Exploring and Distilling Multi-Dimensional Clues for Interpretable Social Bot Detection

Yi Han[1*], **Haiqi Lu**[1], **Lizi Liao**[3], **Shuhan Zhou**[1],
Yuanxing Liu[1], **Weinan Zhang**[1,2†], **Ting Liu**[1]
[1]Research Center for Social Computing and Interactive Robotics
Harbin Institute of Technology, China
[2]Suzhou Research Institute, Harbin Institute of Technology
[3]Singapore Management University, Singapore
{yihan, hqlu, shzhou, yxliu, wnzhang, tliu}@ir.hit.edu.cn, lzliao@smu.edu.sg

Abstract

Social bot accounts have long been disseminating disinformation and engaging in malicious activities on social media platforms. Detecting these social bots has become a critical and urgent task, essential for maintaining a healthy online ecosystem. Existing social bot detection research usually provides detection results directly without corresponding supportive explanations, making it difficult to assess the extent to which such predictions are trustworthy. This is a key concern for online moderation. In this work, we explore the detection interpretation and summarize a four-dimensional clue framework from individual and social perspectives. We propose CDRBot, which primarily employs outcome-reward reinforcement learning to train inspectors to generate faithful, grounded, and readable clues from the *User Information*, *Semantic Features*, *Interactive Situation*, and *Behavioral Pattern*. These clues are then integrated to make final predictions. Experimental results demonstrate that our approach outperforms other baselines in detection performance. The generated clues are faithful, grounded, and readable, and can significantly enhance the performance of large language models in social bot detection. The code is available at: https://github.com/HITSCIR-DT-Code/CDRBot.

1 Introduction

Social bot accounts driven by automated programs are designed to imitate human users on online social media platforms (Wischnewski et al., 2024). It has been revealed that social bots are widely involved in various malicious activities, including disseminating disinformation (Pozzar et al., 2020; Huang et al., 2022; Mishra et al., 2025), fueling extremist movements (Salles et al., 2024; Wan et al., 2025), and manipulating public opinion (Akhtar et al., 2024; Mendoza et al., 2024). These findings indicate that the risk posed by social bots threatening the health of online communities is intensifying (Makovi et al., 2023). Therefore, research on social bot detection holds significant importance for the preservation of the online information ecosystem.

Existing research primarily detects social bots by leveraging heterogeneous information (Metadata, Activities, and Social Relationships) of social accounts (Echeverrï£¡a et al., 2018; Yang et al., 2020). Some methods rely on heuristic rules to extract specific features and identify social bots with machine learning models (Mazza et al., 2019; Hayawi et al., 2022). Another class of research utilizes the graph neural network (GNN) to encode accounts and social networks (Wang et al., 2021; Feng et al., 2022c,a). Despite these advances, a key limitation remains: trustworthy detection requires explanations that are grounded in observable evidence (Guillaro et al., 2023). Most detectors are trained to output labels directly, and their decision process is a black box. This raises concerns about the extent to which the judgments made by black-box models on accounts can be trusted (Huang et al., 2024; Shridhar et al., 2024). For high-stakes moderation and account review, reviewers need to understand why an account is flagged (and what evidence supports that judgment) to conduct follow-up investigations, and resolve disputes (Grimmelikhuijsen, 2023). Put differently, *Prediction alone is insufficient. A detection system should provide interpretable, verifiable, and concise clues that bridge raw data and final predictions.*

We address this gap by reframing bot detection as an evidential reasoning problem. Instead of directly targeting the label, we explicitly extract intermediate, supportive, and human-readable clues that summarize suspicious patterns while remaining grounded in the raw information. However, there is currently no systematic framework for formalizing such clues. To address this, building on established

[*]Work was done during an internship at SMU.
[†]Corresponding author.

findings that bots differ from humans along multiple axes and atomic features (Varol et al., 2017; Ng and Carley, 2025), we introduce a structured four-dimension clue framework (Figure 1): *User Information* (profile and account-level signals), *Semantic Features* (linguistic and topical properties of posted content), *Interactive Situation* (the social contexts, targets, and communities the account engages with), and *Behavioral Pattern* (regularities in activity and interaction). This framework offers a compact interface for both model reasoning and human review, and avoids requiring the exhaustive reporting of thousands of low-level atomic features.

Based on this framework, we distill and filter a batch of high-quality clues for cold-start. Then, we propose CDRBot, a two-stage training framework that learns to distill clues and predict labels (Figure 4). First, we train dimension-specific clue inspectors with reinforcement learning (RL) that map raw account sequences to clue texts. The outcome-based reward comprises a logit-based label correctness signal derived from an external evaluator model and a format control to discourage overly short "shortcut" clues. Since clues generated by inspectors may contain noise or bias, CDRBot trains a fusion predictor that consumes four clues jointly and outputs the final account label, allowing the system to balance evidence across dimensions and reduce the impact of any single noisy clue.

Our experiments demonstrate that CDRBot improves detection performance over various baselines, while producing clue texts that score well on faithfulness, grounding, and fluency. Experiments indicate that clues across different dimensions are almost distinct, validating the rationale of our proposed clue framework. Additional analysis reveals that these clues, as external knowledge, can significantly boost the detection performances of other LLMs. Furthermore, we find that the generated individual-level clues are more explicit and reasonable, outperforming the social-level clues slightly. Nonetheless, the relatively abstract social-level clues still encode meaningful signals that can be further extracted and utilized by reasoning LLMs.

In summary, our contributions are threefold:

- We formalize a multi-dimensional framework for evidential social bot detection, connecting prior feature-discrepancy findings to an interpretable reasoning interface.
- We introduce CDRBot, a two-stage training framework that optimizes clue generation via reinforcement learning and fuses multi-dimensional clues for account prediction.
- We provide empirical results on two benchmarks and evaluate clue quality with automatic and human assessments, demonstrating that the model's intermediate evidence is largely grounded and useful for interpretability.

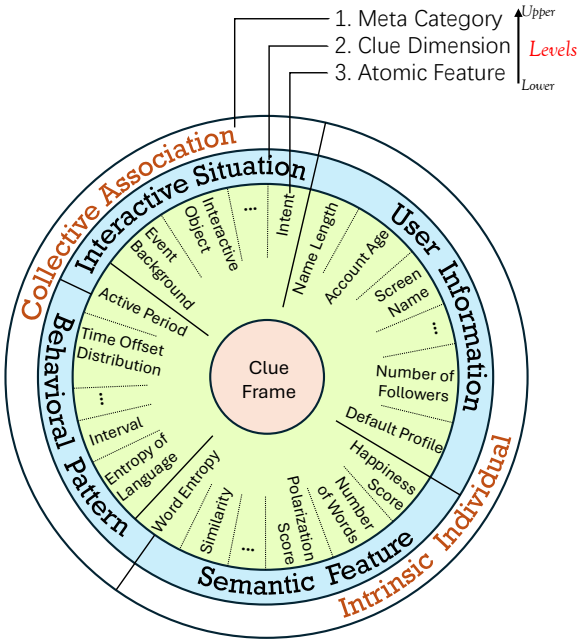

Figure 1: The proposed logic-driven clue framework refines the atomic features (core) into higher abstract dimensions (crust).

2 Clue Framework

2.1 Background

Previous research has extensively analyzed the behavioral characteristics of social bot accounts on online social platforms. Through analyzing bot accounts and genuine user accounts across multiple countries on Twitter, Mazza et al. (2022) discover significant differences between bots and humans in content generation, interaction patterns, and adaptability. For instance, bot accounts often exhibit incomplete or false profile information and show a stronger tendency to retweet rather than create original content. The analysis of Cai et al. (2022) about social bot behavioral patterns during the COVID-19 outbreak reveals that social bots tended to disseminate high-impact controversial topics through meaningless retweets. The proportion of positive and negative emotional content is relatively balanced, indicating that emotional effect is not their primary motivation but rather a strategy to inter-

vene in public opinion. A typical study by Varol et al. (2017) proposed a holistic framework comprising more than a thousand features across six categories to comprehensively distinguish social bot account characteristics.

Existing findings appear to have laid a solid foundation for formalizing the detection interpretation, suggesting that generating an interpretation directly based on these atomic features might be feasible. However, analyzing and reporting such fine-grained features seems daunting due to their sheer quantity. Moreover, a single feature is usually insufficient and these atomic features contain extensive interconnections (e.g., *average emoji entropy per hour of tweets* both relate to *emoji entropy of a single tweet* and *number of tweets per hour*), with complex and obscure correlation that make clue directly based on these features challenging to obtain and interpret.

2.2 Formalization

In this work, we contend that constructing an analytical framework does not necessitate an exhaustive focus on every individual feature. This perspective aligns with Varol et al. (2017) and Pathak et al. (2021), which reveals that most differences in features are vague and difficult to explain, and only a small subset of features exhibit high statistical significance. Another critical consideration is readability and practical applicability. The forms of clue and detection in this study are inspired by the prosecutorial and adjudicative system in judicial processes. For the sake of procedural and outcome rationality, the clue presented to downstream agents must be sufficiently succinct and justified, containing only essential and supportive information.

Inspired by the work of Kiesel et al. (2022), we integrate and summarize discoveries from previous feature-driven studies (Varol et al., 2017; Ng and Carley, 2025), and propose a more refined taxonomy (Figure 1) comprising four high-level analytical dimensions: *User Information*, *Semantic Feature*, *Interactive Situation*, and *Behavioral Pattern*. These dimensions consolidate and abstract the myriad of atomic features identified in prior work, offering a more macroscopic, structured, and comprehensible framework.

Crucially, by organizing features into these coherent dimensions, the framework operates at a higher level of abstraction. Therefore, based on this framework, we do not need to examine each atomic feature in detail. Instead, we can directly extract and reason about the abstract insights encapsulated within each dimension to support downstream detection tasks. Moreover, these dimensions reflect both mutual independence and relatedness, and this relatedness reveals two "higher-order" oppositions: *Intrinsic Individual* and *Collective Association*. This hierarchical taxonomy helps to efficiently disentangle intricate relationships within complex entities (Kiesel et al., 2022; Sap et al., 2020; Ziems et al., 2022). Therefore, in this study, we formalize the supportive clue based on this foundational framework.

Model	Acc.(Raw)	Acc.(Clue)	F → F	F → T	T → F	T → T
R1(Clue Source)	63%	100%	-	-	-	-
Qwen2.5-1.5B	58.41%	99.44%	0.47%	**41.12%**	0.09%	58.32%
Qwen2.5-3B	57.71%	94.21%	5.37%	**36.93%**	0.42%	57.28%
Qwen2.5-7B	58.87%	98.61%	0.42%	**40.70%**	0.97%	57.90%
R1-Distill-1.5B	45.10%	76.80%	11.57%	**43.33%**	11.63%	33.48%
R1-Distill-7B	51.76%	94.45%	2.07%	**46.17%**	3.49%	48.27%

Table 1: The faithfulness of the distilled clues. All models are the instruct version. Acc. evaluates the prediction accuracy when presenting the LLM with the raw sequence or clue. $F \rightarrow T$ represents the percentage that LLM outputs F (false) predictions with raw information but T (true) predictions with clues, and other types of transitions are defined similarly.

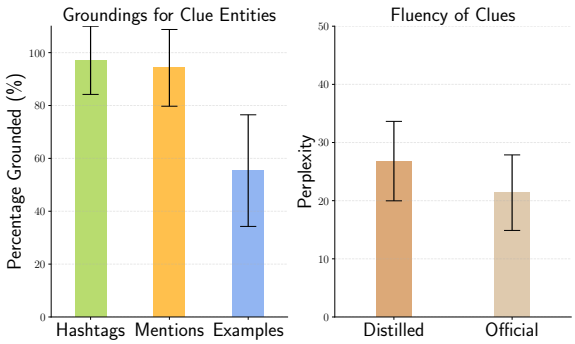

Figure 2: The grounding and fluency evaluations of the distilled clues. The horizontal lines represent the mean and standard deviation, respectively.

2.3 Clue Distillation and Evaluation

To instantiate a batch of initial clues, we sample training data from the Twibot-20 (Feng et al., 2021b). Following Feng et al. (2024), we fuse the heterogeneous account information (metadata, historical activities, social relationships) of each account and concatenate them into a text sequence using templates as the raw account sequence (briefly shown as in Table 17). Leveraging an advanced large reasoning model, Deepseek-R1 (DeepSeek-AI et al., 2025), and following our clue framework,

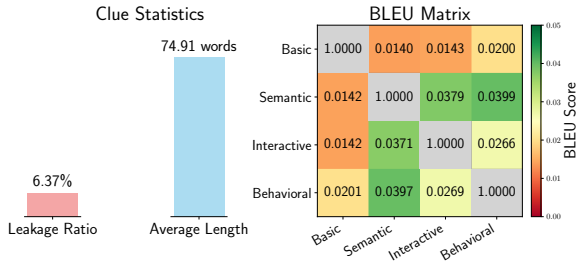

Figure 3: The label leakage ratio and average word length of the clues (left), and the mutual BLEU scores among different dimensions of clues (right).

we conduct post-hoc analysis on raw account sequences and corresponding ground-truth labels to reason and summarize the clue texts for each dimension (§A.1), resulting in a batch of initial distilled clues (Figure 4(a)).

We evaluate such distilled clues in terms of *Predictive Faithfulness*, *Grounding*, *Fluency*, *Leakage*, and *Distinction*. *Predictive Faithfulness* measures whether the distilled clue c preserves the label-relevant evidence in the raw account u. A clue is considered faithful if an LLM can infer the correct category from the clue alone[1]. To reduce the bias introduced by generative decoding, we directly obtain the predicted label $\hat{y}$ from the logits of the last input token, which is:

$$\text{logits} = \text{LLM}(u \text{ or } c|P)[-1, :], \tag{1}$$

$$\hat{y} = \begin{cases} bot, & \text{if } \ell_{\text{bot}} > \ell_{\text{human}} \\ human, & \text{otherwise} \end{cases} \tag{2}$$

where P denotes the direct inference prompt, while ℓ_{bot} and ℓ_{human} represent the logits values corresponding to their respective tokens. Then we report and compare the overall accuracy.

Grounding evaluates the evidential basis of the clue text, to avoid hallucinations and ensure that the clue is strictly refined from the raw account information. Considering that clues should be deeply refined, we employ the regular expression to identify three types of entity reference within clues: Hashtag (#xxx), Mention (@xxx), and Example (e.g., xxx). Then we report the occurrence of these referenced entities in the raw account sequence. For *Fluency*, we report the perplexity based on Llama-3.2-1B [2] and compare it against an official U.S. court opinion [3].

Since in this stage the prediction prompt includes the true label, to assess whether the distilled clues genuinely reflect analytical reasoning rather than label leakage, we examine both the label leakage ratio and the average word length of the clues. Besides, we also report the BLEU score between different dimensions of clues to demonstrate the distinctions of each dimension. More details are shown in §A.3.

Table 1 indicates that, although there is potential bias in the clues distilled from R1, these clues demonstrate high *Faithfulness* across several LLMs, achieving high accuracy rates exceeding 95%. Compared with raw, it reveals that these clues can significantly increase the probability of correct predictions. For grounding, shown in Figure 2, the average grounding of the Hashtag entity and Mention entity reaches over 95%, indicating that most entities in clues originate from source information. We observe that clues are the deep refinement of original content, often including recapitulative words when exemplifying, resulting in a slight word-mismatch issue for the Example entity. Nevertheless, the grounding of the Example entity still approaches 60%. So these clues exhibit excellent grounding performance. Figure 2 also shows that the fluency of clues is close to the fluency of the official document. A low leakage ratio in Figure 3 indicates that the clues do not explicitly embed label-related information, while a longer average word length suggests that the clues are the expected deep condensation, rather than being template-based or shortcut expressions that directly hint at the labels. We observe that the BLEU scores among clues in different dimensions are extremely low (Figure 3), which confirms that these clues contain distinct information, and also validates the rationality of our analytical framework. Overall, these distilled clues demonstrate expected quality. To improve the robustness of latter training, we exclusively utilize clues from the $F \rightarrow T$ case and delete leaked ones for the LLM cold-start.

3 Method

Regarding distilled clues, we employ them solely for the cold-start, guiding LLMs to learn how to extract target dimensional information from raw account sequences (Figure 4 (c)). There is an LLM

[1]For brevity, we use *faithfulness* to refer to *predictive faithfulness* throughout the paper.

[2]To reduce the Self-Preference Bias (Wataoka et al., 2024), we introduce an LLM from a different family to compute the perplexity.

[3]https://law.justia.com/cases/federal/appellate-courts/ca4/17-4194/17-4194-2018-04-10.html

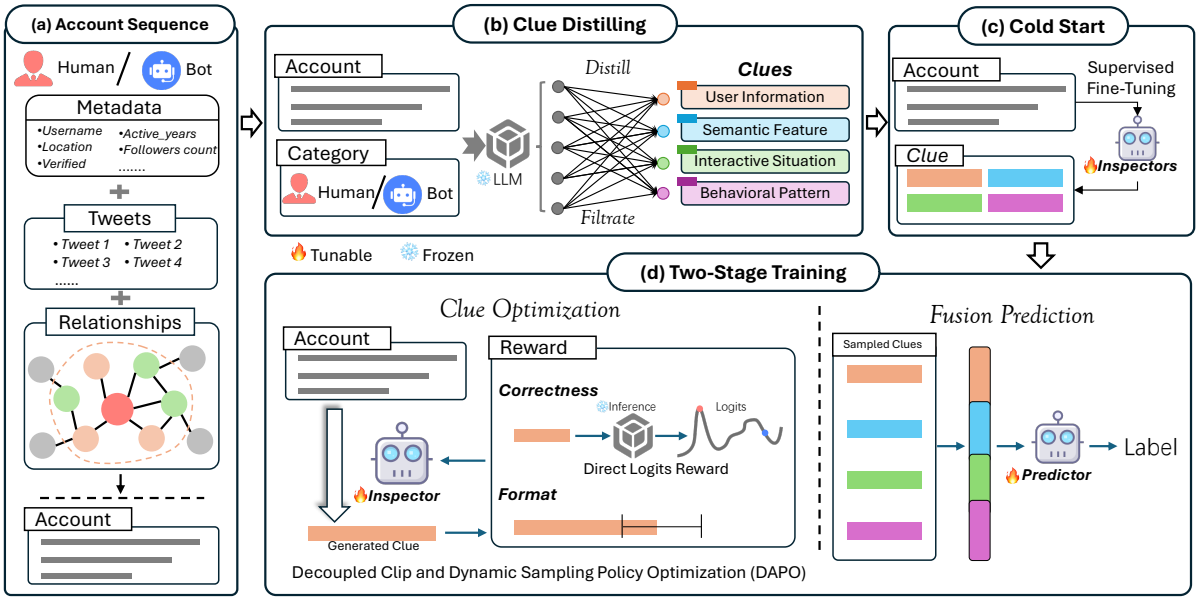

Figure 4: The overview of our proposed method, CDRBot, which has two main parts: (b) the clue distilling and filtration, and (d) two-stage training for clue optimization and prediction. In (c) and (d), for each clue dimension, an inspector corresponds to it.

responsible for each dimensional clue. On this basis, CDRBot adopts a two-stage training framework: (1) optimizing the clues via RL with hybrid outcome rewards, and (2) training an LLM that integrates the clue information for final prediction.

3.1 Clue Optimization

The objective is to obtain the inspector $\pi_\theta : u \to c$, which distills a kind of clue c from u. CDRBot employs an outcome-based reward RL to optimize the generated clues (Figure 4 (d)). The reward for clues consists of both correctness and format score. Based on Equation 2, CDRBot adopts the direct logits label $\hat{y}$ of the clue c and true account label y to get the correctness score $R_{correct}$:

$$R_{correct}(\hat{y}, y) = \mathbb{I}\{\hat{y} = y\}, \tag{3}$$

where $\mathbb{I}$ is the indicator function. Table 1 illustrates that the predicted label $\hat{y}$ of a clue given by small-scale LLM logits is consistent with the source model and nearly unbiased. We choose the Qwen2.5-1.5B with the lowest bias to compute the RL reward. About format, we only impose constraints on token length. Because we found that, during training, without length restrictions, the generated text increasingly tends to produce direct predictions rather than detailed analytical summaries. Formally, that is:

$$R_{format} = \mathbb{I}\{\, a \le \text{TokenLen}(c) \le b \,\}. \tag{4}$$

And with weight α, the final hybrid reward for c is:

$$R = \alpha R_{correct} + (1 - \alpha) R_{format}. \tag{5}$$

CDRBot utilizes the Decouple Clip and Dynamic Sampling Policy Optimization (DAPO) (Yu et al., 2025) to train the clue on rewards. DAPO is similar to the Generalized Reinforcement Policy Optimization (Shao et al., 2024), but has made some improvements on it. For a specific raw information and label pair (u, y) from dataset $\mathcal{D}$, the reference policy $\pi_{\theta_{old}}$ samples a group of G individual outputs $\{o_i\}_{i=1}^{G}$, and the advantage $\hat{A}_{i,t}$ at time step t of the i-th output is calculated by normalizing the group-level rewards $\{R_i\}_{i=1}^{G}$:

$$\hat{A}_{i,t} = \frac{R_i - \texttt{mean}(\{R_i\}_{i=1}^{G})}{\texttt{std}(\{R_i\}_{i=1}^{G})}. \tag{6}$$

Therefore, the optimization objective of the inspector π_θ is:

$$\begin{aligned}\mathcal{J}(\theta) =& \mathbb{E}_{(u,y)\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}(\cdot|u)} \\ & \left[\frac{1}{\sum_{i=1}^{G}|o_i|} \sum_{i=1}^{G}\sum_{t=1}^{|o_i|} \texttt{min}\left(r_{i,t}(\theta)\hat{A}_{i,t}, \right.\right. \\ & \left.\left. \texttt{clip}(r_{i,t}(\theta), 1-\epsilon_{low}, 1+\epsilon_{high})\hat{A}_{i,t} \right) \right] \\ \text{s.t.} \quad & 0 < \sum_{i=1}^{G} \mathbb{I}_{R_i \in \{0,1\}} < G, \end{aligned} \tag{7}$$

where

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}|u, o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|u, o_{i,<t})}. \quad (8)$$

We follow the work of Yu et al. (2025), using a higher ϵ_{high} to expand the upper bound (More details in §A.4). Based on DAPO, the inspector corresponding to each clue dimension can generate expected clues.

3.2 Fusion Prediction

Compared to the distilled clues, the models trained via RL may contain some biases and errors in the generated clues. Additionally, the contributions of clues from different dimensions to the categories are also inconsistent. Therefore, we train an LLM π_ϕ to integrate generated clues from all $k = 4$ dimensions and produce the final predictions (Figure 4 (d)). Formally, that is:

$$\begin{aligned} &\max_{\phi} \mathbb{E}_{(u,y)\sim\mathcal{D},C} \log \pi_\phi(y \mid C), \\ &\text{where } C = \left(c^{(1)}, c^{(2)}, c^{(3)}, c^{(4)}\right), \\ &\qquad c^{(k)} \sim \pi_\theta^{(k)}(\cdot \mid u), \quad k = 1, \ldots, 4. \end{aligned} \quad (9)$$

4 Experiments

4.1 Training

We utilize the Qwen2.5-14B-Instruct as the basis of the inspector and the Qwen2.5-3B-Instruct to make final predictions (Qwen et al., 2025). For the weight of the hybrid reward (Equation 5), we set $\alpha = 0.7$. More details are shown in A.4.

4.2 Datasets and Baselines

We conduct our experiments on two comprehensive and widely-used social bot detection benchmarks **Twibot-20** (Feng et al., 2021b), and **Twibot-22** (Feng et al., 2022b). Both datasets contain raw social account information collected from X (Twitter). These raw social account information involves a wide range of areas, languages, entities, and relationships.

We extensively compare our approach, CDRBot, with social bot detection methods, which generally can be categorized into three groups. *Feature-based* methods typically extract numerical and semantic features from the user metadata and textual content to make a prediction. In this paper, we report the performances of **SGBot** (Yang et al., 2020), **DeeProBot** (Hayawi et al., 2022), and **BotBuster** (Ng and Carley, 2023). *Graph-based* methods generally utilize the GNN-based method to encode the account information and the social topological network to detect social bots. We report performances of **GAT** (Wang et al., 2021), **BotRGCN** (Feng et al., 2022c), **RGT** (Feng et al., 2022a), and **BotDGT** (He et al., 2024). *LLM-based* methods generally leverage the LLM to detect the social bot based on the input account information with its abundant internal knowledge. We report the performance of **WDBS** (Feng et al., 2024), and its **SFT** version. For fair comparison, the prompt settings and the base LLM remain the same (Qwen2.5-14B-Instruct). Besides, given the increasing performance of reasoning models on complex problems, we report the performance of two advanced reasoning LLMs, **Qwen3-Max** (Yang et al., 2025) and **o3-mini** (OpenAI, 2025). We merely instruct them to reason and make predictions based on raw account sequences.

4.3 Evaluation Metrics

For the detection performance, we report the *Accuracy*, *F1-score*, *Precision*, and *Recall*. We also evaluated the qualities of the generated clues. In addition to automated evaluation, we also conduct the human evaluation (§ A.5) for clues in terms of *Faithfulness*, *Grounding*, and *Distinction*, which corresponds to § 2.3.

5 Results And Analysis

5.1 Social Bot detection

The performance of methods in bot detection is reported in Table 2. Experimental results indicate that our CDRBot achieves the best overall results on both benchmarks. In detail, about feature-based methods, SGBot shows the highest recall on the Twibot-20. This aligns with intuition, as the feature model is designed to prioritize capturing the unusual features. Such a sensitive design for features leads to a high recall. Besides, we observe that graph-based methods generally perform better than feature-based ones, with their main advantage being the incorporation of relationship information among users. This advantage also validates the rationale behind the social relationship clues design. However, graph-based methods exhibit serious metrics imbalance on the Twibot-22, mainly because the human account proportion in Twibot-22 is much greater than the bot accounts proportion, which leads to a low F1. Regarding LLM-based methods, the performances of two reasoning mod-

Type	Method	Twibot-20				Twibot-22			
		Accuracy	F1-score	Precision	Recall	Accuracy	F1-score	Precision	Recall
Feature-Based	SGBot	79.50±0.72	84.15±0.53	75.64±0.70	**93.54**±0.36	75.53±0.25	37.45±0.24	74.31±0.16	25.42±0.07
	DeeProBot	73.14±0.01	77.05±0.02	71.61±0.01	83.50±0.04	76.50±0.07	24.74±0.08	**80.00**±0.27	14.99±0.05
	BotBuster	78.55±0.44	82.12±0.61	79.85±0.74	84.00±0.53	74.33±0.17	52.26±1.82	63.32±1.47	45.64±1.70
Graph-Based	GAT	83.24±0.48	85.22±0.46	81.89±1.03	89.54±0.85	78.65±0.19	55.86±1.38	71.24±0.80	46.04±2.17
	BotRGCN	85.83±0.38	87.44±0.42	83.98±0.34	91.20±1.03	77.51±0.54	49.28±3.67	73.32±1.71	37.28±4.44
	RGT	86.53±0.47	87.74±0.62	86.08±0.64	88.90±0.43	77.01±0.21	47.25±0.83	72.80±0.76	34.99±0.90
	BotDGT	86.56±0.61	88.08±0.63	84.64±0.07	91.82±1.35	**79.04**±0.04	56.67±0.30	72.38±0.24	46.57±0.49
LLM-Based	Qwen3-Max	57.51±0.05	52.21±0.06	70.48±0.14	57.51±0.05	65.65±0.10	65.58±0.12	65.52±0.15	65.65±0.10
	o3-mini	57.48±0.59	55.12±0.61	62.88±0.93	57.48±0.59	69.37±0.59	65.36±0.79	64.60±0.99	69.37±0.59
	WDBS	65.99±0.56	65.89±0.56	65.89±0.57	65.99±0.56	49.64±0.02	50.89±0.03	64.50±0.00	49.64±0.02
	WDBS+SFT	78.33±1.39	78.25±1.38	78.34±1.42	78.33±1.39	71.64±0.11	67.28±0.14	65.22±0.18	70.64±0.11
	CDRBot (Ours)	**89.10**±0.08	**89.06**±0.09	**89.09**±0.08	89.09±0.08	77.73±0.38	**73.64**±0.47	76.15±0.24	**71.54**±0.42

Table 2: The experimental performances of different social bot detection methods on the TwiBot-20 and TwiBot-22 benchmarks. Each method has three independent runs, and we report the average value as well as the standard deviation. The **bold** and underline highlight the best and second-best results, respectively.

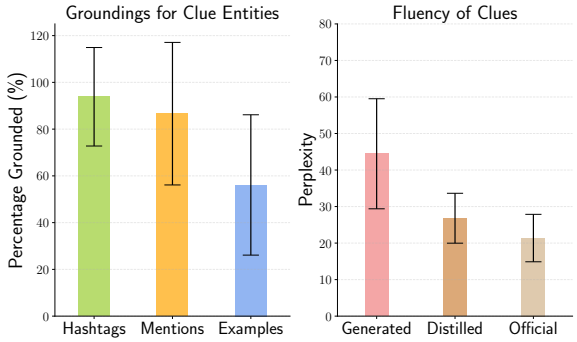

Figure 5: The grounding and fluency evaluations of the generated clues. The horizontal lines represent the mean and standard deviation, respectively.

Evaluation	User Inf.	Seman. Fea.	Intera. Sit.	Behav. Pat.
Faithfulness	0.825	0.855	0.775	0.843
Grounding	0.928	0.932	0.912	0.908

Table 3: The human evaluation results of the Faithfulness and Grounding of generated clues.

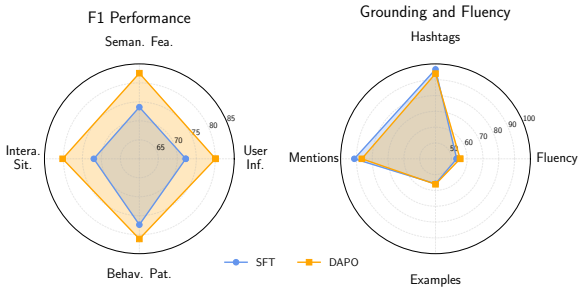

Figure 6: Comparisons of clues generated under SFT and DAPO training. Since a lower PPL indicates better fluency, we report *100 – PPL* so that higher values consistently represent better performance across all metrics.

els indicate that current advanced general reasoning models struggle to achieve good performance through heuristic reasoning on this task, highlighting the limitations of current general reasoning models. Overall, our method outperforms other baselines in terms of social bot detection.

5.2 Clue Evaluation

We evaluate and report the *Faithfulness*, *Grounding*, *Fluency*, and *Distinction* of generated clues by automatic (Figure 5) and human evaluation (Table 3). Good detection performance (Table 2) and human faithfulness results (Table 3) show that the generated clues exhibit good faithfulness, indicating that most generated clues could accurately preserve the category-relevant evidential information from the raw account. Besides, most entity words in clues are derived from raw account sequences, demonstrating that the clues genuinely condense and summarize the original information. Compared to the distilled clues (Figure 2), all metrics show a performance decline, and the perplexity also slightly increases. This phenomenon suggests that there is still room for optimization in the model under DAPO training. This may be due to insufficient training on our experiments or limitations of the base model. And the distinction score by human evaluation is 0.876, which demonstrates all four dimensional clues are necessary and the training process is as expected. More discussions are listed in §C and § B.1.

Besides, Figure 6 further demonstrates that DAPO substantially enhances the faithfulness of clues across all dimensions compared with the SFT, while preserving comparable grounding and fluency. These results highlight that DAPO not only

Clue	Twibot-20		Twibot-22	
	Accuracy	F1-score	Accuracy	F1-score
User Inf.	81.44±0.83	80.85±0.82	74.58±0.11	72.02±0.58
Seman. Fea.	82.58±0.73	82.55±0.74	74.45±0.05	72.03±0.17
Intera. Sit.	80.48±1.42	80.21±1.60	72.11±0.04	64.29±0.14
Behav. Pat.	81.17±0.65	81.10±0.62	71.74±0.06	64.09±0.09
Majority Vote	85.08±0.04	85.12±0.18	74.78±0.11	73.15±0.29

Table 4: Social bot detection ablation performance of each dimensional clue.

Clue	Qwen3-Max		o3-mini	
	Accuracy	F1-score	Accuracy	F1-score
Raw	57.51±0.05	52.21±0.06	57.48±0.59	55.12±0.61
User Inf.	77.07±0.12	77.21±0.12	77.60±0.72	77.74±0.72
Seman. Fea.	81.20±0.20	81.05±0.19	**81.93**±0.83	**81.95**±0.83
Intera. Sit.	79.00±0.00	79.08±0.00	80.53±0.99	80.64±1.01
Behav. Pat.	81.40±0.00	81.51±0.00	80.80±0.20	80.91±0.20
Ensemble	**81.53**±0.23	**81.63**±0.23	81.80±0.42	81.92±0.42

Table 5: The social bot detection performance of advanced reasoning LLMs based on given clues. The **bold** and underline highlight the best and second-best results, respectively.

strengthens the analytical rigor of the distilled clues but does so without compromising their evidential grounding or linguistic naturalness, underscoring its value as a more reliable clue-distillation paradigm.

5.3 Clue Difference Analysis

Table 4 evaluates the detection performance of each clue dimension separately, where the predicted labels for clues in each dimension are obtained by prompting the LLM to make direct predictions based on the clues (Equation 2). Experimental results indicate that clues from each dimension can achieve relatively good performance in social bot detection, and such performance can be further improved through a simple strategy (Majority Voting). Additionally, we found that the overall performance of the first two clues (individual level) is slightly higher than that of the last two clues (social level). This indicates that the clues at the individual level are more explicit and clear, directly supporting the model in obtaining correct predictions. In contrast, the slightly lower performance in the two social level dimensions suggests that the clues in this part are relatively ambiguous and less explicit, leading to the lower overall performance. This indicates that our experiments are not sufficient to explore clues at the social level. This may be limited by the quality of the distillation data or the performance of the base LLM. In short, there is still room for further exploration.

5.4 Boosting the LLM Detection

Since Table 2 indicates that advanced reasoning LLM performs poorly in bot detection, we explore whether these generated clues are helpful for these models. We prompt two reasoning LLMs (Table 20) to reason with given clues and then make predictions. The detection results on Twibot-20 are listed in Table 5. It shows that all clues significantly increase the detection of two reasoning LLMs, which validates that the generated clues are helpful. Besides, compared with directly predicting labels based on clues (Table 4), the performance of both reasoning LLMs slightly decreases across different dimensions. This phenomenon suggests that reasoning on clues during the prediction process introduces extra uncertainty or bias, thereby leading to a slight overall reduction in performance. Although the decoding bias could be reduced by multiple sampling (Muldrew et al., 2024), this phenomenon also indirectly demonstrates the rationality of the setting of the direct logits label.

Additionally, contrary to Table 4, the overall performance of the latter two clue dimensions (social level) in Table 5 is slightly better than that of the first two clue dimensions. This indicates that while reasoning LLMs introduce uncertainty, they can extract crucial information from some ambiguous or unclear clues through a heuristic reasoning process. Consequently, this leads to an improvement in the relative performance of the latter two clue dimensions. This phenomenon also reveals the potential optimization direction for our approach, such as incorporating deeper and longer reasoning processes to increase detection performance.

6 Related Work

Early research on social bot detection relies on extracting features from user metadata to distinguish bots from humans (Echeverrï£¡a et al., 2018; Mazza et al., 2019; Yang et al., 2020). Other studies analyze textual characteristics of bot-generated content to improve detection (Wei and Nguyen, 2019; Kudugunta and Ferrara, 2018; Feng et al., 2021a; Dukić et al., 2020; Ma et al., 2025). More recent work models social relationships using graph-based methods, leveraging GNNs to capture social network structures (Pham et al., 2022; Feng et al.,

2022a,c; Peng et al., 2024; Huang et al., 2025; Feng et al., 2022b), and exploring the dynamic network characteristics He et al. (2024). Additionally, recent studies explore using LLMs for social bot detection (Cai et al., 2024; Feng et al., 2024).

7 Conclusion

In this paper, we introduced an explainable clue framework for social bot detection. We proposed a two-stage training process to generate high-quality multi-dimensional clues and integrated them for predicting social bots. Experiments demonstrated that our method outperformed baselines in detection performance and produced faithful, grounded, and fluent clues.

Limitations

This study explores the large language models' distillation and reasoning for social bot detection. Compared to traditional feature-based and graph-based methods, this approach leads to higher computational resource demands, as well as increased time and energy consumption. Another limitation is that the datasets used in this study are sourced from the X (Twitter) platform, which lacks sufficient discussion on other platforms. Furthermore, regarding the two large-scale benchmarks used in this study, some of the labels are obtained through weakly supervised classifiers. Although these labels are widely used, we remain uncertain about their faithfulness, as some of them are likely to be incorrect. This paper focuses solely on specific algorithms or approaches, and does not involve a discussion of the inherent limitations of existing datasets.

Ethical Considerations

When comparing the fluency of distilled and generated clues, we used an official court document as a reference. We declare that all legal materials employed in this study are sourced from publicly accessible federal court rulings of the United States. Under U.S. law, judicial opinions issued by federal courts belong to the public domain and are not subject to copyright restrictions. Therefore, citing, analyzing, and discussing the relevant judgments fully complies with academic ethics and legal norms. This study does not utilize any copyrighted non-public materials, nor does it distort or misrepresent the original content.

Besides, all datasets used in this study have been formally authorized by the original authors or data providers, ensuring that the data usage process complies with relevant legal and ethical requirements. All sample data presented in this paper have undergone rigorous screening and processing, and do not contain any personally identifiable information, nor do they disclose any personal privacy.

Furthermore, we emphasize that the social bot detection method developed and discussed in this study is solely intended to provide preliminary screening references for relevant personnel. Its purpose is to offer explanations and analysis to assist manual review, rather than to replace human judgment. Any final review, judgment, or decision should be formulated with the participation of professional human reviewers. This study neither encourages nor endorses the use of model outputs as the sole basis for automated decision-making.

Acknowledgements

The authors would like to thank all the anonymous reviewers for their insightful comments. This work is supported by the National Science and Technology Major Project (No. 2025ZD1606200 and Sub-project No. 2025ZD1606203) and the National Natural Science Foundation of China (No. 92470205). Also, this research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-RP-2024-035).

References

Mohammad Majid Akhtar, Rahat Masood, Muhammad Ikram, and Salil S Kanhere. 2024. Sok: False information, bots and malicious campaigns: Demystifying elements of social media manipulations. In *Proceedings of the 19th ACM Asia Conference on Computer and Communications Security*, pages 1784–1800.

Meng Cai, Han Luo, Xiao Meng, and Ying Cui. 2022. Differences in behavioral characteristics and diffusion mechanisms: A comparative analysis based on social bots and human users. *Frontiers in Physics*, 10:875574.

Zijian Cai, Zhaoxuan Tan, Zhenyu Lei, Zifeng Zhu, Hongrui Wang, Qinghua Zheng, and Minnan Luo. 2024. Lmbot: Distilling graph knowledge into language model for graph-less deployment in twitter bot detection. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, page 57–66, New York, NY, USA. Association for Computing Machinery.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

David Dukić, Dominik Keča, and Dominik Stipić. 2020. Are you human? detecting bots on twitter using bert. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 631–636.

Juan Echeverrï£¡a, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Gianluca Stringhini, and Shi Zhou. 2018. Lobo: Evaluation of generalization deficiencies in twitter bot classifiers. In *Proceedings of the 34th Annual Computer Security Applications Conference*, ACSAC '18, page 137–146, New York, NY, USA. Association for Computing Machinery.

Shangbin Feng, Zhaoxuan Tan, Rui Li, and Minnan Luo. 2022a. Heterogeneity-aware twitter bot detection with relational graph transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3977–3985.

Shangbin Feng, Zhaoxuan Tan, Herun Wan, Ningnan Wang, Zilong Chen, Binchi Zhang, Qinghua Zheng, Wenqian Zhang, Zhenyu Lei, Shujie Yang, Xinshun Feng, Qingyue Zhang, Hongrui Wang, Yuhan Liu, Yuyang Bai, Heng Wang, Zijian Cai, Yanbo Wang, Lijing Zheng, and 3 others. 2022b. Twibot-22: Towards graph-based twitter bot detection. In *Advances in Neural Information Processing Systems*, volume 35, pages 35254–35269. Curran Associates, Inc.

Shangbin Feng, Herun Wan, Ningnan Wang, Jundong Li, and Minnan Luo. 2021a. Satar: A self-supervised approach to twitter account representation learning and its application in bot detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 3808–3817, New York, NY, USA. Association for Computing Machinery.

Shangbin Feng, Herun Wan, Ningnan Wang, Jundong Li, and Minnan Luo. 2021b. Twibot-20: A comprehensive twitter bot detection benchmark. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 4485–4494, New York, NY, USA. Association for Computing Machinery.

Shangbin Feng, Herun Wan, Ningnan Wang, and Minnan Luo. 2022c. Botrgcn: Twitter bot detection with relational graph convolutional networks. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '21, page 236–239, New York, NY, USA. Association for Computing Machinery.

Shangbin Feng, Herun Wan, Ningnan Wang, Zhaoxuan Tan, Minnan Luo, and Yulia Tsvetkov. 2024. What does the bot say? opportunities and risks of large language models in social media bot detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3580–3601, Bangkok, Thailand. Association for Computational Linguistics.

Stephan Grimmelikhuijsen. 2023. Explaining why the computer says no: Algorithmic transparency affects the perceived trustworthiness of automated decision-making. *Public Administration Review*, 83(2):241–262.

Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. 2023. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20606–20615.

Kadhim Hayawi, Sujith Mathew, Neethu Venugopal, Mohammad M Masud, and Pin-Han Ho. 2022. Deeprobot: a hybrid deep neural network model for social bot detection based on user profile data. *Social Network Analysis and Mining*, 12(1):43.

Buyun He, Yingguang Yang, Qi Wu, Hao Liu, Renyu Yang, Hao Peng, Xiang Wang, Yong Liao, and Pengyuan Zhou. 2024. Dynamicity-aware social bot detection with dynamic graph transformers. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 5844–5852. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Di Huang, Jinbao Song, and Xingyu Zhang. 2025. Semi-supervised social bot detection with relational graph attention transformers and characteristics of the social environment. *Information Fusion*, 118:102956.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Hanchi Sun, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bertie Vidgen, Bhavya Kailkhura, Caiming Xiong, and 52 others. 2024. Position: TrustLLM: Trustworthiness in large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 20166–20270. PMLR.

Zhen Huang, Zhilong Lv, Xiaoyun Han, Binyang Li, Menglong Lu, and Dongsheng Li. 2022. Social bot-aware graph neural network for early rumor detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6680–6690, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of*

the Association for Computational Linguistics (Volume 1: Long Papers), pages 4459–4471, Dublin, Ireland. Association for Computational Linguistics.

Sneha Kudugunta and Emilio Ferrara. 2018. Deep neural networks for bot detection. *Information Sciences*, 467:312–322.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Longxuan Ma, Xiao Wu, Yuxin Huang, Shengxiang Gao, and Zhengtao Yu. 2025. 3R: Enhancing sentence representation learning via redundant representation reduction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31630–31643, Suzhou, China. Association for Computational Linguistics.

Kinga Makovi, Anahit Sargsyan, Wendi Li, Jean-François Bonnefon, and Talal Rahwan. 2023. Trust within human-machine collectives depends on the perceived consensus about cooperative norms. *Nature communications*, 14(1):3108.

Michele Mazza, Marco Avvenuti, Stefano Cresci, and Maurizio Tesconi. 2022. Investigating the difference between trolls, social bots, and humans on twitter. *Computer Communications*, 196:23–36.

Michele Mazza, Stefano Cresci, Marco Avvenuti, Walter Quattrociocchi, and Maurizio Tesconi. 2019. Rtbust: Exploiting temporal patterns for botnet detection on twitter. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, page 183–192, New York, NY, USA. Association for Computing Machinery.

Marcelo Mendoza, Eliana Providel, Marcelo Santos, and Sebastián Valenzuela. 2024. Detection and impact estimation of social bots in the chilean twitter network. *Scientific reports*, 14(1):6525.

Himanshu Mishra, Pallavi Singh, Melbin Kurien, and Suhail Javed Quraishi. 2025. Detecting bots in misinformation campaigns: Strategies and societal implications. In *2025 International Conference on Pervasive Computational Technologies (ICPCT)*, pages 917–921.

William Muldrew, Peter Hayes, Mingtian Zhang, and David Barber. 2024. Active preference learning for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Lynnette Hui Xian Ng and Kathleen M Carley. 2023. Botbuster: Multi-platform bot detection using a mixture of experts. In *Proceedings of the international AAAI conference on web and social media*, volume 17, pages 686–697.

Lynnette Hui Xian Ng and Kathleen M Carley. 2025. A global comparison of social media bot and human characteristics. *Scientific Reports*, 15(1):10973.

OpenAI. 2025. Openai o3 and o4-mini system card. Accessed 2025-12-30.

Arjunil Pathak, Navid Madani, and Kenneth Joseph. 2021. A method to analyze multiple social identities in twitter bios. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2).

Hao Peng, Jingyun Zhang, Xiang Huang, Zhifeng Hao, Angsheng Li, Zhengtao Yu, and Philip S Yu. 2024. Unsupervised social bot detection via structural information theory. *ACM Transactions on Information Systems*, 42(6):1–42.

Phu Pham, Loan TT Nguyen, Bay Vo, and Unil Yun. 2022. Bot2vec: A general approach of intra-community oriented representation learning for bot detection in different types of social networks. *Information Systems*, 103:101771.

Rachel Pozzar, Marilyn J Hammer, Meghan Underhill-Blazey, Alexi A Wright, James A Tulsky, Fangxin Hong, Daniel A Gundersen, and Donna L Berry. 2020. Threats of bots and other bad actors to data quality following research participant recruitment through social media: Cross-sectional questionnaire. *J Med Internet Res*, 22(10):e23021.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Débora Salles, Priscila Muniz de Medeiros, Bruno Martins, Lorena Regattieri, and Rose Marie Santini. 2024. The role of social bots in the brazilian environmental debate:: an analysis of the 2020 amazon forest fires in twitter. *The International Review of Information Ethics*, 33(1).

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Kumar Shridhar, Koustuv Sinha, Andrew Cohen, Tianlu Wang, Ping Yu, Ramakanth Pasunuru, Mrinmaya Sachan, Jason Weston, and Asli Celikyilmaz. 2024. The ART of LLM refinement: Ask, refine, and trust.

In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5872–5883, Mexico City, Mexico. Association for Computational Linguistics.

Onur Varol, Emilio Ferrara, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):280–289.

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 31481–31504, Suzhou, China. Association for Computational Linguistics.

Yanbang Wang, Pan Li, Chongyang Bai, and Jure Leskovec. 2021. Tedic: Neural modeling of behavioral patterns in dynamic social interaction networks. In *Proceedings of the Web Conference 2021*, pages 693–705.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in LLM-as-a-judge. In *Neurips Safe Generative AI Workshop 2024*.

Feng Wei and Uyen Trang Nguyen. 2019. Twitter bot detection using bidirectional long short-term memory neural networks and word embeddings. In *2019 First IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, pages 101–109.

Magdalena Wischnewski, Thao Ngo, Rebecca Bernemann, Martin Jansen, and Nicole Krämer. 2024. "i agree with you, bot!" how users (dis)engage with social bots on twitter. *New Media & Society*, 26(3):1505–1526.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Kai-Cheng Yang, Onur Varol, Pik-Mai Hui, and Filippo Menczer. 2020. Scalable and generalizable social bot detection through data selection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):1096–1103.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, and 16 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *ArXiv*, abs/2503.14476.

Ming Zhou, Dan Zhang, Yuandong Wang, Yangli-ao Geng, Yuxiao Dong, and Jie Tang. 2025. Lgb: Language model and graph neural network-driven social bot detection. *IEEE Transactions on Knowledge and Data Engineering*, 37(8):4728–4742.

Caleb Ziems, Jane Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022. The moral integrity corpus: A benchmark for ethical dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3755–3773, Dublin, Ireland. Association for Computational Linguistics.

A Appendix

A.1 Distilling Details

We sample 6801 accounts from the training data of Twibot-20, and prompt it to Deepseek-R1 with template (Table 18). The base model we use is DeepSeek-R1-0528 and we choose the reasoning mode. To get the optimal outputs, we distill clues from Deepseek-R1 with multi-turns and ensure that all generated clues can instruct R1 to obtain true label. So in Table 1, the accuracy of clues for R1 is 100%.

A.2 Dataset Details

We conduct experiments on two widely used social bot detection benchmarks. For Twibot-20, we use the raw dataset, which contains 8,278, 2,365, and 1,183 accounts in the training, validation, and test splits, respectively. For Twibot-22, we follow the same settings as Zhou et al. (2025), using a balanced set of 81,432 training accounts, while the validation and test splits stay unchanged.

A.3 Clue Evaluation

For faithfulness evaluation, we use a prompt (Table 20) to ask LLMs to make a prediction based on a raw account sequence or clues. All generation configs are default. Based on the special design of the prompt, we can use the output logits of the last input token to make a prediction. In the tokenizers of all listed LLMs in Table 1, both words "Bot" and "Human" each correspond to only one token. This allows us to obtain the prediction result by comparing the logits of these two tokens. About the fluency, we use a publicly available official opinion of the U.S. court. We compute the perplexity based on the main paragraph on pages 2-4.

To assess label leakage in the clues, we detect whether there are independent label terms in the clue text through word matching. Since our prompt contains fixed descriptions of account labels (e.g., "is a bot account", "is a human account"), we are more concerned with whether these phrases, such as "a bot" or "a human", can also be found in the clues. Therefore, we calculate the occurrence rate of these independent terms and report the overall ratio of label leakage.

Regarding the calculation details of BLEU, we utilized the *sentence_bleu* function from the *nltk* package with the default weights and settings. In the BLEU matrix (Figure 3), the number at row i and column j represents the BLEU score obtained when using the i-th dimensional clue as the candidate and the j-th dimensional clue as the reference. Therefore, this matrix is asymmetric.

A.4 Training Details

About DAPO, in our experiments, we set $\epsilon_{high} = 0.32$ as suggested. When calculating $\hat{A}_{i,t}$ by Equation 6, to enabling more robust reward shaping, we calculate $\hat{A}_{i,t}$ on *batch* instead of on *group*.

For the length reward, based on the average of the initial clues and early experiments, we set $a = 64$ and $b = 196$, which is sufficient to accommodate the helpful clue information.

About the training implementation, we choose the Transformer Reinforcement Learning (TRL) framework[4] to do SFT (SFT Trainer) and DAPO (GRPO Trainer) training. To train a 14B LLM on 6×A800-SXM4-80GB GPUs, we enable the *accelerate* and *Deepspeed-Stage-3*[5] to speed up training and optimize the cuda memory. To efficiently generated long-sequence clues, we enable vLLM to speed up generation (Kwon et al., 2023). Our main experimental environments are:

Parameter	Value
learning_rate	8e-6
epsilon_high	0.32
scale_rewards	group
loss_type	dapo
max_completion_length	196
num_train_epochs	12
num_generations	8
per_device_train_batch_size	1
gradient_accumulation_steps	18

Table 6: The DAPO training config.

Parameter	Value
deepspeed_multinode_launcher	standard
offload_optimizer_device	none
offload_param_device	none
zero3_init_flag	true
zero_stage	3
mixed_precision	auto
num_processes	4

Table 7: The Deepspeed config used in training.

A.5 Human Evaluation

Six annotators with undergraduate backgrounds are recruited to evaluate each dimensional of clues.

[4]https://huggingface.co/docs/trl/main/en/index
[5]https://github.com/deepspeedai/DeepSpeed

Before the annotation process, annotators are thoroughly informed of the evaluation details, and we have obtained their consent. The human evaluation for all criterion are based on the 3-point Likert scale, and higher is better. About the evaluation aspects, the *Faithfulness* focus on whether the clue analysis is reasonable or meaningful, and align with the account category (0-Not Faithful, 1-Partly Faithful, and 2-Faithful), *Grounding* meanings whether the clue content originates from the original content (0-Not Grounded, 1-Partly Grounded, and 2-Grounded), and *Distinction* measures whether all clues discuss different aspects (0-All clues are highly similar, 1-Only some clues are similar, and 2-All clues are distinct), respectively. The annotation process is paid, and the hourly wage significantly exceeds the local average.

We first present the annotators with ten account annotations for reference. The annotators have an acceptable inter-rater agreement (the Fleiss' kappas of all evaluation aspects are range from 0.442 to 0.702). We scale each score into $[0, 1]$ and report the average value for clues. Each evaluation result is calculated by two hundred sampled clues from the Twibot-20. The human evaluation results are listed in Table 3.

A.6 AI Usage Statement

All writings and figures in this paper are produced by humans, and we use a toolkit called Grammarly[6] during final proofreading to check spelling.

B Additional Analysis

B.1 Clue Explanation

Given that we distill a batch of clue texts from the raw account information based on our clue theory framework, we are more interested in studying what these clues reveal in different dimensions. What kind of explanations can support the LLM/Human in making correct predictions, and whether the spontaneous reasoning-derived explanations for each dimension are truly reasonable and meaningful? Based on this, we conduct part-of-speech tagging (POS) on all clue texts across the four dimensions[7], filter out stop words, and extract the structural pairs of "*adjective + noun*". Word cloud diagrams are then created for the clues in each dimension to visualize what each dimension specifically analyzes.

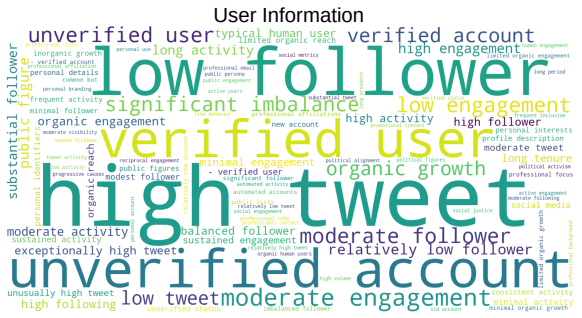

Figure 7: The word cloud of the main analyzed targets in the *User Information*.

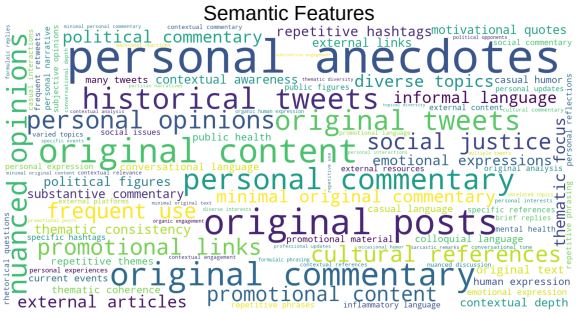

Figure 8: The word cloud of the main analyzed targets in the *Semantic Features*.

Figure 9: The word cloud of the main analyzed targets in the *Interactive Situation*.

The results and word clouds are shown in Figure 7, 8, 9, and 10. These word clouds indicate that these distilled clues from each dimension spontaneously and profoundly focus on distinct, interpretable aspects of account behavior, generating human-understandable clues for the social bot detection. To further examine the semantic focus and explanatory roles of these clues, we provide a simple analysis of each dimension, illustrating how spontaneously derived explanations capture meaningful and complementary behavioral signals.

[6] https://app.grammarly.com/

[7] The Python toolkit *spacy* and model *en_core_web_sm* are used for POS.

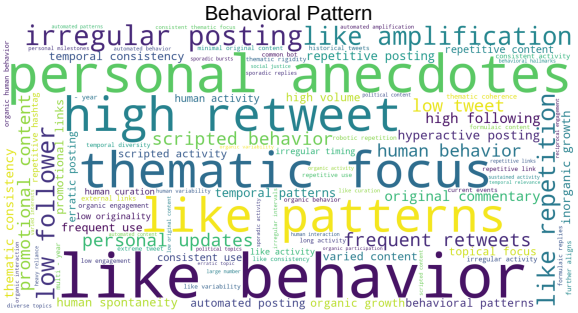

Figure 10: The word cloud of the main analyzed targets in the *Behavioral Pattern*.

User Information. In the *user information* dimension, the distilled clues primarily analyze the static composition and credibility signals embedded in account profiles. The model spontaneously highlights anomalous configurations, such as accounts exhibiting a *high tweet* count alongside a *low follower* base, which is widely recognized as a hallmark of inauthentic, broadcast-oriented behavior. Moreover, the clues contrast *unverified accounts* with *verified users* and emphasize growth-related patterns, including *significant imbalance* and the absence of *organic growth*. Collectively, these indicators point to profiles that lack genuine social capital or platform-authenticated identity, enabling the model to form an intuitive and interpretable assessment of account legitimacy directly from profile-level metadata.

Semantic Features. Within the *semantic features* dimension, the clues naturally focus on linguistic patterns and content originality. They distinguish human-like expressions, such as *personal anecdotes* and *original commentary*, from bot-like signals, including *promotional content* and *repetitive hashtags*. At the same time, the model surfaces markers of authentic human expression, such as *emotional expressions* and *contextual depth*, while identifying mechanistic language usage through patterns like *repetitive phrases* and *formulaic phrasing*. These semantic cues provide a transparent and interpretable lens for determining whether an account's textual output reflects genuine personal cognition or scripted, automated dissemination.

Interactive Situation. The *interactive situation* dimension evaluates the quality and authenticity of social relationships by examining engagement behaviors. The distilled clues emphasize interaction depth and reciprocity, favoring accounts that demonstrate *sustained dialogue* and *direct engagement* with *other users*. In contrast, they reveal shallow or asymmetric interaction patterns commonly associated with social bots, such as one-way engagement with *public figures*, the absence of *reciprocal conversations*, or the prevalence of *generic replies*. Through these signals, the clues effectively characterize an account's interaction style and social positioning, distinguishing genuine community participants from isolated propagation-oriented entities.

Behavioral Pattern. In the *behavioral pattern* dimension, the clues capture the temporal dynamics and activity signatures of accounts. They automatically identify irregularities in posting rhythms, including *irregular posting* and *hyperactive posting*, as well as amplification behaviors such as *high retweet counts* and *like amplification*. Additionally, the clues highlight content-level regularities, including *low originality* and *scripted behavior*. By focusing on temporal consistency, this dimension exposes automated scheduling and engagement strategies that are difficult for bots to conceal but readily recognizable by humans as non-organic behavioral patterns.

Overall, the distilled clues from the four dimensions spontaneously converge on complementary and interpretable key features of account behavior without requiring manual intervention. And these clues collectively construct a multi-layered, intuitive, and human-intelligible explanatory framework. They not only effectively uncover anomalous signals of social bots across different dimensions but also demonstrate the model's ability to compress complex behaviors into discriminative explanatory evidence, thereby providing robust and transparent support for explainable social bot detection.

B.2 Error Analysis

The generated cues may, in certain scenarios, support predictions that are diametrically opposed to the true account type. Table 8 present two error cases from different cue dimensions, and we will analyze the potential causes of these errors based on these cases.

For the first case, the error stems from over-reliance on statistical features and verification badges. The account's characteristics closely match those of a typical institutional or promotional account, leading the model to classify it as a bot. However, this account is actually the official

Account: u1xxxxxxxx3 **Type**: Human
User Information:
The user "□" (□) is a verified account associated with the TV show, active for seven years with over 2.1 million followers. Its profile highlights the final season of the series, includes a listed count of 2,480, and features a detailed location "□" mirroring the show's setting. The account has tweeted over 26,000 times, reflecting sustained activity consistent with a long-running, high-profile media property. Notably, its followings_count (364) is minimal relative to its follower base, typical of institutional or promotional accounts prioritizing outreach over reciprocal engagement.
Clue Predict: **Bot**.

Account: u12xxxxxxxxxxxxx867 **Type**: Bot
Semantic Features:
The tweets exhibit informal, emotionally expressive language with slang ("□ off", "□"), humor (e.g., "□" self-identification), and personal anecdotes (e.g., crying sessions with □). Content includes contextual references to real-world topics (e.g., food preferences, memes like "□") and uses emojis to convey feelings, demonstrating natural, topic-diverse communication without repetitive or overly formulaic structures. Messages lack promotional links or repetitive themes, showing spontaneous, contextually grounded expression.
Clue Predict: **Human**.

Table 8: An error case where the user information clue mislead the model into a wrong classification.

account of a TV series. Although it possesses these surface-level features, its operators might be a real human team. The model overlooks the fact that account types (e.g., media outlets) might be maintained by humans and fails to integrate other clues, such as content, for a comprehensive judgment, resulting in a misjudgment.

For the second case, the error stems from over-reliance on the natural language features. The account's tweets utilize rich slang, emojis, and humorous expressions, presenting a diverse and emotional linguistic style that closely mimics the daily communication of human users. The model is thus misled by this superficially natural and emotional expression, classifying it as human. However, this account is actually a meticulously designed bot capable of imitating human language patterns but may exhibit anomalies in other dimensions (such as posting regularity or interaction patterns). The model makes its judgment based solely on textual content and fails to identify the potential traces of automation behind it, leading to a misjudgment.

Based on the cases above, we see that the clues from a single dimension often have inherent limitations. To determine whether an account is a social bot, it is essential to combine clues from various dimensions for a comprehensive judgment.

C Case study

Account: u*******58
{Raw account information} #Omitted for ethic considerations

Clues:
User Information:
The user □ (@□) has operated for □ years in Hawaii, with □ followers and □ accounts followed. Despite its longevity, it lacks verification and has a relatively low follower-to-following ratio. The profile emphasizes free business listings and promotional hashtags like # □. Its bio is transactional, focused on driving business engagement rather than personal or community-driven messaging. The account's listed count □ is modest compared to its tweet volume □, suggesting limited curation or audience interest in organizing its content.

Semantic Feature:
The account predominantly shares external links to articles, promotions, and advertisements, often with minimal or repetitive captions (e.g., URLs without context). Content themes are fragmented, ranging from business strategies and local □ promotions to unrelated topics like energy trends, with no cohesive narrative beyond general business advocacy. Hashtags like #□ and location tags (#□) are reused mechanically. Many posts are retweets of commercial entities, lacking original commentary or analysis, and some tweets contain only hyperlinks, indicating automated or templated sharing.

Table 9: The **distilled** clues for a social account which is marked as the **Bot**. Some parts are omitted by * and □ for information protection.

Account: u*******58
{Raw account information} #Omitted for ethic considerations

Clues:
Interactive Situation:
Interactions are heavily skewed toward retweets of businesses (e.g., @□, @□) or growth-focused accounts (@□). Mentions of other users are primarily promotional (e.g., tagging @□ in ads) or event-related (@□). Replies are rare and superficial (e.g., "@□ ...it may already be surging ahead of schedule. :)"), with no sustained conversations. The account does not engage meaningfully with followers, suggesting a focus on broadcasting rather than building relationships.

Behavioral Pattern:
The account exhibits automated patterns, including high-volume link-posting, repetitive promotional hashtags, and retweets of commercial partners. Tweets often lack contextual text (e.g., standalone URLs) and follow a templated structure (e.g., "□ @□ #□..."). Activity appears geared toward □ or directory-style listings (e.g., hardwood flooring ads with phone numbers). Despite a decade of activity, there is no evolution in content style or engagement strategy, consistent with machines designed for persistent, low-variation output.

Table 10: The **distilled** clues for a social account which is marked as the **Bot**. Some parts are omitted by * and □ for information protection.

Account: u*****645
{Raw account information} #Omitted for ethic considerations

Clues:
User Information:
□ is a verified user based in Auckland, New Zealand, with □ years of activity. He has □ followers, follows □ accounts, and has posted □ tweets. His profile identifies him as a tech/business commentator, futurist, keynote speaker, and CEO associated with entities like @□ and @□. His professional affiliations, long-term activity, and moderate follower-to-following ratio align with a human user in media and tech leadership roles.

Semantic Feature:
Tweets reflect expertise in technology and business, including podcast promotions, industry commentary, and discussions about Bluetooth contact tracing, device backups, and software comparisons. Content includes contextualized opinions (e.g., supporting □ blocking harmful content), personal anecdotes (e.g., airport delays), and professional engagements (e.g., □). Use of hashtags like # □ and # □, along with links to podcasts and news articles, demonstrates thematic consistency with his stated roles.

Table 11: The **distilled** clues for a social account which is marked as the **Human**. Some parts are omitted by * and □ for information protection.

Account: u*****645
{Raw account information} #Omitted for ethic considerations

Clues:
Interactive Situation:
Engages in bidirectional conversations, such as advising @□ on software evaluation, asking @□ about device vendors, and responding to queries about voicemail alternatives. Mentions collaborators like @□ and @□ in professional contexts. Retweets colleagues (e.g., @□) and participates in community discussions (e.g., @□ events). Direct interactions with followers and industry peers suggest authentic dialogue.

Behavioral Pattern:
Demonstrates sustained, varied activity: sharing news, retweeting podcasts, commenting on tech trends, and reacting to real-world events (e.g., □ policy). Balances self-promotion with organic engagement, such as congratulating peers or discussing travel frustrations. Follows niche tech professionals (e.g., □) and media figures, aligning with his professional interests. Regular participation in time-sensitive discussions (e.g., live radio appearances) and multi-year project continuity (e.g., podcast episode milestones) further indicate the behavioral complexity.

Table 12: The **distilled** clues for a social account which is marked as the **Human**. Some parts are omitted by * and □ for information protection.

Account: u**********487808
{Raw account information} #Omitted for ethic considerations

Clues:
User Information:
The user exhibits characteristics typical of bot-like profiles, including a mismatch between the account's claimed activation date (□ year) and an unusually high tweet count (□), far exceeding typical human engagement levels. Despite claiming to be a constitutionalist conservative with ties to real-world politics (e.g., □, □), the account lacks verification and has a low listed count (1), while follower and following numbers (□ and □, respectively) suggest minimal organic growth. The profile description blends generic political slogans (#□, #□) with personal claims (e.g., "□"), but the lack of activity specificity (e.g., no verified location, minimal profile depth) contrasts with the high tweet volume, a pattern often linked to automated activity.

Semantic Feature:
The user's tweets and retweets exhibit repetitive, hyper-partisan content with emotionally charged language (e.g., "insane," "ignorant," "arrested and put in jail immediately") and slogans (e.g., #□, #□). Posts lack substantive discussions or nuanced opinions and rely heavily on memes, emojis (□), and inflammatory claims (e.g., "□," "□" accusations). Content topics are fragmented and contextually shallow, focusing on reactive political narratives (e.g., □, □, □ references) without sustained thematic depth or personal anecdotes beyond ideological posturing. Links and media (images, videos) are shared without explanatory context, and replies often consist of aggressive rebuttals (e.g., "□") rather than dialogue, indicating a quasi-botsman pattern of repetitive, contextually disconnected advocacy.

Table 13: The **generated** clues for a social account which is marked as the **Bot**. Some parts are omitted by * and □ for information protection.

Account: u**********487808
{Raw account information} #Omitted for ethic considerations

Clues:
Interactive Situation:
Interactions are predominantly one-directional, consisting of mass mentions, retweets, and formulaic replies lacking personalized engagement. The user frequently tags accounts with clearly partisan or promotional profiles (e.g., @□, @□), suggesting participation in coordinated networks rather than organic communities. Replies, such as the response to @□, are generic and fail to address the original content, indicating a focus on volume over quality. Mentions of the user's own handle (e.g., @□) and repetitive tagging patterns further suggest inorganic behavior.

Behavioral Pattern:
The account exhibits hallmark bot-like behaviors: high tweet volume with low variability, repetitive use of hashtags and narratives, and a focus on retweeting viral or polarizing content. The tweet-to-follow ratio suggests inorganic growth, possibly through mass-following or automation. Mentions of users like @□ and @□ (the target user herself) serve to amplify specific accounts or echo chambers. The lack of personal anecdotes, off-topic tweets, or responses to current events outside the partisan echo chamber further aligns with bot or semi-automated account patterns.

Table 14: The **generated** clues for a social account which is marked as the **Bot**. Some parts are omitted by * and □ for information protection.

Account: u*****9213
{Raw account information} #Omitted for ethic considerations

Clues:
User Information:
☐ is a verified reporter affiliated with ☐ in ☐, with a history spanning six years, over ☐ tweets, and coverage of political, local, and national events. His profile highlights professional ties to media outlets, educational background from ☐, and personal interests like sports fandom (☐, ☐). The account follows a typical journalistic structure, including balanced follower-following ratios (☐ followers, ☐ following) and ☐ listed_count (☐), consistent with specialized regional reporting.

Semantic Feature:
The user's tweets exhibit contextual awareness and topical diversity, including political commentary (e.g., references to the ☐, ☐, and elections) and sports enthusiasm (e.g., ☐ sports teams and ☐). Content includes descriptive details, contextual references (e.g., ☐ starts playing), and reactions to real-world events (e.g., sports achievements), demonstrating natural language fluency and topic-specific knowledge. Tweets lack repetitive or generic structures and show emotional expression (e.g., excitement over a sports event), indicating human-like interests and opinions. Links and media are contextually justified (e.g., news articles, event photos), further supporting authenticity. Mentions of public figures and institutions (e.g., politicians, sports teams) are contextually appropriate and topic-relevant.

Table 15: The **generated** clues for a social account which is marked as the **Human**. Some parts are omitted by * and ☐ for information protection.

Account: u*****9213
{Raw account information} #Omitted for ethic considerations

Clues:
Interactive Situation:
Interactions are predominantly transactional and professional, with frequent mentions of verified accounts, public figures, and niche communities (e.g., political leaders, sports teams). The user retweets from credible sources and engages in context-specific discussions, such as reporting on live events or sharing industry-relevant content. Following accounts like sports teams and religious figures suggests diverse but human-like interests. Mentions of specific individuals (e.g., colleagues, officials) indicate genuine participation in a journalistic network.

Behavioral Pattern:
The account exhibits high-volume tweeting consistent with a journalist covering time-sensitive topics, retweeting credible sources and sharing real-time updates. Followed accounts include verified officials, media outlets, and public figures, reflecting a balanced and purpose-driven curation. Tweets lack spam or repetitive links, and the mix of retweets, replies, and original posts aligns with human variability. The follower-to-following ratio and tweet frequency suggest active, sustained engagement rather than bot-driven automation. Additionally, the account's six-year history with steady activity and topic consistency further supports human authenticity.

Table 16: The **generated** clues for a social account which is marked as the **Human**. Some parts are omitted by * and ☐ for information protection.

D Prompt

Metadata:
Username: xxx,
Userid: xxx,
Location: xxx,
Follower_count: xxx,
Following_count: xxx,
Tweet_count: xxx,
Listed_count: xxx,
Verified: True,
Active_years: xxx years

Description:
xxxxxxx

History Tweets:
The user has the following historical tweets.
Tweet1: xxx
...

Follower:
The information of the followers of the target user is as follows:
xxx
...

Following:
The information of the followings of the target user is as follows:
xxx
...

Mention: The users who was mentioned in the user's historical tweets are as follows:
User1: xxx
...

Table 17: The account sequence template that used for contains available information from the dataset. In the paper, we only extract and concatenate the one-hop neighbor account information as the social relationship.

{Account Sequence}

The user is known to be a {Ground Truth} user. Condense the above user information to form a discriminative logic that can be used to assist in determining this user is a {Ground Truth} user.

Note that you now need to reason and summarize the clues indicating that this account is considered to be a {Ground Truth} based on four dimensions: User Information (examining the basic metadata of the account for obvious patterns or anomalies), Semantic Features (analyzing the textual characteristics of the content posted from various angles), Interactive Situation (summarizing the specific social scenarios or communities in which the account engages and identifying its interests or focus areas), and Behavioral Pattern (extracting behavioral patterns from historical activities to detect any distinctive regularities). Formulate and summary the discriminative logic, and follow the format strictly:

1. user information: (start with this and generated in a paragraph of text, do not add formatting information such as bold)

2. semantic features: (generated in a paragraph of text, do not add formatting information such as bold)

3. interactive situation: (generated in a paragraph of text, do not add formatting information such as bold)

4. behavioral pattern: (generated in a paragraph of text, do not add formatting information such as bold)

Table 18: The prompt contains the concatenated account information and ground truth label, which is presented to most advanced LLM to distill the clues.

You are a helpful assistant. I will provide you with an analysis of four dimensions of a social media account. Based on given analyses, you need to determine whether this account is a human account or a social robot account. These analyses are:
1. basic information: {clue}

2. semantic features: {clue}

3. interactive situation: {clue}

4. behavioral pattern: {clue}

You need to balance these analyses based on your knowledge to reach a conclusion.

Table 19: The fusing prompt that instruct the LLM to make a final prediction.

Given the following information of a social media account, please judge whether the account is a human account or a bot account. The information of this social account is:

{text}

Give your judgment directly with either "Human" or "Bot".

Table 20: The inference prompt that asks the LLM to make a direct prediction.