

Detection Is Not The End: Personalized Modeling of User Susceptibility to Misinformation

Peng Qi and Zehong Yan and Wynne Hsu and Mong-Li Lee

National University of Singapore

{peng.qi, dcshsuw, dcsleeml}@nus.edu.sg, zyan@u.nus.edu

Abstract

The spread of misinformation poses serious societal harms. While detection has been extensively studied, modeling user susceptibility remains a challenging yet important direction for strengthening users' resilience. In this work, we formally define the Personalized Misinformation Susceptibility Prediction (MiSP) task, predicting whether a specific user is susceptible to (*i.e.*, believe) a particular misinformation claim. We construct benchmark datasets derived from social science surveys and social-platform data, and design diverse evaluation settings covering different seen/unseen user–claim scenarios. Using this benchmark, we examine LLMs' ability to simulate personalized susceptibility under realistic multi-attribute user profiles, and further evaluate LM-, LLM-, and GNN-based approaches, revealing their limitations. Motivated by these findings, we propose VulneraSpotter, which integrates semantic information and behavioral patterns via heterogeneous and homogeneous GNNs, enabling effective generalization to previously unseen users and claims. Extensive experiments demonstrate that VulneraSpotter outperforms existing baselines in predicting personalized susceptibility, uncovers meaningful group-level patterns, and shows promising potential for mitigating misinformation spread. This work provides a benchmark and modeling foundation for user-centric misinformation understanding and mitigation.

1 Introduction

Existing studies on misinformation (Chen and Shu, 2024) have largely focused on automatic detection, *i.e.*, judging whether a given claim is true or false. Although veracity assessment is indispensable, it alone is insufficient to comprehensively mitigate the spread of misinformation, particularly in the early stages of emerging events when evidence is too scarce to support reliable verification (Wei et al., 2025). An important complementary measure is

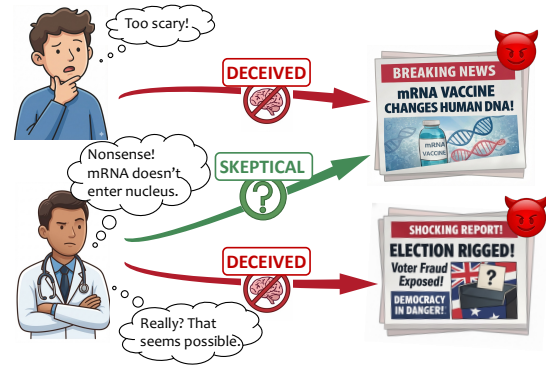


Figure 1: Different people are susceptible to different types of misinformation, highlighting the necessity of personalized modeling for user susceptibility to misinformation.

to enhance users' resilience by modeling how individuals perceive misinformation, enabling interventions such as psychological inoculation (Roosenbeek et al., 2022) and personalized pre-debunking (Gabriel et al., 2024).

Prior works (Gabriel et al., 2022; Shen et al., 2025) study susceptibility from the perspective of claim content, *i.e.*, whether the given misinformation can be identified by the general public. However, susceptibility is inherently personalized: individuals differ in their ability to discern misinformation due to variations in education, life experiences, and cognitive abilities. As shown in Figure 1, health-related misinformation might mislead the general public but would be recognized by medical professionals, whereas doctors may in turn be vulnerable to political misinformation. Our preliminary analysis further confirms such personalization in real-world scenarios (Appendix B).

For personalized misinformation susceptibility, prior work primarily relies on users' sharing behaviors as supervision signals (Liu et al., 2024; Teng et al., 2022). However, sharing is a very noisy proxy: users may share false headlines they judge as inaccurate, while non-sharing may reflect limited exposure or reluctance to repost rather than

disbelief (Arechar et al., 2023). The subjectivity of susceptibility also makes labels difficult to validate with external annotators. Other than sharing-based approaches, Ma et al. (2024) use LLMs to simulate how single demographic attributes influence susceptibility. However, simulating each attribute in isolation fails to reflect the complex interactions among multiple attributes that naturally coexist in real users. Also, the accuracy of such simulations is hard to assess due to the lack of alignment with real-world datasets with ground-truth labels. Thus, personalized susceptibility modeling remains challenging, largely due to the scarcity of reliable benchmarks with accurate labels.

To bridge this research gap, we *formally define the Personalized Misinformation Susceptibility Prediction (MiSP) task*, which predicts whether a given user is susceptible to a piece of misinformation. To provide accurate learning signals, we build benchmarks from two complementary sources: we repurpose social-science survey datasets with direct susceptibility labels (Arechar et al., 2023; Spampatti et al., 2024), and construct a social-platform benchmark by inferring such labels from users’ stances toward misinformation claims (Nielsen and McConville, 2022). These benchmarks enable direct modeling of susceptibility rather than relying only on indirect sharing behaviors. We further *define four testing settings covering different combinations of previously seen/unseen users and claims* to comprehensively simulate realistic scenarios.

Building on this benchmark, we first *investigate whether LLMs (e.g., GPT-4o) can simulate personalized susceptibility under realistic multi-attribute user profiles*, an aspect unexplored in prior works that typically consider only one attribute at a time. Using non-parametric prompting, including zero-shot, direct few-shot, and our proposed story-based few-shot prompting, LLMs struggle to exhibit personalized simulation and often assume that users can correctly identify misinformation regardless of their distinctive profiles. We further *evaluate LM, parametric and non-parametric LLM, and GNN-based methods*, showing that accurate susceptibility prediction requires both semantic understanding (e.g., what are the characteristics of this user or claim?) and historical behavioral patterns learned from training data (e.g., what types of misinformation this user is likely to believe, and what kinds of users tend to be deceived by this claim?).

Motivated by these findings, we *propose a novel model, VulneraSpotter, which explicitly models*

and integrates both the semantic and historical behavior features to predict personalized misinformation susceptibility. VulneraSpotter uses heterogeneous user-claim and homogeneous user-user/claim-claim GNNs to integrate semantic and behavioral signals and propagate behavior patterns of seen nodes to semantically similar unseen nodes. Experiments demonstrate that VulneraSpotter not only predicts individual susceptibility more accurately than existing baselines, but also captures meaningful group-level susceptibility patterns by revealing how user attributes relate to susceptibility. Beyond detection accuracy, we further demonstrate the potential of our susceptibility prediction model to reduce misinformation harms through a simplified application scenario. The contributions of this work can be summarized as:

- **Benchmark Dataset and Task Formulation.** We build benchmark datasets with susceptibility labels from both social-science surveys and social-platform discussions, enabling direct modeling of susceptibility rather than relying only on indirect sharing behaviors. We further formalize the MiSP task and define diverse testing scenarios, providing a foundation for studying personalized susceptibility for misinformation mitigation.
- **Extensive Baseline Evaluation.** We systematically evaluate LM, LLM, and GNN-based baselines. In particular, our analysis reveals the challenge of LLM simulation under realistic multi-attribute user profiles.
- **Strong Graph-based Classifier.** We propose VulneraSpotter, which integrates semantic and behavioral signals with hetero- and homogeneous GNNs, achieving strong generalization across realistic scenarios. Experiments further verify its effectiveness at both the individual and group levels and indicate its potential to reduce misinformation propagation.

2 Task Formulation and Benchmark Construction

Definition 1 (Personalized Misinformation Susceptibility Prediction (MiSP)). *Given labeled pairs $\mathcal{P} = \{(u_k, i_k, y_k)\}_{k=1}^n$, where u_k is a user profile, i_k is a misinformation claim, and $y_k \in \{0, 1\}$ indicates whether the user is susceptible to the claim (1) or not (0), the task is to learn a classifier $\mathcal{F}$ to predict $y_k = \mathcal{F}(u_k, i_k)$.*

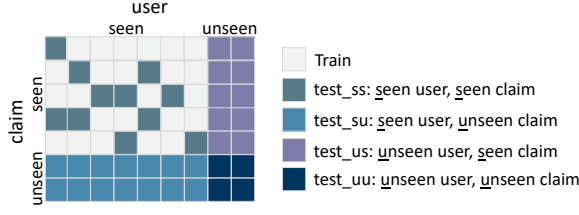


Figure 2: Different test scenarios based on whether users and claims have appeared in the training set.

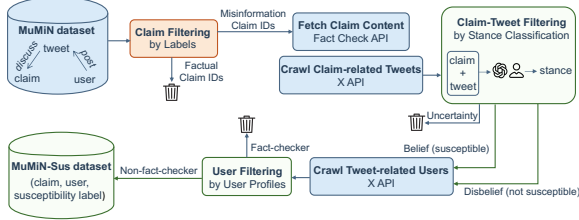


Figure 3: Construction pipeline of the social-platform susceptibility dataset MuMiN-Sus.

In real-world scenarios, users and claims encountered at test time may be seen or unseen during training, introducing different generalization challenges. We therefore design four distinct testing scenarios (Figure 2) to capture these situations (Appendix C). For example, for a user–claim pair (u, i) in test_{ss} to be classified, the training set contains labeled examples involving user u with other claims and claim i with other users. From test_{ss} to test_{uu}, behavioral information becomes sparser and the model relies more heavily on content semantics, resulting in substantial performance differences for the same method across scenarios. Thus, we evaluate MiSP performance under all four settings to assess the model’s generalization ability.

Benchmark Dataset Construction. A major obstacle for the task is the scarcity of benchmark datasets with reliable susceptibility labels. To support systematic evaluation, we propose a three-dataset benchmark from two complementary sources. First, we reorganize two social-science survey datasets on COVID-19 (Arechar et al., 2023) and Climate (Spampatti et al., 2024) misinformation. These surveys directly ask participants to judge the truthfulness of misinformation claims and provide user attributes such as demographics and cognitive styles. Second, we construct MuMiN-Sus, a social-platform dataset derived from MuMiN (Nielsen and McConville, 2022), which links fact-checked claims with tweets and users. Because MuMiN lacks direct susceptibility labels, we derive them through stance classification (Figure 3):

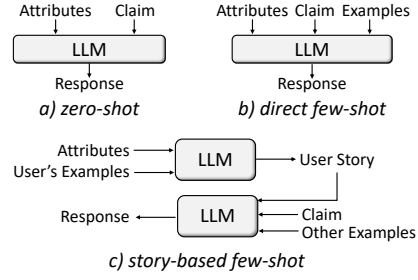


Figure 4: LLM-based susceptibility simulation methods.

we annotate each claim–tweet pair as belief, disbelief, or uncertainty with GPT followed by manual verification. We map belief to susceptible and disbelief to not susceptible, and remove uncertain pairs. We further collect user profiles and filter fact-checker-related accounts, whose profile cues and overwhelmingly not-susceptible labels would otherwise introduce shortcut bias. The resulting dataset consists of user–claim pairs with stance-derived susceptibility labels and user profiles. Appendix D provides detailed dataset introduction and statistics.

3 Exploratory Study of LLM Simulations

Inspired by the success of LLM-based role-playing for user behavior simulation (Wang et al., 2024; Park et al., 2023), Ma et al. (2024) simulate how user attributes influence misinformation susceptibility with LLM agents. However, this work considers only one user attribute at a time, while real users necessarily have multiple co-existing attributes (e.g., gender and education) whose susceptibility signals can be contradictory. Therefore, using our constructed COVID benchmark as a representative case, we conduct an exploratory analysis to evaluate whether LLMs can simulate susceptibility under realistic multi-attribute user profiles.

As shown in Figure 4, we use both zero-shot and few-shot prompting with LLMs to simulate user susceptibility. GPT-4o is instructed to play the given user and judge the truthfulness of the given misinformation claim. Beyond direct few-shot, we propose an enhanced story-based few-shot approach, inspired by the story-based zero-shot in (Ma et al., 2024). We first generate a life-experience narrative for the target user, based on their attributes and past annotations. For response generation, we supply examples of how other users annotated the target claim and other semantically similar examples. Appendix E provides prompt and example-selection details.

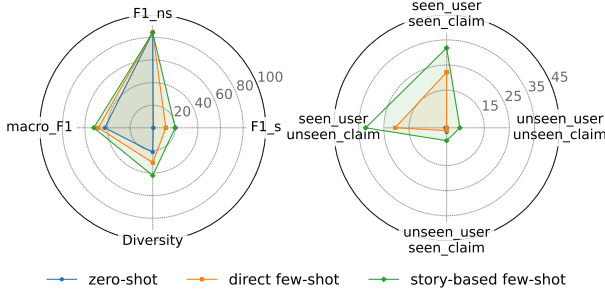


Figure 5: Performance (%) comparison of three LLM-based susceptibility simulation methods: four metrics on the full test set (left) and F1_s on four test subsets (right).

Results. We report macro-F1 and class-wise F1 for “susceptible” (F1_s) and “not susceptible” (F1_ns) classes. Given that LLM simulations often display homogenized predictions (Cheng et al., 2023), we quantify the Diversity of the predicted label distribution (Appendix A). Figure 5 shows that although F1_ns exceeds 80%, F1_s remains below 20% and is nearly zero in the zero-shot setting. This suggests that, under the tested non-parametric prompting settings, LLMs tend to predict that users will not be deceived, rather than reliably reproducing personalized susceptibility patterns. Story-based few-shot outperforms direct few-shot, which in turn outperforms zero-shot, primarily through gains in F1_s. These results suggest that the semantics of user profiles and claim content are insufficient for modeling personalized susceptibility, whereas historical behaviors offer stronger guidance. The right panel further shows that, with the same number of examples, using a user’s historical annotations yields better performance than using a claim’s historical annotations or semantically similar examples. This suggests that user-specific historical behavior is particularly informative for personalized susceptibility modeling. Overall, these findings motivate models that go beyond profile and claim semantics alone and explicitly integrate historical user–claim interaction patterns.

4 Proposed VulneraSpotter Model

4.1 Overview

Motivated by the above analysis, the MiSP task poses two key challenges. *The first is to effectively model and integrate semantic and behavioral information.* In MiSP, a user’s susceptibility pattern can be inferred from the claims they have interacted with, while a claim’s susceptibility profile can be inferred from the users who have responded to it. Thus, users and claims mutually characterize each

other through interaction behaviors, with their representations iteratively refined in a mutually reinforcing manner. Such high-order dependencies can be naturally captured through graph aggregation. Accordingly, we construct a user–claim bipartite graph from observed interactions and employ heterogeneous graph learning to integrate semantic and behavioral information.

The second challenge lies in generalizing to unseen users and claims, since behavioral information is only available for seen nodes during training. To address this, we build homogeneous user–user and claim–claim graphs based on semantic similarity, enabling behavioral patterns learned from seen nodes to propagate to unseen ones through graph message passing. The underlying assumption is that semantically similar users (or claims) exhibit similar behavioral tendencies. We further apply feature alignment on seen nodes to align semantic representations with both behavioral and final aggregated embeddings, thereby calibrating the semantic space with behavior-relevant patterns. Thus, despite lacking behavioral signals, unseen nodes can still benefit from more distinctive semantic representations. We propose the VulneraSpotter model (Figure 6) to address these challenges.

4.2 Heterogeneous Graph Learning

To better model behavioral and semantic information, we construct behavior- and semantic-aware graphs with the same structure but distinct node features to learn complementary embeddings.

Graph Construction. Given users $u \in \mathcal{U}$ and claims $i \in \mathcal{I}$, we model the observed interactions as a user–claim bipartite graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where the node set $\mathcal{V} = \mathcal{U} \cup \mathcal{I}$ and an edge $(u, i) \in \mathcal{E}$ indicates that user u has interacted with claim i . For the behavior-aware graph, each user and claim node is initialized with a one-hot ID embedding $\mathbf{e}_u^{(0)}$ and $\mathbf{e}_i^{(0)}$, following Wang et al. (2019). For the semantic-aware graph, nodes are initialized with semantic features of user profiles and claim contents, denoted as $\mathbf{f}_u^{(0)}$ and $\mathbf{f}_i^{(0)}$, respectively. Motivated by LLMs’ strong ability to capture preference-related semantics (Wei et al., 2024), we adopt an LLM-based encoder to obtain semantic features. To make the learned representations more predictive of misinformation susceptibility, we attach a task-specific classification head and fine-tune the encoder with LoRA (Appendix F).

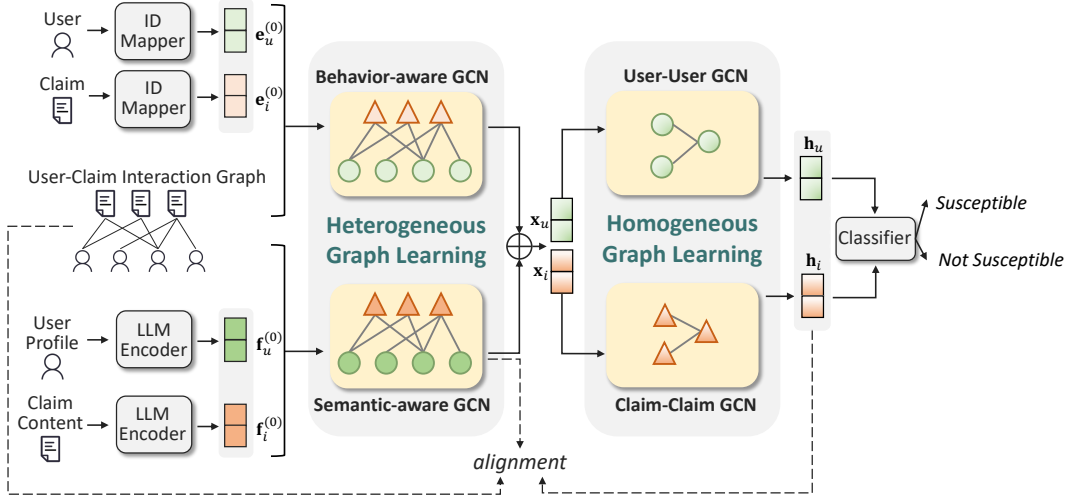


Figure 6: VulneraSpotter Architecture. We first model high-order user-claim interactions with a heterogeneous graph, and then capture intra-user and intra-claim relations using homogeneous graphs. The resulting embeddings are combined for binary susceptibility classification.

Graph Aggregation. To inject user-claim collaborative signals into node representations, we adopt LightGCN (He et al., 2020b), a light yet effective variant of GCN, for recursive message passing over the interaction graph. Taking the behavior-aware graph as an example, node representations are iteratively updated by aggregating features from their neighbors:

$$\mathbf{e}_u^{(k+1)} = \sum_{i \in \mathcal{N}_u} \frac{\mathbf{e}_i^{(k)}}{\sqrt{|\mathcal{N}_u|} \sqrt{|\mathcal{N}_i|}}, \quad \mathbf{e}_i^{(k+1)} = \sum_{u \in \mathcal{N}_i} \frac{\mathbf{e}_u^{(k)}}{\sqrt{|\mathcal{N}_i|} \sqrt{|\mathcal{N}_u|}}, \quad (1)$$

where $\mathcal{N}_u$ and $\mathcal{N}_i$ are the neighborhood sets of u and i in $\mathcal{G}$. After K propagation layers, we obtain the final behavior-aware embeddings by averaging representations across layers, denoted as $\mathbf{e}_u = \sum_{k=0}^{K-1} \frac{\mathbf{e}_u^k}{K}$ and $\mathbf{e}_i = \sum_{k=0}^{K-1} \frac{\mathbf{e}_i^k}{K}$. Unseen nodes skip this stage since they have no observed interactions in $\mathcal{G}$. The semantic-aware GCN operates similarly, yielding semantic-aware embeddings $\mathbf{f}_u$ and $\mathbf{f}_i$. Since behavior-aware and semantic-aware embeddings provide complementary information, we aggregate them using a weighted sum to obtain the fused user and claim representations $\mathbf{x}_u$ and $\mathbf{x}_i$.

4.3 Homogeneous Graph Learning

Graph Construction. Given a set of claims $i \in \mathcal{I}$, we construct a claim-claim homogeneous graph $\mathcal{G}_I = \{\mathcal{I}, \mathcal{E}_I\}$, where $\mathcal{E}_I = \{s_{i,i'} \mid i, i' \in \mathcal{I}\}$. For the edge $s_{i,i'}$ connecting the claim i and i' , its weight is calculated by the cosine similarity between the initial semantic features $\mathbf{f}_i^{(0)}$ and $\mathbf{f}_{i'}^{(0)}$. This results in a fully connected graph, to which we employ kNN pruning (Chen et al., 2009) to

yield a sparse one $\tilde{\mathcal{G}}_I$, by retaining only the top- k most similar neighbors for each node. This sparsification better aligns with real-world settings, where each claim (or user) is typically related to only a limited number of counterparts. We construct the user-user homogeneous graph $\tilde{\mathcal{G}}_U$ in the same manner. Following (He et al., 2024), these graphs are built using only seen nodes during training and expanded to include all nodes at inference time.

Graph Aggregation. Similar to the heterogeneous graph, we apply LightGCN for graph aggregation. To prevent unrealistic propagation from unseen to seen nodes, we introduce an additional mask during message passing:

$$\mathbf{h}_i^{(l+1)} = \sum_{i' \in \mathcal{N}_i} \frac{\mathbf{h}_{i'}^{(l)}}{\sqrt{|\mathcal{N}_i|} \sqrt{|\mathcal{N}_{i'}|}} \cdot \mathbb{I}(i \notin \mathcal{I}_{\text{seen}} \text{ or } i' \notin \mathcal{I}_{\text{unseen}}), \quad (2)$$

where $\mathbf{h}_i^0$ is initialized with $\mathbf{x}_i$, which is learned from heterogeneous graph learning, and $\mathbb{I}$ denotes the indicator function. After L convolutional layers, we obtain the final claim representations $\mathbf{h}_i = \mathbf{h}_i^L$ and user representations $\mathbf{h}_u$.

4.4 Optimization

We concatenate the final user and claim representations for binary classification and optimize the classifier with the cross-entropy loss $\mathcal{L}_{\text{cls}}$. To improve generalization to unseen nodes, we further regularize seen-node representations with two self-supervised objectives (Wei et al., 2023): contrastive learning aligns semantic embeddings with final embeddings that integrate semantic and be-

	Method	test					test_ss	test_su	test_us	test_uu	
		macro_F1	F1_s	F1_ns	AUC	Acc.	macro_F1	macro_F1	macro_F1	macro_F1	
COVID	LM	RoBERTa	0.569	0.407	0.731	0.572	0.630	0.594	0.563	0.551	0.560
		Longformer	0.523	0.324	0.722	0.525	0.606	0.526	0.509	0.532	0.520
	LLM	GPT-4o_zs	0.424	0.006	<u>0.843</u>	0.551	0.705	0.429	0.416	0.430	0.421
		GPT-4o_fs	0.481	0.117	0.846	0.526	0.715	0.543	0.522	0.426	0.421
		GPT-4o_storyfs	0.521	0.199	<u>0.843</u>	0.544	<u>0.727</u>	0.592	0.579	0.442	0.446
		LLaMA-2_zs	0.222	0.016	<u>0.427</u>	0.501	<u>0.276</u>	0.229	0.228	0.213	0.214
		LLaMA-2_fs	0.415	0.312	0.518	0.556	0.671	0.406	0.615	0.335	0.314
		LLaMA-2_storyfs	0.434	0.387	0.481	0.571	0.683	0.432	0.641	0.342	0.321
		LLaMA-2_peft	0.601	<u>0.453</u>	0.748	<u>0.615</u>	0.643	0.618	<u>0.572</u>	0.591	
	GNN	NGCF	0.609	0.387	0.831	0.601	0.715	<u>0.703</u>	0.704	0.504	0.453
		Dragon	<u>0.613</u>	0.387	0.840	0.604	0.725	0.716	0.679	0.422	0.430
		Firzen	0.607	0.374	0.840	0.599	0.716	0.696	0.693	0.422	0.424
		VulneraSpotter (ours)	0.647*	0.488*	0.807	0.648*	0.734*	0.697	<u>0.698</u>	0.625*	<u>0.570</u>
Climate	LM	RoBERTa	0.587	0.665	0.507	0.589	0.601	0.623	0.543	0.625	0.552
		Longformer	0.571	0.623	0.520	0.581	0.577	0.624	0.519	0.625	0.510
	LLM	GPT-4o_zs	0.404	0.243	0.566	0.544	0.448	0.467	0.336	0.464	0.330
		GPT-4o_fs	0.486	0.400	0.573	0.564	0.500	0.533	0.410	0.542	0.438
		GPT-4o_storyfs	0.535	0.498	<u>0.572</u>	0.589	0.516	0.593	0.472	0.570	0.468
		LLaMA-2_zs	0.296	0.032	0.559	0.502	0.394	0.343	0.282	0.274	0.280
		LLaMA-2_fs	0.398	0.235	0.561	0.558	0.500	0.396	0.633	0.274	0.280
		LLaMA-2_storyfs	0.417	0.268	0.566	0.570	0.521	0.436	<u>0.657</u>	0.301	0.289
		LLaMA-2_peft	0.617	0.664	0.571	<u>0.627</u>	0.623	<u>0.656</u>	0.586	0.651	0.575
	GNN	NGCF	0.591	<u>0.733</u>	0.449	0.593	0.640	0.676	0.565	<u>0.634</u>	0.399
		Dragon	0.600	0.647	0.552	0.609	0.606	0.649	0.608	0.617	0.495
		Firzen	0.621	0.717	0.525	0.619	<u>0.645</u>	0.638	0.593	0.623	<u>0.625</u>
		VulneraSpotter (ours)	0.659*	0.745*	0.573	0.657*	0.681*	0.644	0.711*	0.618	0.653*
MuMin-Sus	LM	RoBERTa	0.735	0.778	0.691	0.816	0.742	0.832	0.792	0.667	0.646
		Longformer	0.744	0.798	0.692	0.817	0.756	0.849	0.793	0.678	0.654
	LLM	GPT-4o_zs	0.527	0.422	0.632	0.611	0.551	0.548	0.476	0.559	0.519
		GPT-4o_fs	0.593	0.513	0.675	0.742	0.603	0.671	0.563	0.649	0.489
		GPT-4o_storyfs	0.642	0.597	0.688	0.694	0.648	0.828	0.684	0.577	0.442
		LLaMA-2_zs	0.299	0.016	0.582	0.499	0.413	0.305	0.298	0.299	0.294
		LLaMA-2_fs	0.588	0.638	0.538	0.588	0.594	0.724	0.512	0.616	0.492
		LLaMA-2_storyfs	0.623	0.659	0.587	0.759	0.615	0.751	0.542	0.635	0.564
		LLaMA-2_peft	0.758	0.815	0.705	0.837	0.773	0.864	<u>0.811</u>	0.691	0.665
	GNN	NGCF	0.735	0.783	0.688	0.836	0.744	0.901	0.768	<u>0.750</u>	0.521
		Dragon	0.744	0.792	0.696	0.841	0.756	0.869	0.782	0.732	0.593
		Firzen	0.769	<u>0.827</u>	<u>0.711</u>	<u>0.863</u>	<u>0.788</u>	0.858	0.786	0.745	<u>0.685</u>
		VulneraSpotter (ours)	0.803*	0.865*	0.740*	0.902*	0.821*	<u>0.889</u>	0.838*	0.751	0.714*

Table 1: Performance comparison. The best results are in **boldface**, while the second-best is underlined. * denotes the improvement over the best baseline is statistically significant (*i.e.*, paired two-tailed t -test with $p < 0.05$).

havioral information, while adversarial training aligns semantic-induced interaction patterns with behavior-induced ones. The training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{cl}}\mathcal{L}_{\text{cl}} + \lambda_{\text{adv}}\mathcal{L}_{\text{G}} + \lambda_{\text{reg}}\|\theta\|^2, \quad (3)$$

where $\mathcal{L}_{\text{cl}}$ and $\mathcal{L}_{\text{G}}$ denote the contrastive and adversarial losses, respectively, and the last term prevents overfitting. We alternatively optimize $\mathcal{L}$ and the discriminator loss $\mathcal{L}_{\text{D}}$; detailed formulations are provided in Appendix G.

5 Performance Study

5.1 Experimental Setting

Evaluation Setup. We conduct experiments on the proposed three benchmark datasets. We report macro-F1, class-wise F1, AUC, and Accuracy on the full test set, together with macro-F1 on each

subset. Since the datasets are class-imbalanced and the full test set reflects overall performance across all testing scenarios, we treat full-set macro-F1 as the primary comparison metric. Implementation details are provided in Appendix F.

Baselines. To establish a comprehensive benchmark, we evaluate multiple types of methods that leverage information from different perspectives: 1) *LM (language models)-based methods*, which serve as representatives of text classification models that leverage semantic information. Here, we test RoBERTa (Liu et al., 2019) and Longformer (Beltagy et al., 2020), which take the concatenated user profile and claim content as input and perform classification. 2) *LLM-based methods*: We evaluate both closed-source and open-source LLMs, including GPT-4o (OpenAI, 2024) and LLaMA-2 (Meta,

	Method	All	test_ss	test_su	test_us	test_uu
COVID	VulneraSpotter	0.647	0.697	0.698	0.625	0.570
	w/o semantic	0.605	0.698	0.663	0.422	0.422
	w/o behavior	0.630	0.721	0.709	0.522	0.490
	w/o homo-graph	0.618	0.717	0.650	0.422	0.546
	w/o SSL loss	0.628	0.679	0.697	0.551	0.568
	w/o LLM encoder	0.610	0.724	0.699	0.425	0.421
Climate	VulneraSpotter	0.659	0.644	0.711	0.618	0.653
	w/o semantic	0.600	0.668	0.593	0.633	0.383
	w/o behavior	0.642	0.639	0.659	0.634	0.632
	w/o homo-graph	0.649	0.655	0.658	0.630	0.645
	w/o SSL loss	0.648	0.648	0.670	0.626	0.642
	w/o LLM encoder	0.618	0.653	0.643	0.571	0.600
MuMiN-Sus	VulneraSpotter	0.803	0.889	0.838	0.751	0.714
	w/o semantic	0.754	0.904	0.811	0.742	0.560
	w/o behavior	0.781	0.855	0.823	0.745	0.701
	w/o homo-graph	0.788	0.860	0.831	0.683	0.678
	w/o SSL loss	0.792	0.837	0.832	0.702	0.697
	w/o LLM encoder	0.758	0.886	0.804	0.693	0.649

Table 2: Ablation studies in VulneraSpotter (macro-F1).

2023). Each LLM is tested under zero-shot (zs), direct few-shot (fs) and story-based few-shot (storyfs) settings. Open-source LLMs are also evaluated under parameter-efficient fine-tuning (peft) setting using LoRA. 3) *GNN-based methods*: Given the relevance of collaborative signals in the MiSP task, we borrow representative graph-based methods from recommendation systems, including NGCF (Wang et al., 2019), Dragon (Zhou et al., 2023) and Firzen (He et al., 2024). We retain their original graph structures and message-passing designs, and adapt the final prediction objective from recommendation ranking to binary susceptibility classification.

5.2 Performance Comparison

From Table 1, we observe that VulneraSpotter achieves superior classification performance on the full test set and delivers consistently competitive performance across different generalization scenarios compared to baseline methods, demonstrating its effectiveness and robustness for personalized misinformation susceptibility prediction. Regarding different categories of baselines, LM- and LLM-based methods—which primarily rely on semantic information—exhibit relatively balanced performance across subsets involving seen or unseen nodes. Parameter-efficient fine-tuning substantially improves LLM-based methods, yet they still underperform GNN-based approaches that explicitly model behavioral interactions, particularly in settings involving seen nodes. However, despite their strengths in leveraging user–claim interactions, GNN-based methods suffer pronounced performance drops on subsets with unseen nodes, due to their limited ability to generalize semantic representations. For example, Dragon’s macro-F1

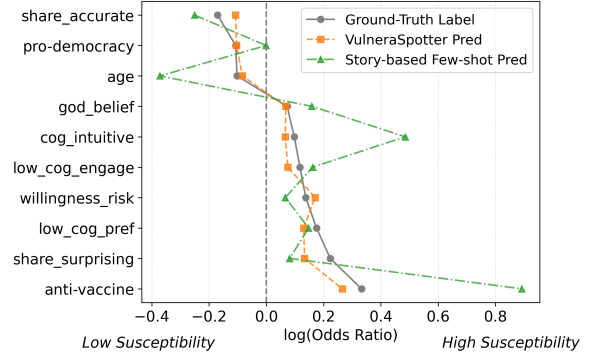


Figure 7: User attribute effects on misinformation susceptibility derived from ground-truth labels and model predictions. cog=cognition. pref=preference.

in the COVID dataset decreases from 0.716 on test_ss to 0.430 on test_uu, highlighting persistent generalization challenges. Overall, these results validate the MiSP task and underscore the necessity of jointly leveraging semantic and behavioral information and robust generalization mechanisms.

5.3 Ablation Studies

We conduct ablation studies to assess the contribution of each component in VulneraSpotter by removing them separately: *w/o semantic* (removing the semantic-aware GCN and homogeneous graph learning), *w/o behavior* (removing heterogeneous graph learning), *w/o homo-graph* (removing homogeneous graph learning), *w/o SSL loss* (removing contrastive and adversarial losses), and *w/o LLM encoder* (replacing the fine-tuned LLM with OpenAI’s text-embedding-3 model¹ as the feature encoder). Table 2 shows that: i) All components contribute positively to the aggregated full-test performance, although some components may trade off performance across different generalization scenarios. ii) Semantic and behavioral information provide complementary benefits, with semantic information playing a more dominant role, particularly in scenarios involving unseen nodes. For example, removing semantic features reduces macro-F1 by 15-27 percentage points on the test_uu set. iii) The fine-tuned LLM encoder, homogeneous graph learning, and the SSL loss consistently improve performance, especially on subsets with unseen nodes. It shows that the task-adapted semantic encoder provides the primary cold-start representation, while homogeneous graph propagation and SSL are also important for generalization by providing complementary behavior-aware refinement.

¹<https://platform.openai.com/docs/guides/embeddings/>

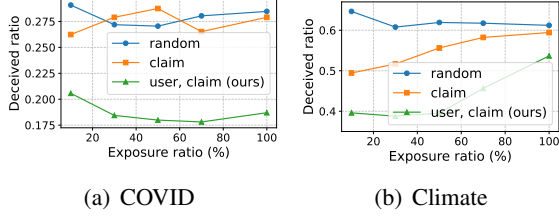


Figure 8: Misinformation harms under different intervention strategies. Exposure ratio is the proportion of candidate users who are exposed to misinformation, while deceived ratio is the proportion of exposed users who believe the misinformation.

5.4 Attribute Effect Analysis

We further analyze how user attributes relate to predicted susceptibility through logistic regression, using the COVID dataset as an example. As shown in Figure 7, VulneraSpotter’s predictions yield attribute–susceptibility associations that closely match those derived from ground-truth labels in both direction and effect size (*e.g.*, valuing news accuracy reduces susceptibility, whereas anti-vaccine attitudes increase it). These findings are also consistent with prior research (Arechar et al., 2023). In contrast, story-based few-shot prompting with GPT-4o (a strong prompt-based LLM simulation) fails to capture several key attribute effects (*e.g.*, pro-democracy attitudes) and overestimates others (*e.g.*, age, anti-vaccine beliefs). These findings illustrate not only the mechanism underlying our model but also its accuracy in capturing group-level susceptibility patterns, complementing the individual-level results presented in Table 1.

5.5 Effects in Mitigating Misinformation Harms

An intuitive application of user misinformation susceptibility modeling is risk-aware exposure control. We consider a simplified scenario where some false claims bypass detection, and the platform must decide which candidate user–claim exposures to retain or downrank. Susceptibility prediction can complement detection by reducing exposures to users who are more likely to believe undetected misinformation. For each false claim, candidate users are ranked by predicted susceptibility, and only a fixed proportion of exposures is retained as the exposure budget. Harm is measured by the deceived ratio, *i.e.*, the proportion of truly susceptible users among retained exposures.

We compare three strategies: *random*, which assigns susceptibility labels with 50% probability; *claim-based* (Gabriel et al., 2022), which estimates

susceptibility solely based on claim content; and *user, claim-based (ours)*, which predicts susceptibility conditioned on both the user and the claim. As shown in Figure 8, our method consistently yields a lower deceived ratio than the others. While real-world news recommendations are considerably more complex than this simplified simulation, these results provide encouraging evidence that modeling users’ personalized susceptibility is a viable direction for mitigating misinformation harm.

6 Related Work

Prior works mostly focus on automatic detection, with limited attention to how users perceive misinformation. Gabriel et al. (2022) propose a pragmatic formalism to model users’ multi-dimensional reactions to misinformation, but their analysis targets general user groups and overlooks individual differences. Social science studies (Arechar et al., 2023; Ali et al., 2020; Spampatti et al., 2024) highlight how user attributes (such as analytic thinking or accuracy motivations) influence susceptibility, yet they mainly rely on correlational analyses and lack practical predictive models. To bridge this gap, Liu et al. (2024); Teng et al. (2022) propose computational methods to predict susceptibility from sharing behaviors, but these methods lack direct ground-truth validation, which limits the reliability of the learned scores. Recent efforts (Ma et al., 2024; Pratelli and Petrocchi, 2025) use LLMs to simulate user susceptibility by constructing agents from demographic descriptions (such as age and personality), but these simulations vary only one attribute at a time and miss the multi-attribute complexity of real-world users. In contrast, our work targets personalized susceptibility, proposing a real-world benchmark and novel prediction model, and assessing diverse approaches under realistic multi-attribute profiles.

7 Conclusion

In this paper, we formalize the task of personalized misinformation susceptibility prediction as a complementary perspective to detection for mitigation. To establish this benchmark, we construct susceptibility-labeled benchmarks from social-science surveys and social-platform discussions, define diverse seen/unseen user-claim evaluation settings, and evaluate LM-, LLM-, and GNN-based methods. Experiments reveal both the feasibility and the inherent challenges of the MiSP

task. Our proposed model jointly integrates semantic and behavioral information through hetero- and homogeneous GNNs, achieving superior prediction performance with robust generalization.

Acknowledgments

This work was supported by the Ministry of Education, Singapore, under its MOE AcRF TIER 3 Grant (MOE-MOET32022-0001). The authors thank the developers of the open-source frameworks, models, and libraries that supported the implementation and experiments in this work.

Limitations

Label Granularity. We formulate MiSP as a binary classification task, following prior work (Teng et al., 2022). However, susceptibility can be more fine-grained in practice. Our survey datasets contain six ordinal annotations, which could support future extensions to ordinal regression or multi-class susceptibility prediction. Our framework can also be naturally extended to support more fine-grained definitions, while keeping the core model architecture unchanged and requiring only minor modifications to the existing prediction head.

Real-world Validation. Our mitigation analysis is based on a simplified news recommendation simulation rather than deployment in real platforms. Although it demonstrates the potential utility of susceptibility modeling, real-world systems involve additional factors such as user interest, exposure history, engagement incentives, and social influence. Future work should evaluate susceptibility-aware modeling in more realistic recommendation or intervention pipelines.

Ethical Statement

Data Ethics. Our experiments rely on publicly available datasets from previously published works (Arechar et al., 2023; Spampatti et al., 2024; Nielsen and McConville, 2022). The survey datasets contain anonymized responses and no personally identifiable information; all data were collected with informed consent, and the original studies received ethical review exemptions from their respective institutional review boards. For the social-platform dataset derived from MuMiN, we use only publicly available content and public user profile metadata in accordance with the source dataset’s terms and platform policies. Because public profile metadata may still contain information

that identifies individual users, we do not attempt to infer or reconstruct individual identities, and we do not release raw identifiable profile information or tweet text. The released artifacts will contain only permitted tweet/user/claim IDs and derived labels, allowing rehydration only under the corresponding platform policies.

Applications and Misuse. User susceptibility modeling is intended as a defensive tool for misinformation mitigation. One application is safer recommendation: susceptibility scores can serve as risk-aware signals alongside user interest and content quality. Another application is personalized debunking and educational intervention (Lo et al., 2022; Wang et al., 2022): for users predicted to be vulnerable to specific misinformation topics or claim types, systems may prioritize fact-checking resources, pre-debunking messages, or trustworthy contextual information, thereby strengthening long-term resilience. These applications support protective interventions rather than restrict users’ autonomy or access to information.

Although our work is motivated by defensive applications, susceptibility modeling could also be misused for targeted manipulation. To reduce this risk, following prior studies (Liu et al., 2024; Ma et al., 2024), we will restrict model access to approved users under controlled settings by application review. These safeguards should also be complemented by platform-level governance, as ethical risks primarily arise from deployment context rather than the modeling technique itself.

Discrimination. A possible concern is that susceptibility modeling may reinforce discrimination on certain user groups. We emphasize that susceptibility is not a fixed personal attribute, but a contextual interaction between users and misinformation claims: users may be vulnerable to some claims while remaining resistant to others. Nevertheless, because predictions may correlate with demographic or behavioral attributes, the susceptibility modeling should only support protective interventions, such as downranking high-risk misinformation exposures or providing contextual warnings, and should be accompanied by transparency and user autonomy safeguards.

References

Shahmir H Ali, Joshua Foreman, Ariadna Capasso, Abbey M Jones, Yesim Tozan, and Ralph J Di-

- Clemente. 2020. Social media as a recruitment platform for a nationwide online survey of COVID-19 knowledge, beliefs, and practices in the United States: methodology and feasibility analysis. *BMC medical research methodology*, 20:1–11.
- Antonio A Arechar, Jennifer Allen, Adam J Berinsky, Rocky Cole, Ziv Epstein, Kiran Garimella, Andrew Gully, Jackson G Lu, Robert M Ross, Michael N Stagnaro, Yunhao Zhang, Gordon Pennycook, and David G Rand. 2023. Understanding and combatting misinformation across 16 countries on six continents. *Nature Human Behaviour*, 7(9):1502–1513.
- Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.
- Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone. 1984. *Classification and Regression Trees*. Wadsworth International Group.
- Canyu Chen and Kai Shu. 2024. Combating misinformation in the age of LLMs: Opportunities and challenges. *AI Mag.*, 45(3):354–368.
- Jie Chen, Haw-ren Fang, and Yousef Saad. 2009. Fast approximate knn graph construction for high dimensional data via recursive lanczos bisection. *J. Mach. Learn. Res.*, 10:1989–2012.
- Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023. CoMPosT: Characterizing and evaluating caricature in LLM simulations. In *EMNLP*, pages 10853–10875.
- Chroma Team. 2022. Chroma. <https://github.com/chroma-core/chroma>. Accessed: Jan 4, 2026.
- Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. 2022. Misinfo reaction frames: Reasoning about readers’ reactions to news headlines. In *ACL*, pages 3108–3127.
- Saadia Gabriel, Liang Lyu, James Siderius, Marzyeh Ghassemi, Jacob Andreas, and Asuman E. Ozdaglar. 2024. MisinfoEval: Generative AI in the era of “alternative facts”. In *EMNLP*, pages 8566–8578.
- Hulingxiao He, Xiangteng He, Yuxin Peng, Zifei Shan, and Xin Su. 2024. Firzen: Firing strict cold-start items with frozen heterogeneous and homogeneous graphs for recommendation. In *ICDE*, pages 4657–4670.
- Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020a. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9726–9735.
- Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yong-Dong Zhang, and Meng Wang. 2020b. LightGCN: Simplifying and powering graph convolution network for recommendation. In *SIGIR*, pages 639–648.
- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.
- Yanchen Liu, Mingyu Derek Ma, Wenna Qin, Azure Zhou, Jiaao Chen, Weiyan Shi, Wei Wang, and Diyi Yang. 2024. Decoding susceptibility: Modeling misbelief to misinformation through a computational approach. In *EMNLP*, pages 15178–15194.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.
- Kuan-Chieh Lo, Shih-Chieh Dai, Aiping Xiong, Jing Jiang, and Lun-Wei Ku. 2022. VICTOR: an implicit approach to mitigate misinformation via continuous verification reading. In *WWW*, pages 3511–3519.
- Weicheng Ma, Chunyuan Deng, Aram Moossavi, Lili Wang, Soroush Vosoughi, and Diyi Yang. 2024. Simulated misinformation susceptibility (SMISTS): Enhancing misinformation research with large language model simulations. In *Findings of ACL*, pages 2774–2788.
- Meta. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.
- Dan Saattrup Nielsen and Ryan McConville. 2022. MuMiN: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In *SIGIR*, pages 3141–3153.
- OpenAI. 2024. GPT-4o system card. *CoRR*, abs/2410.21276.
- 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.
- Manuel Pratelli and Marinella Petrocchi. 2025. Evaluating the simulation of human personality-driven susceptibility to misinformation with llms. *CoRR*, abs/2506.23610.
- Jon Roozenbeek, Sander Van Der Linden, Beth Goldberg, Steve Rathje, and Stephan Lewandowsky. 2022. Psychological inoculation improves resilience against misinformation on social media. *Science advances*, 8(34):eabo6254.
- Zhiqi Shen, Shaojing Fan, Danni Xu, Terence Sim, and Mohan Kankanhalli. 2025. Modeling human responses to multimodal AI content. *CoRR*, abs/2508.10769.
- Tobia Spampatti, Ulf JJ Hahnel, Evelina Trutnevyte, and Tobias Brosch. 2024. Psychological inoculation strategies to fight climate disinformation across 12 countries. *Nature Human Behaviour*, 8(2):380–398.

Xian Teng, Yu-Ru Lin, Wen-Ting Chung, Ang Li, and Adriana Kovashka. 2022. Characterizing user susceptibility to COVID-19 misinformation on Twitter. In *ICWSM*, volume 16, pages 1005–1016.

Noah Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhang Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of ACL*, pages 14743–14777.

Shoujin Wang, Xiaofei Xu, Xiuzhen Zhang, Yan Wang, and Wenzhuo Song. 2022. Veracity-aware and event-driven personalized news recommendation for fake news mitigation. In *WWW*, pages 3673–3684.

Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural graph collaborative filtering. In *SIGIR*, pages 165–174.

Guangyu Wei, Ke Han, Yueming Lyu, Yu Luo, Yue Jiang, Caifeng Shan, and Nicu Sebe. 2025. Towards real-time fake news detection under evidence scarcity. *CoRR*, abs/2510.11277.

Wei Wei, Chao Huang, Lianghao Xia, and Chuxu Zhang. 2023. Multi-modal self-supervised learning for recommendation. In *WWW*, pages 790–800.

Wei Wei, Xubin Ren, Jiabin Tang, Qinyong Wang, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. 2024. LLMRec: Large language models with graph augmentation for recommendation. In *WSDM*, pages 806–815.

Shanshan Zhong, Zhongzhan Huang, Daifeng Li, Wushao Wen, Jinghui Qin, and Liang Lin. 2024. Mirror gradient: Towards robust multimodal recommender systems via exploring flat local minima. In *WWW*, pages 3700–3711.

Hongyu Zhou, Xin Zhou, Lingzi Zhang, and Zhiqi Shen. 2023. Enhancing dyadic relations with homogeneous graphs for multimodal recommendation. In *ECAI*, volume 372 of *Frontiers in Artificial Intelligence and Applications*, pages 3123–3130.

Xin Zhou. 2023. MMRec: Simplifying multimodal recommendation. In *Proceedings of the 5th ACM International Conference on Multimedia in Asia Workshops*, pages 1–2.

A Definition of Diversity Score

Conventional classification metrics primarily reflect overall prediction accuracy, but they fail to characterize the distribution of model predictions, such as whether predictions are concentrated on a single class or evenly distributed across classes. To capture this aspect, we define a Diversity score based on the Gini impurity of the predicted label

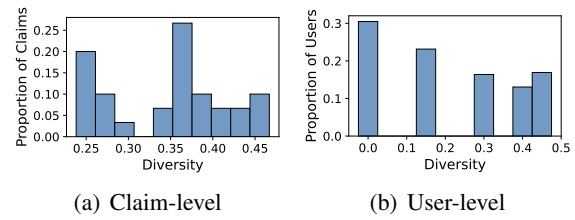


Figure 9: Diversity distribution of users' truthfulness ratings for misinformation claims.

distribution. Gini impurity measures the degree of class heterogeneity in a distribution (Breiman et al., 1984). A lower Gini value indicates a more homogeneous label distribution. In binary classification, the minimum Gini value $G = 0$ corresponds to all instances being assigned to the same class, whereas the maximum value $G = 0.5$ indicates a perfectly balanced distribution between the two classes. In our datasets, the ground-truth label diversity, measured by Gini impurity, is 0.40 for the COVID dataset and 0.48 for the Climate dataset.

B Preliminary Analysis about Cognitive Personalization

To examine whether cognitive personalization exists in users' responses to misinformation, we analyze the diversity of truthfulness ratings assigned by different users to different misinformation claims in the COVID dataset.

We first compute the *claim-level* diversity distribution, as shown in Figure 9(a). Specifically, for each claim, we aggregate truthfulness ratings from different users and compute the corresponding diversity score. We observe that the diversity scores of all claims are non-zero, indicating that different users assign different ratings to the same claim, and that no claim is universally perceived as either true or false. The diversity scores vary across claims, ranging from 0.2 to 0.5, suggesting that the difficulty of identifying misinformation differs across claims. Claims that are easier to recognize as misinformation tend to exhibit higher rating consistency across users and thus lower diversity scores.

Similarly, Figure 9(b) shows the *user-level* diversity distribution. We find that approximately 30% of users have a diversity score of 0, indicating fully consistent ratings across all claims. This suggests that these users exhibit either consistently strong discernment (all correct judgments) or consistently weak discernment (all incorrect judgments). For the remaining users with diversity greater than

Dataset	#Interactions			#Users					#Claims					Sparsity	Pos/Neg ratio
	All	Train	Test	All	Train/Seen	Test	Test/Seen	Test/Unseen	All	Train/Seen	Test	Test/Seen	Test/Unseen		
COVID	34,565	29,365	5,200	4,527	3,688	2,754	1,915	839	30	25	30	25	5	75%	1:3.2
Climate	48,932	41,612	7,320	6,816	6,206	3,441	2,831	610	10	7	10	7	3	28%	1.54:1
MuMiN-Sus	5,960	4,404	1,556	3,318	2,759	1,034	475	559	2,247	1,907	808	468	340	99%	1.47:1

Table 3: Dataset statistics. We report the number of user-claim interactions, users, and claims across training and test splits. Sparsity refers to the proportion of missing (unobserved) user-claim interactions relative to the total number of possible user-claim pairs. Pos/Neg ratio indicates the distribution of positive (“susceptible”) versus negative (“not susceptible”) labels.

zero, their ability to identify misinformation varies across claims.

Overall, these observations provide evidence of cognitive differentiation in misinformation perception, which forms a foundation for studying personalized misinformation susceptibility prediction.

C Four Test Settings

We define four test settings to represent different realistic scenarios as follows:

- **test_{ss}** (seen user, seen claim): Both the involved users and claims appear during training, but the test samples consist of new combinations of them. This is the simplest evaluation scenario, as the model can leverage users’ and claims’ historical behavioral patterns in addition to content semantics to capture user preferences. For example, if the training data contain interactions such as user u with misinformation claims i_1 , i_2 , and i_3 , and misinformation claim i with users u_1 , u_2 , and u_3 , then the model can infer both which types of claims user u is likely to be deceived by and which types of users are likely to be deceived by claim i , thereby facilitating the test task of determining whether user u would be deceived by claim i .
- **test_{su}** (seen user, unseen claim): For a user-claim pair (u, i) in test_{su} to be classified, the training set contains labeled examples involving user u with other claims but none for claim i . The model must generalize user susceptibility patterns to unseen misinformation claims.
- **test_{us}** (unseen user, seen claim): For a user-claim pair (u, i) in test_{us} to be classified, the training set contains labeled examples involving claim i with other users but none for user u . The model must generalize claim susceptibility patterns to unseen users.
- **test_{uu}** (unseen user, unseen claim): Neither the users nor the claims appeared during training,

making this the most challenging scenario. The model must rely solely on semantic cues from user profile and claim content to predict susceptibility (*e.g.*, doctors tend not to be susceptible to medical misinformation), rather than behavioral patterns learned from training data as in test_{ss}.

D Datasets

As shown in Table 3, we construct multi-domain datasets covering COVID, Climate, and general-domain misinformation from two sources. The COVID and Climate survey datasets provide dense user-claim interactions, with tens of thousands of annotations over a small set of claims (30 COVID claims and 10 Climate claims); by design, each user annotates a fixed number of claims and each claim receives a fixed number of annotations. In contrast, the social-platform dataset MuMiN-Sus contains substantially more (thousands) claims but only thousands of total interactions, resulting in a highly sparse user-claim matrix. This sparsity reflects realistic platform settings, where users naturally engage with only a tiny fraction of the misinformation claims circulating online. For evaluation, we apply downsampling to equalize test subset sizes.

COVID (Arechar et al., 2023): This dataset was collected via large-scale surveys conducted across 16 countries on six continents, with recruited participants quota-matched to the national distributions of age and gender. Each participant was asked to rate the truthfulness of 10 true and 10 false news headlines, which are randomly sampled from a larger pool of 45 headlines (30 false), using a six-point Likert scale. These headlines were collected from reputable sources, including the World Health Organization and fact-checking websites from multiple countries. Participants also reported demographic information (*e.g.*, country, age, and education), news-sharing habits (*e.g.*, preferred news domain and factors considered most important when evaluating news), cognitive style (*e.g.*, reflective

vs. intuitive), and personal beliefs (*e.g.*, toward vaccines and democracy). In our experiments, we retain only responses to misinformation from participants who passed two rounds of attention checks to ensure high data quality. The original six-point truthfulness scale, ranging from “completely inaccurate” to “completely accurate”, is mapped to binary labels of “not susceptible” and “susceptible” for clarity.

Climate (Spampatti et al., 2024): Participants in this dataset were recruited using a procedure similar to that employed for the COVID dataset. Each participant reported personal information, including demographics, cognitive style, and affect toward climate action, and was asked to classify 20 climate-related claims (10 true and 10 false) as either “true” or “false”. These statements were initially generated by ChatGPT through summarizing from an initial pool of 300 candidate statements, and were manually verified. As with the COVID dataset, we retain only responses to misinformation and map the judgments to “not susceptible” and “susceptible”, respectively.

MuMiN-Sus: The proposed MuMiN-Sus dataset is derived from the MuMiN dataset (Nielsen and McConville, 2022), a multilingual and multimodal social network dataset built by linking fact-checked claims across multiple topics with social-media discussions and users. MuMiN first uses the Google Fact Check Tools API to collect fact-checked claims from a large set of fact-checking organizations across languages. Because the original verdicts are unstructured free text, MuMiN normalizes them into three categories: misinformation, factual, and other. For each claim, MuMiN extracts claim keyphrases and queries X/Twitter to collect related source tweets, applying filters such as excluding tweets with insufficient retweet activity. The surrounding Twitter threads and user information are then collected from these source tweets. To ensure that the collected social-media discussions correspond to the target fact-checked claims, candidate tweet/thread content is semantically linked to claims using sentence embeddings and similarity matching. Due to X/Twitter platform restrictions, the released dataset provides identifiers such as claim IDs and tweet IDs rather than all original social content. We therefore hydrate the relevant records ourselves; some claims, tweets, and users cannot be recovered because they have been deleted or made unavailable.

We construct the susceptibility-labeled MuMiN-Sus dataset following the pipeline shown in Figure 3. We first retain claim IDs whose fact-checking verdicts indicate misinformation and retrieve their claim content through the Google Fact Check Tools API. We then crawl the corresponding tweets and annotate each claim–tweet pair for stance toward the claim. Specifically, we use GPT-5.5 for initial stance annotation, followed by manual verification to reduce labeling noise. The stance labels include belief, disbelief, and uncertainty. We map belief to “susceptible” and disbelief to “not susceptible”, while removing uncertain pairs. To reduce annotation noise, our explicitly asks GPT whether the tweet author appears to believe the claim. Sarcastic, discussion, neutral, or otherwise ambiguous tweets are labeled uncertainty and excluded. Two annotators independently verified a stratified sample of 300 claim–tweet pairs covering all three stance categories. The annotators achieved a Cohen’s κ of 0.83, indicating substantial agreement. After GPT-based pre-annotation and manual verification, the stance label distribution is 4,943 belief pairs, 6,905 disbelief pairs, and 13,908 uncertain pairs. Thus, more than half of the claim–tweet pairs are removed due to ambiguous or insufficient stance signals. This conservative filtering strategy sacrifices coverage but improves label reliability by retaining only pairs with clear endorsement or rejection of the misinformation claim.

Next, we collect public profiles for the remaining tweet authors, usually including username, description, verification type, location, follower/following counts, and tweet counts. Among the recovered users, fact-checker-related accounts account for only about 3% of all users, but contribute approximately 46% of not-susceptible claim–user pairs. This disproportionate contribution is expected because fact-checkers frequently participate in discussions of popular misinformation claims to debunk them. However, these accounts have distinctive profile patterns and are overwhelmingly not susceptible, which may introduce shortcut bias into susceptibility prediction. Since our goal is to model susceptibility among ordinary users rather than professional fact-checking or debunking accounts, we filter these accounts from the final dataset.

E Prompt for LLM Simulation

We employ three prompting strategies, including zero-shot, direct few-shot, and story-based few-

shot, to guide LLMs in imitating a given user when evaluating the authenticity of a misinformation claim. The prompts are shown below, with minor variations introduced to accommodate the specific individual attributes available in each dataset.

For few-shot prompting, we provide nine examples. During example selection, we prioritize labeled examples involving the same user and/or same claim as the target sample, and select the remaining examples based on high semantic similarity. Similarity is measured via cosine similarity of OpenAI’s text-embedding-3 representations of the concatenated text consisting of the user profile and claim content, with the embedding database built using Chroma (Chroma Team, 2022). As illustrated in the prompts below, different examples are accompanied by different instructions to guide the model’s behavior.

System Prompt in Zero-shot Setting

Let’s roleplay. I will provide you with a persona’s basic information, habits of social media usage, and other personal beliefs. Then, I will present a claim with unknown authenticity. You need to think as the given persona and give this persona’s judgment on the given claim’s accuracy (0=False, 1=True) and explain your reasoning. Please make sure to simulate the thinking pattern of this persona, rather than the objectively correct reasoning process. Please respond in the following format: [0=False/1=True] Reasons: ...

Profile: {profile}
Claim: {claim}

System Prompt in Direct Few-shot Setting

Let’s roleplay. I will provide you with a persona’s basic information, habits of social media usage, and other personal beliefs. Then, I will present a claim with unknown authenticity. You need to think as the given persona and give this persona’s judgment on the given claim’s accuracy (0=False, 1=True) and explain your reasoning. Please make sure to simulate the thinking pattern of this persona, rather than the objectively correct reasoning process. Please respond in the following format: [0=False/1=True] Reasons: ...

Profile: {profile}
Claim: {claim}

Here are some judgments of this persona on other claims for your reference. Please refer to this persona’s behavior pattern. {examples_sameuser}

Here are other users’ judgments on this claim for your reference. Please refer to the behavior pattern of the user that is most similar to you. {examples_sameclaim}

Here are other users’ judgments on other claims for your

reference. Please refer to the cases with users and claims that are most similar to yours. {examples_similar}

System Prompt in Story-based Few-shot Setting

Step 1: User Story Generation

Based on the provided persona profile and their historical judgments on COVID-related misinformation, please generate a coherent and realistic life experience narrative that captures the persona’s cognitive understanding and reasoning patterns toward misinformation.

Profile: {profile}

Historical judgments: {examples_sameuser}

Step 2: User Response Generation

Life story: {gen_story}

Claim: {claim}

Let’s roleplay. Suppose you are this persona with the life experience described above. You need to think as the given persona and give this persona’s judgment on the given claim’s accuracy (0=False, 1=True) and explain your reasoning. Please make sure to simulate the thinking pattern of this persona, rather than the objectively correct reasoning process. Please respond in the following format: [0=False/1=True] Reasons: ...

Here are the ratings from individuals with profiles similar to yours for this claim or a similar claim. Since your judgments are generally aligned with theirs, you are encouraged to pay closer attention to their evaluations: {examples_exclude_sameuser}

F Implementation Details

To obtain task-specific semantic features, we fine-tune LLaMA-2 with a linear classification head via LoRA, where the backbone is frozen and only the LoRA adapters and classification head are trained. Specifically, LoRA adapters are applied to the query and value projection matrices of the self-attention layers with a fixed rank $r = 8$. After fine-tuning, we extract token-level hidden states from the final transformer layer as textual representations. During graph aggregation, node features are projected to a d -dimensional latent space ($d = 128$) and kept learnable throughout the training process. Our implementation is based on the MMRec library (Zhou, 2023) and built with PyTorch. For optimization, we used the Adam optimizer (Kingma and Ba, 2015) and applied the mirror gradient strategy (Zhong et al., 2024) to enhance model robustness. We conducted a grid search to determine optimal hyperparameters for different datasets. The best learning rates were found to be e^{-4} (COVID) and e^{-5} (Climate, MuMiN-Sus). To sparsify the seman-

tic graph, the number of nearest neighbors was set to 10 and 5. Other hyperparameters were tuned as follows: $\lambda = 0.35/0.17$, $\lambda_{cl} = 0.01$, $\lambda_{adv} = e^{-4}$, and $\lambda_{reg} = e^{-4}$. The number of layers for the heterogeneous graph and homogeneous graph networks was set to 2 and 3, respectively. Since the MuMiN-Sus datasets contains multilingual claims and user profiles, we use Google Translator API to obtain the English versions instead of sole rely on multilingual encoders following (Nielsen and McConville, 2022).

G Optimization Details

To mitigate the absence of behavioral signals for unseen nodes, we employ self-supervised learning (Wei et al., 2023) on seen nodes, including contrastive learning and adversarial training, to regularize semantic representations and enhance their behavior-awareness.

Contrastive Learning. To encourage semantic representations to capture behavior-relevant information, we align the semantic embeddings (*i.e.*, $\mathbf{f}_u$ and $\mathbf{f}_i$) with the final features (*i.e.*, $\mathbf{h}_u$ and $\mathbf{h}_i$) that integrate both semantic and behavioral information. Specifically, we employ the InfoNCE loss (He et al., 2020a) to maximize their mutual information:

$$\mathcal{L}_{cl} = -\frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \log \frac{\exp(\text{sim}(\mathbf{f}_u, \mathbf{h}_u))}{\sum_{u' \in \mathcal{U}} (\exp(\text{sim}(\mathbf{f}_u, \mathbf{h}_{u'}) + \exp(\text{sim}(\mathbf{f}_{u'}, \mathbf{h}_u))} \\ - \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \log \frac{\exp(\text{sim}(\mathbf{f}_i, \mathbf{h}_i))}{\sum_{i' \in \mathcal{I}} (\exp(\text{sim}(\mathbf{f}_i, \mathbf{h}_{i'}) + \exp(\text{sim}(\mathbf{f}_{i'}, \mathbf{h}_i))}. \quad (4)$$

Adversarial Training. We further employ adversarial training to align semantic features with final representations at the interaction-graph level. An auxiliary discriminator $\mathcal{D}$ distinguishes the semantic-related interaction graph $\mathbf{G}_{sem}$ and the behavior-related interaction graph $\mathbf{G}_{beh}$. The former is constructed from semantic embeddings $\mathbf{f}_u$ and $\mathbf{f}_i$ via cosine similarity. The latter is derived from the observed interaction graph $\mathbf{G}_{ori}$ and augmented with a virtual interaction graph constructed from $\mathbf{h}_u$ and $\mathbf{h}_i$. Following Wei et al. (2023), we densify the observed graph using Gumbel-Softmax:

$$\mathbf{G}_{beh} = \tilde{\mathbf{G}}_{ori} + \lambda \frac{\mathbf{h}_u \cdot \mathbf{h}_i^T}{\|\mathbf{h}_u\|_2 \cdot \|\mathbf{h}_i\|_2}, \quad (5)$$

where $\tilde{\mathbf{G}}_{ori}$ is the densified behavioral graph, and λ is a weight hyperparameter. The generator and discriminator are optimized as:

$$\mathcal{L}_G = -\mathbb{E}[\mathcal{D}(\mathbf{G}_{sem})], \quad \mathcal{L}_D = -\mathbb{E}[\mathcal{D}(\mathbf{G}_{beh})] + \mathbb{E}[\mathcal{D}(\mathbf{G}_{sem})]. \quad (6)$$

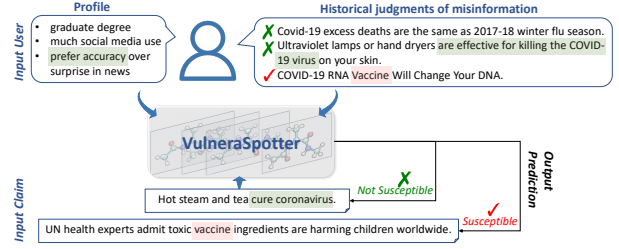


Figure 10: An illustrative example of personalized susceptibility prediction by VulneraSpotter. ✓ indicates belief (susceptible) for the misinformation, whereas ✗ indicates disbelief (not susceptible).

H Case Study

To intuitively illustrate the capability of our VulneraSpotter model in personalized susceptibility modeling, we present a representative example in Figure 10. For the illustrated user, the profile attributes suggest a certain ability to discern misinformation. However, the user’s historical responses to previously encountered misinformation reveal a tendency to believe claims emphasizing vaccine-related risks, indicating a potential anti-vaccine inclination. By jointly considering the semantics of the user profile and claim content, as well as historical judgment behaviors, our model correctly predicts that the user is likely to be deceived by the second claim (vaccine-related misinformation), while exhibiting resistance to the first claim (misinformation about COVID-19 prevention methods).

I GenAI Usage Disclosure

This work involved the use of generative AI tools (here, ChatGPT) in two aspects of the research process. First, ChatGPT was used to refine the phrasing of some sentences and identify grammatical issues in the manuscript. All AI-assisted edits were carefully reviewed by the authors to ensure the correctness and fidelity of the intended meaning. Second, ChatGPT was employed to assist in drafting parts of the data processing codes, which were carefully reviewed and validated by the authors to ensure logical correctness. Generative AI tools were not used to generate scientific content, analyses, experimental results, or methodological designs. No confidential or proprietary data was provided to the generative AI tools.