833	0.5504	985	515	735	765	0.5825
	H-Scale: A-CLT	0.6030	0.4860	1246	254	937	563	0.6061
	H-Scale: SOR-ELM	0.5847	0.5098	1106	394	852	648	0.5854
	Scale-A: MOA-SOR	0.5617	0.5006	1026	474	841	659	0.5468
	Scale-A: TPB-NAM	0.5940	0.5250	1109	391	827	673	0.5864
	Scale-A: A-CLT	0.6053	0.5552	1077	423	761	739	0.6090
	Scale-A: SOR-ELM	0.5573	0.4226	1186	314	1014	486	0.5349
	H-Scale-A: MOA-SOR	0.5167	0.4993	827	673	777	723	0.5135
	H-Scale-A: TPB-NAM	0.5963	0.5337	1096	404	807	693	0.6018
	H-Scale-A: A-CLT	0.5820	0.5430	1001	499	755	745	0.5810
	H-Scale-A: SOR-ELM	0.5617	0.4320	1185	315	1000	500	0.5398
GROK	GP	0.5053	0.5023	767	733	751	749	0.5324
	CoT	0.5358	0.5316	817	683	709	790	0.5770
	Scale: MOA-SOR	0.5730	0.6216	667	833	448	1052	0.6320
	Scale: TPB-NAM	0.5890	0.6516	614	886	347	1153	0.6451
	Scale: A-CLT	0.5930	0.6285	746	754	467	1033	0.6327
	Scale: SOR-ELM	0.5042	0.4992	771	728	759	741	0.5302
	H-Scale: MOA-SOR	0.4992	0.4853	789	710	792	708	0.5139
	H-Scale: TPB-NAM	0.5685	0.6484	512	987	307	1193	0.6423
	H-Scale: A-CLT	0.5697	0.6041	724	776	515	985	0.6105
	H-Scale: SOR-ELM	0.4533	0.4482	694	806	834	666	0.4762
	Scale-A: MOA-SOR	0.5419	0.5467	796	701	672	828	0.5581
	Scale-A: TPB-NAM	0.5897	0.6350	689	793	422	1057	0.6372
	Scale-A: A-CLT	0.5953	0.5990	878	621	592	906	0.6233
	Scale-A: SOR-ELM	0.5258	0.5219	801	699	723	776	0.5519
	H-Scale-A:MOA-SOR	0.4930	0.4809	774	724	796	704	0.5113
	H-Scale-A: TPB-NAM	0.5650	0.5651	845	653	649	846	0.5974
	H-Scale-A: A-CLT	0.5778	0.5739	879	618	647	852	0.6042
	H-Scale-A: SOR-ELM	0.4980	0.4833	788	710	793	703	0.5254

Table 12: Prompt-level prediction results after median-based grouping by user information text length

Model	Prompt	Sparse Users		Rich Users	
		F1	AUC	F1	AUC
DS	GP	0.5667	0.5971	0.5900	0.6405
	CoT	0.5495	0.6271	0.5877	0.6774
	Scale	0.6583	0.6698	0.6112	0.6952
	H-Scale	0.6601	0.6549	0.5929	0.6618
	Scale-A	0.7091	0.7057	0.6171	0.6885
	H-Scale-A	0.6078	0.7047	0.5453	0.6623
GLM	GP	0.4539	0.5664	0.4683	0.5996
	CoT	0.6574	0.6439	0.5800	0.6841
	Scale	0.6537	0.6870	0.6292	0.7068
	H-Scale	0.6984	0.7189	0.6368	0.7149
	Scale-A	0.7127	0.7032	0.6328	0.7117
	H-Scale-A	0.7172	0.7179	0.6133	0.6854
GPT	GP	0.3529	0.5358	0.3333	0.5184
	CoT	0.3771	0.5761	0.3464	0.5763
	Scale	0.2612	0.6129	0.2631	0.6180
	H-Scale	0.3864	0.6242	0.3427	0.5852
	Scale-A	0.3935	0.5968	0.3293	0.5532
	H-Scale-A	0.3781	0.5962	0.2962	0.5440
GROK	GP	0.3490	0.4778	0.4455	0.5765
	CoT	0.4707	0.5277	0.5183	0.6183
	Scale	0.6426	0.5540	0.5953	0.6547
	H-Scale	0.5682	0.5228	0.5546	0.6097
	Scale-A	0.6036	0.5341	0.5896	0.6402
	H-Scale-A	0.4829	0.5160	0.4966	0.6143

D.2 Topic Analysis Results

The topic analysis examines the explanations generated by the LLM under different prompt strategies. We analyze the explanation texts using LDA (Latent Dirichlet Allocation) to identify recurring cognitive and behavioral themes, and then compare how these themes vary across prompting settings. The experiment is implemented with the Python packages `scikit-learn` and `jieba`. Specifically, we first use `jieba` for Chinese word segmentation and then construct the bag-of-words representation with `max_df=0.6` and `min_df=20`. For LDA training, we set `n_components=8` and `random_state=42`. This experiment is used to evaluate whether different prompts guide the model toward different explanatory perspectives.

The topic distributions indicate that prompt design changes not only the output format but also the factors emphasized in the model's explanations. The hypothesis-guided settings cluster more clearly around theory-variable and hypothesis-related topics, especially Topic 5 and Topic 8, suggesting that explicit theoretical hypotheses guide the model to produce explanations that are more aligned with the intended psychological constructs. In contrast, the scale-based settings more often focus on structured analytical themes such as critical thinking, cautious verification, and authenticity judgment. The CoT and no-scale baseline settings place greater weight on broader themes, including user interests, low attention to the rumor, and general participation domains. Overall, these patterns support the conclusion that different prompt strategies systematically reshape the topic distribution of the reasoning process.

D.3 Scale Filtering Results

The scale filtering experiment evaluates whether prompt cost can be reduced by selecting a smaller subset of psychological scale items. This experiment is conducted using the GLM model with the Scale-A prompting strategy, as the preceding results show that GLM combined with Scale-A achieves relatively strong and stable overall performance. Therefore, this setting is selected as the representative configuration for evaluating whether scale-item filtering can preserve predictive effectiveness while reducing prompt length. For each psychological dimension, we respectively select the most relevant one-third, two-thirds, and full set of scale items according to the user profile and behavioral records of serval randomly selected samples. The filtering is performed using GLM-4.6 with prompts as below, and then compare prediction performance across the three scale lengths.

Scale Filtering Prompt

Please generate a user profile based on the user information, and then select the questionnaire items from the scale that are most relevant to the generated user profile. For each dimension of the scale, select a fixed proportion of items:{given proportion}, rounded up to the nearest integer, and output the results in a structured format.
User Information: {user_information}
Scale Questionnaire: {scale_questionnaire}

Table 13: Topic Distribution and Assigned Groups (Translated Version)

Topic ID	Keywords	Assigned Group	COT	Scale	Hypo	All
1	Focus, Extremely Low, Field, Concentration, Attention, Lottery, Game, Participation, Politics	cot-oriented	0.232190	0.170750	0.160461	0.173068
2	Sense of Justice, Sympathy, HIA, Generate, Extreme, Source, Empathy, HRB, Justice, PI	hypo-oriented	0.129676	0.155581	0.164309	0.156542
3	Criticism, System, Doubt, Critical, Values, Thinking, Media, Patriotic, Fit, Official	scale-oriented	0.157125	0.163011	0.113801	0.140628
4	Cautious, Emergency, Proactive, Health Preservation, Sensitivity, Media, Thrilling, Activity, Energy, Influence	scale-oriented	0.143124	0.153244	0.090826	0.124558
5	Regulation, Subjective, PC, Control, Action, Sense of Responsibility, Government, Account, AT, BI	hypo-oriented	0.066224	0.068945	0.165915	0.111451
6	Warning, Relatives and Friends, Livelihood, Energy, Safety, Psychology, Doubtful, Protection, Blessing, Tragedy	cot-oriented	0.136967	0.125644	0.083877	0.108469
7	Authenticity, Gossip, Curiosity-seeking, Enthusiastic, Ignore, Highly, Verification, Headline, Inflammatory, Visual	scale-oriented	0.086632	0.121577	0.080104	0.099366
8	EN, Self-regulation, EX, SP, Increase, Flaw, SRD, Entertaining, Although, Suppression	hypo-oriented	0.048062	0.041248	0.140707	0.085917

Table 14: Scale filtering results under different scale sizes

Theory	#scales	$F1_{no}$	$F1_{med}$	AUC
TPB-NAM	1/3	0.6757	0.6525	0.7021
	2/3	0.6794	0.6822	0.7088
	Full	0.6625	0.6751	0.7042
A-CLT	1/3	0.6854	0.6834	0.7254
	1/2	0.6736	0.6905	0.7252
	Full	0.6761	0.6745	0.7059

The performance differences across scale lengths are relatively rare, and sometimes filtered scales would perform a little better, suggesting that profile-matched scale filtering can reduce prompt length and computational cost while preserving most of the predictive signal provided by psychological measurement.

D.4 Theoretical Path Validation using LLM-simulated Data

To further examine whether the LLM-simulated scale scores preserve theory-consistent psychological relationships, we input the LLM-generated scale results into the original statistical models, such as regression and structural equation models (SEMs) associated with the corresponding theoretical frameworks. We then compare the estimated paths with those significant paths reported from human questionnaire responses in the original studies, resulting in number of Mathched Paths and Path Hit Rate. In addition, we define an Emergent Significance Count (ESC), which captures the number of structural paths that were non-significant in the original human-data-based survey but become statistically significant when estimated using LLM-simulated responses.

The results in Table 15 show that the structural relationships inferred from LLM-simulated scale scores are largely consistent with those obtained from human questionnaire data in the original studies. However, the degree of consistency varies across scales, indicating that LLM-based psychological measurement may better reproduce some theoretical structures than others. Notably, in several cases the LLM-simulated data not only preserves existing significant paths but also induces additional statistically significant relationships that were absent in the original human-based analyses, as reflected by non-zero ESC values. This suggests that LLM-generated responses may amplify or regularize latent structural dependencies beyond those observed in human survey data. Overall, these find-

Table 15: Path Significance consistency between LLM-simulated scale scores-based results and original human-survey data-based results. “Original Paths” refers to number of significant paths reported from human questionnaire responses in the original studies; Matched Paths refers the number of matched significant paths using LLM-simulated data; “ Path Hit Rate" is the hit rate; “’ESC’ denotes the number of paths that were non-significant using the original survey but statistically significant using LLM-generated data.

Scale/Theory	Model	Original Paths	Matched Paths	Path Hit Rate	ESC
MOA-SOR	DS	5	5	100.00%	0
MOA-SOR	GLM	5	5	100.00%	0
MOA-SOR	GPT	5	5	100.00%	0
MOA-SOR	GROK	5	4	80.00%	0
TPB-NAM	DS	13	10	76.92%	2
TPB-NAM	GLM	13	11	84.62%	2
TPB-NAM	GPT	13	11	84.62%	2
TPB-NAM	GROK	13	11	84.62%	2
A-CIT	DS	11	7	63.64%	2
A-CIT	GLM	11	7	63.64%	1
A-CIT	GPT	11	6	54.55%	2
A-CIT	GROK	11	8	72.73%	2
SOR-ELM	DS	7	6	85.71%	1
SOR-ELM	GLM	7	6	85.71%	1
SOR-ELM	GPT	7	6	85.71%	1
SOR-ELM	GROK	7	6	85.71%	1

ings provide preliminary support for the psychological validity of LLM-simulated scale responses, while also highlighting scale-dependent variation and the need for careful validation when interpreting emergent effects.