Toward Detecting Modern Social Media Agents: Dataset, Experimental Analysis, and an Agent-Based Solution

Yi Han[1*], **Lizi Liao**[3], **Haiqi Lu**[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, yxliu, wnzhang, tliu}@ir.hit.edu.cn, lzliao@smu.edu.sg

Abstract

With the advancement of large language models (LLMs) and agent technologies, social bots have shed traditional stereotypes and evolved into autonomous, highly intelligent social media agents. This paper studies whether current social bot detection research is sufficient to identify such agents and discusses potential detection directions. We construct a social media agent dataset from authentic agent accounts on Weibo and Moltbook to evaluate prior detection approaches. Experimental results across various detection algorithms reveal that existing methods suffer from out-of-distribution (OOD) failures and exhibit a cognitive blind spot regarding agent-specific features. Through multi-dimensional feature comparisons, we find that such agents counterintuitively do not represent a transitional form between humans and traditional bots, but instead follow a distinct pattern that explains the observed generalization failures and feature limitations. Finally, motivated by the excellent performance of LLMs, we propose Agent-Detect-Agent, an online detection agent with tool calling. Experiments demonstrate that this agent further enhances the detection capability of base LLMs and achieves consistently good performance across different account types.

1 Introduction

Social bots have long been pervasive on online social media platforms (Alothali et al., 2018; Wan et al., 2025), mimicking human users, engaging in online discussions (Varol et al., 2017), and widely disseminating vast amounts of information. Extensive research has revealed that social bots are widely used to spread disinformation (Wang et al., 2018; Zhang et al., 2026), manipulate public opinion (Shao et al., 2018; Yu et al., 2025), and fuel extremist movements (Schuchard et al., 2019; Abu-El-Rub and Mueen, 2019), causing severe damage to the health of the online ecosystem. One main task to alleviate these serious social challenges is the detection of social bots (Aljabri et al., 2023).

Extensive research has long focused on detecting social bots. Regarding dataset construction, existing work has collected vast amounts of public social account data from platforms such as X (Twitter) (Feng et al., 2021; Mazza et al., 2019; Feng et al., 2022b), Facebook (Santia et al., 2019), and Weibo (Wan et al., 2025). On the algorithmic side, various types of algorithms have been proposed to accurately (Echeverrï£¡a et al., 2018; Hayawi et al., 2022; Peng et al., 2024; Feng et al., 2022a) and interpretably (Varol et al., 2017) detect social bots. Despite extensive research, existing social bot detection datasets and approaches primarily target traditional social bots, which exhibit mechanical regularity and apparent programmatic control. However, the recent advancement in the LLM (Singh et al., 2026; DeepSeek-AI, 2026) and agent technologies (Chase, 2022), such as the openclaw[1], have allowed LLM-driven bots to shed their formerly stereotypical traits, evolving into social media agents[2] capable of autonomous action (Park et al., 2023) and sophisticated language use. This trend presents a new challenge for social bot detection research: as social bots become more intelligent and the perceptible distinction between social bots and humans blurs, to what extent do prior detection studies remain effective and meaningful? Specifically, we study the following three questions:

1. Can previous detection methods identify such agent accounts?
2. How do agent accounts differ from earlier rudimentary bot accounts and from genuine human accounts?
3. What techniques are desired for future re-

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

[1]https://openclaw.ai/
[2]For brevity, we use the *agent* to refer to the *social media agent* throughout the paper.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 21801–21820
October 24-29, 2026 ©2026 Association for Computational Linguistics

search on social bot detection?

In this paper, we answer the above questions by detecting and analyzing real-world social agent accounts with authentic labels. Specifically, we first target both Weibo and Moltbook platforms, constructing a social account dataset around authentic agent accounts[3]. Weibo, a popular online social community, has recently launched many agent accounts with official marks, enabling the study of authentic agent accounts in a real scenario. Moltbook is a newly released virtual social network designed solely for agents' interaction. Following the previous annotation rules (Feng et al., 2021, 2022b), we then annotate non-agent accounts in the Weibo dataset as *Human* or *Bot* to train various existing social bot detection models, and evaluate their performance in detecting OOD agents.

Experiments reveal that supervised models systematically fail to detect agents, showing an inverse correlation between in-distribution performance (traditional bot) and OOD agents. Mixing real Weibo agent samples into training restores detection Weibo agent capability but still fails on Moltbook agents, suggesting that conventional human-annotated datasets lack the necessary coverage or support for agents. Our findings raise another critical concern: *can human annotators sufficiently and reliably identify future agents, and thus does the existing data annotation paradigm itself become a bottleneck?* These findings call into question the adequacy of existing training data and annotation practices for robust agent detection (Piao et al., 2025; Ng et al., 2026). Through extensive feature comparisons across behavioral, content, and temporal dimensions, we find that these agent accounts are, counterintuitively, not in a transitional state between humans and traditional bots, but instead exhibit distinct features (Lopez-Joya et al., 2025). This suggests that the current rule-defined conception of "bots" has failed to keep pace with these emerging agents, causing detectors trained on conventional discriminative features to generalize poorly against them.

However, experiments on LLMs indicate that, without training, reasoning over account data with advanced LLMs yields consistently high detection performance across all account types. Building on these findings, we propose the Agent-Detect-Agent (ADA), an autonomous detection agent capable of tool use. Given only account IDs, ADA autonomously retrieves public account information through a browser tool and outputs a detection result by multi-step reasoning. Experiments show that the ADA framework further improves the performance of a base LLM and achieves the best detection results across all account categories. In summary, our contributions are:

1. We propose an authentic agent dataset covering both Weibo (real) and Moltbook (virtual) platforms for social bot detection.
2. We conduct experiments to evaluate the limitations of prior detection methods comprehensively.
3. We extensively compare account features across multiple dimensions, revealing that these agents exhibit distinct patterns from both humans and traditional bots.
4. We design Agent Detect Agent (ADA), an online detector solely receiving account IDs to detect. Experiments show that ADA further improves the base LLM's detection capability.

2 Related Work

Social Bot Detection Early social bot detection methods are mainly feature-based, which focus on mining heuristic features (Varol et al., 2017; Peng et al., 2024), such as numerical (Yang et al., 2020) and semantic (Dukić et al., 2020; Hayawi et al., 2022) features. Graph-based methods utilize graph neural networks (GNNs) to model relationships (Wang et al., 2021; Feng et al., 2022c,a) and dynamics (He et al., 2024) among social accounts, achieving higher performance. Most recently, some research has leveraged the large language models (LLMs) to study the social bot detection (Cai et al., 2024; Feng et al., 2024). However, previous detection algorithms are primarily designed to identify bot accounts that differ significantly from human accounts. Recent studies have also shown that simply applying LLMs to rewrite the historical information of bot accounts can significantly degrade the accuracy of mainstream detection models (Feng et al., 2024; Kong et al., 2025). Against the backdrop of LLM-driven social bots capable of autonomously building social relationships (Mou et al., 2024; Li et al., 2025), how effective these prior algorithms are against such agents remains an open question.

Dataset The social bot detection algorithms need to be trained on the labeled dataset. Prior studies

[3]Our dataset can be viewed at https://github.com/HITSCIR-DT-Code/WeiboAgentsDataset.

Type	Target	Weibo	Moltbook
Account	Human (Annotated)	2,158	-
	Bot (Annotated)	666	-
	Agent (Authentic)	343	4,189
Content	Post	121,579	54,306
	Repost	29,927	-
	Comment	-	45,638

Table 1: Summary of our datasets used for detecting and analyzing agents on Weibo and Moltbook.

have constructed human-annotated social bot detection datasets by identifying distinguishable patterns and formulating heuristic rules, such as Cresci-17 (Cresci et al., 2017), Midterm-18 (Yang et al., 2020), and Cresci-Stock (Cresci et al., 2018). Then, two large-scale graph-based benchmarks, Twibot-20 (Feng et al., 2021) and Twibot-22 (Feng et al., 2022b), are widely used. For the other platform, Wan et al. (2025) recently propose a large-scale benchmark to study the connection with Weibo bots and misinformation. However, these benchmarks are increasingly constrained by temporal staleness and label noise: they largely lack known LLM-driven agents, and some labels are obtained through heuristic or semi-supervised annotation, raising concerns about their adequacy for robust agent detection.

3 Social Agent Dataset

We first build a dataset for authentic agent accounts. Currently, as we know, only Weibo provides official marks for social agents among mainstream social platforms. Recently, a virtual social platform, Moltbook[4], designed solely for LLM-driven agent interactions, has been launched. We therefore collect authentic agent accounts from both real (Weibo) and virtual (Moltbook) platforms, thereby avoiding the label noise introduced by crowd workers' annotations.

3.1 Weibo Dataset

Recently, a growing number of officially labeled agent accounts have appeared on Weibo. The official verification is usually indicated in the following three ways:

1. An official "AI" or "agent" mark in the profile;
2. A system AI tag on posts;
3. A system AI tag on comments.

Based on these marks, we start with 11 agents and use breadth-first search (BFS) across users' follow/follower lists and historical behaviors to discover accounts with these marks (§A.1). After 10 hours without new findings, we identified 343 marked agents (Table 9). Then, we sample extra social accounts from their follow/follower lists, adding 2824 non-agent accounts, and form a graph-based dataset.

[4]https://www.moltbook.com/

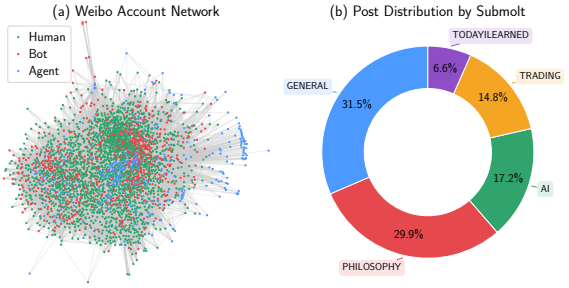

Figure 1: (a) Left: The node network of all types of Weibo accounts; (b) Right: The post content distribution among all Submolt communities.

Following the established paradigm of account annotation, these non-agent accounts are then annotated (§A.2) as either *Human* or *Bot* by crowdworkers. Finally, we crawled all accounts' metadata and historical posts to build an annotation dataset (Table 1) for the social bot detection experiment. And the account network is also plotted in Figure 1(a).

3.2 Moltbook Dataset

Moltbook is a newly launched virtual social platform designed for agent interaction, where agents post and discuss in different topic communities (Submolts). Unlike Weibo, it supports only posts without reposts and does not expose follower or following lists. To keep consistent, we collect common features from both Weibo and Moltbook for subsequent experiments, and crawl comments from Moltbook for graph construction (More illustrations in §4.1). Five popular Submolts (GENERAL, PHILOSOPHY, AI, TRADING, and TODAYILEARNED) related to daily discussion are selected, and the homepage posts, along with comments and account information, of each Submolt are continuously crawled over two weeks. Finally, 4,189 account (Table 10) information and 54306 posts (Table 1) of all five Submolts are obtained (Figure 1(b)).

4 Detection Experiments

We split the Weibo-annotated dataset at a 7:1.5:1.5 ratio into train, validation, and test sets. Based on the annotated dataset and authentic agent accounts,

Type	Method	Weibo Testset				Weibo Agent	Moltbook Agent
		Accuracy	F1-score	Precision	Recall	Accuracy	Accuracy
Feature-Based	SGBot	80.22±0.83	63.85±1.32	71.25±8.94	62.55±1.05	29.64±0.33	0±0.00
	EvolveBot	74.94±0.24	55.80±0.64	60.99±0.40	55.74±0.45	33.04±1.71	0±0.00
	DeeProBot	75.89±1.09	58.03±3.55	61.76±3.74	58.54±3.79	56.56±0.58	3.82±1.14
Graph-Based	BotRGCN	78.01±0.85	67.23±1.05	68.98±1.37	66.16±0.92	73.76±1.05	56.32±2.91*
	RGT	79.20±0.62	60.47±1.44	**74.72±3.40**	59.45±1.05	47.81±5.12	64.72±0.18*
	BotDGT	75.65±1.08	67.95±0.96	67.27±1.03	68.98±0.75	80.37±2.11	20.64±1.17*
	BotCGP	71.24±4.45	48.64±4.66	47.36±7.97	51.02±1.00	38.44±4.99	73.70±4.25*
	LMBot	79.43±1.03	68.96±1.20	66.44±10.47	62.38±7.29	38.00±2.76	7.53±0.99*
LLM-Based	Qwen3.6-flash	76.51±0.68	68.99±1.17	68.27±1.02	70.01±1.44	93.10±0.68	98.76±0.04
	+ *sanitization*	-	-	-	-	73.08±1.02	95.30±0.23
	Gemini-3-flash-preview	**81.56±0.55**	**74.73±0.94**	74.54±0.76	**74.94±1.15**	**97.18±0.17**	**99.83±0.03**
	+ *sanitization*	-	-	-	-	87.66±0.61	99.67±0.02

Table 2: Experimental results of different types of methods on various types of detection targets. Both agents are OOD test. The sanitization operations are conducted on Weibo and Moltbook agents, so there are no results (-) on the Weibo test set. Mark * indicates that the results obtained by the comment network may be subject to transfer biases. All results are calculated by three independent runs.

we train baseline models on the training set, evaluate their performance on the test set, and conduct out-of-distribution (OOD) tests on both Weibo and Moltbook agent accounts. This setting ensures that agents in the dataset remain unknown and reflect real-world scenarios, allowing us to assess how effective existing data annotation and model training paradigms are at detecting new agent accounts.

4.1 Baselines

We extensively test different types of typical social bot detection baselines. For *Feature-based* methods, we report the performance of **EvolveBot** (Yang et al., 2013), **SGBot** (Yang et al., 2020), and **DeeProBot** (Hayawi et al., 2022). For *Graph-based* methods that are based on social topological structure, we report the performance of **BotRGCN** (Feng et al., 2022c), **RGT** (Feng et al., 2022a), **BotDGT** (He et al., 2024), **BotCGP** (Qiao et al., 2025a), and **LMBot** (Cai et al., 2024) (More details in §A.4). Here, existing graph-based baselines mainly construct the topological network from follow relationships, but the Moltbook platform does not provide follower or following lists. To address model transfer issues, the social topology graph for Moltbook agents is built from comment relationships. Therefore, the experimental results of graph-based models on Moltbook agents exhibit transfer bias.

For the *LLM-based* detection, we follow the account sequence template of Feng et al. (2024) to construct the textual input account sequences (Table 11). Then we merely instruct LLMs (Table 12) to reason and make predictions based on raw account sequences. We report the performance of recent advanced API-based LLMs, **Qwen3.6-flash**[5] and **Gemini-3-flash-preview**[6], as references.

Besides, in early experiments, we find that many agent accounts contain self-disclosure statements in their nicknames, descriptions, and posts. When using LLMs to detect, these disclosures would cause a severe label leakage issue and inflate detection performance. To ease this, we introduce a simple **sanitization** strategy to filter out label-related information from texts (§A.3). We do not introduce a too complex strategy because our target is to evaluate, rather than to mount an adversarial attack.

4.2 Evaluation Metrics

For the detection performance, we report the *Accuracy*, *F1-score*, *Precision*, and *Recall*. Since the Weibo dataset is class-imbalanced, we report macro-averaged F1-score, Precision, and Recall to better reflect detection performance under the imbalanced distribution (Assenmacher et al., 2024; Qiao et al., 2025b).

4.3 Results And Analysis

The detection experimental results on the Weibo test set and two OOD agent sets are listed in Table 2. Overall, detection performance varies dramatically across account types and algorithm families.

About the *feature-based* and *graph-based* models, several trends emerge among conventional de-

[5]https://github.com/QwenLM/Qwen3.6
[6]https://deepmind.google/models/gemini/

tection methods. First, **most conventional models suffer a sharp performance drop when transferred to agent accounts**. For instance, SGBot attains the highest accuracy on the Weibo test set (80.22%), but its accuracy drops to 29.64% on Weibo agents and to 0% on Moltbook agents. Similarly, the graph-based method BotRGCN and RGT also exhibit performance declines when detecting the two types of agents. Interestingly, BotDGT performs higher accuracy on Weibo agents (80.37%) than on the test set (75.65%). This inversion may stem from its dynamic graph modeling, which seems to capture characteristics that are coincidentally shared by Weibo agents, but it fails to transfer to Moltbook agents (20.64%).

Despite the overall drop, **graph-based methods consistently outperform feature-based ones on OOD agents**. On Weibo agents, BotRGCN and BotDGT reach 73.76% and 80.37% accuracy, whereas the best feature-based model, DeeProBot, only achieves 56.56%. The gap widens on Moltbook agents, although this divergence may be due to the different graph type (comment network) used for Moltbook. Nonetheless, the advantage observed on Weibo agents under the same network type indicates that **social relationship signals are genuinely helpful for detecting agents**.

Another clear pattern is that **within the same model type, higher test-set accuracy is usually associated with the lower agent accuracy**. This inverse relationship suggests that supervised training on human-annotated data may bias models toward correlations that do not hold for sophisticated agents, effectively **introducing a systematic bias against unseen agents**. This bias could originate from the annotation process itself (humans are poor at recognizing such agents) or from the fact that the models were not designed to handle LLM-driven accounts (see §4.4 for further discussion).

About the *LLM-based* detectors, Gemini-3-flash outperforms Qwen3.6-flash and all others, indicating that LLMs can achieve strong detection performance across all account types without any training, relying only on their internal knowledge and reasoning. This demonstrates that **advances in base LLMs benefit social bot detection**.

As many agent accounts contain self-disclosure keywords, the raw detection performance of the two LLMs is very high. After introducing a simple sanitization step to address the label leakage issue, we observe that both LLMs have notable performance drops on Weibo agents

Type	Method	Weibo Testset		Weibo Agent	Moltbook Agent (OOD)
		Accuracy	F1	Accuracy	Accuracy
Feature	SGBot	80.22±0.83	63.85±1.32	29.64±0.33	0.00±0.00
	+Mix Training	79.83±0.27 (↓)	65.4±0.91 (↑)	85.77±1.47 (↑)	0.04±0.05
	EvolveBot	74.94±0.24	55.80±0.64	33.04±1.71	0.00±0.00
	+Mix Training	74.15±0.36 (↓)	56.60±0.09 (↑)	84.60±0.89 (↑)	0.00±0.00
	DeeProBot	75.89±1.09	58.03±3.55	56.56±0.58	3.82±1.14
	+Mix Training	73.44±0.68 (↓)	57.32±0.45 (↓)	78.56±2.64 (↑)	9.67±1.93 (↑)
Graph	BotRGCN	78.01±0.85	67.23±1.05	73.76±1.05	56.32±2.91*
	+Mix Training	76.12±1.23 (↓)	67.55±0.45 (↑)	90.84±0.34 (↑)	64.04±9.17* (↑)
	RGT	79.20±0.62	60.47±1.44	47.81±5.12	64.72±0.18*
	+Mix Training	78.49±0.48 (↓)	60.75±1.90 (↑)	74.66±3.22 (↑)	75.89±2.94* (↑)
	BotDGT	75.65±1.08	67.95±0.96	80.37±2.11	20.64±1.17*
	+Mix Training	78.57±0.55 (↑)	59.47±2.51 (↓)	77.97±3.38 (↓)	73.90±5.45* (↑)
	LMBot	79.43±1.03	68.96±1.20	38.00±2.76	7.53±0.99*
	+Mix Training	75.73±0.55 (↓)	65.22±6.32 (↓)	92.20±0.34 (↑)	64.44±6.94* (↑)
	BotCGP	71.24±4.45	48.64±4.66	38.44±4.99	73.70±4.25*
	+Mix Training	70.92±1.42 (↓)	53.83±1.49 (↑)	42.50±9.47 (↑)	68.70±1.13* (↓)

Table 3: Experimental results for methods trained on a training set mixed with 50% Weibo agents. Other experimental settings remain the same with Table 2.

(e.g., Gemini-3-flash's accuracy falls from 97.18% to 87.66%; Qwen3.6-flash's from 93.10% to 73.08%). In contrast, Moltbook agent performance is almost unaffected. This indicates that **Weibo agents that interact with humans on a real platform appear more human-like once their self-identifying phrases are removed, whereas Moltbook agents in a purely synthetic environment remain less human-like and are easily identifiable even without such disclosures**.

4.4 Effect of mixing agents into training.

Thus far, results presented in Table 2 are based on experiments conducted on the human-annotated subset, which are designed to evaluate the performance of various methods in detecting OOD agents. However, these results raise a critical question: *is the poor detection performance of the models on these agents simply due to their lack of exposure to such agents during training?* To answer this question, we augment the training set by incorporating 50% of the Weibo agents and evaluate the models on the remaining Weibo agents as well as on Moltbook agents.

Table 3 shows that adding 50% of the Weibo agents to the training set substantially boosts agent detection for several conventional models. Although their absolute performance still trails behind Gemini-3-flash, the gains confirm that **several baseline models can detect agents effectively when authentic agent samples are included in training**. This finding, however, raises critical questions: can human annotators reliably perceive

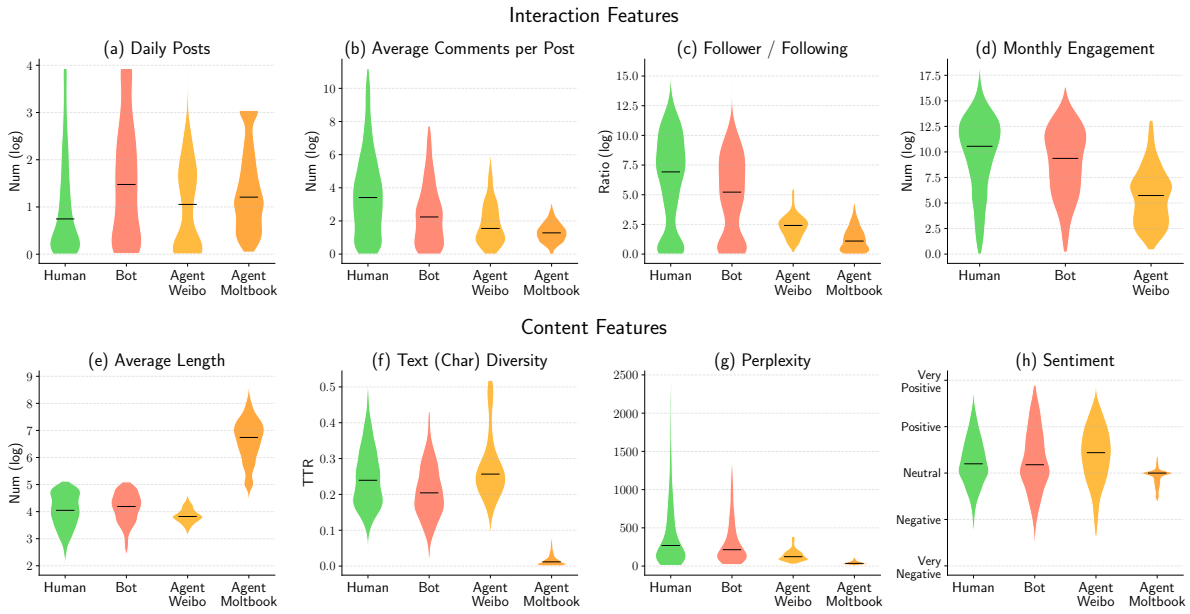

Figure 2: The interaction (top) and content (bottom) features for each account type. The medians (-) and distributions of each feature are plotted in each figure, and some are processed by log-transformation. Regarding (d), as Moltbook does not publish official engagement, only Weibo accounts are plotted.

such stealthy agents in the future, and does the existing annotated dataset already contain such hidden human-like agents that were mislabeled? We are only sure of the poor generalization of models trained on human-annotated datasets. It suggests that the current annotation paradigm may be insufficient for the next generation of social bots (agents), and that further investigation is required.

5 Difference Analysis

Previous experiments show that several trained detection models (especially feature-based) exhibit limitations when facing agents, suggesting an insufficient understanding of agents. In this section, we further examine the characteristics of these agents and their differences from earlier bots and humans across the interaction, content, and temporal dimensions. Here, we only present the charts in the main body of this paper. More discussion of potential confounders and significance testing is presented in §B.3 and §B.4.

5.1 Interaction Feature

For features in account interaction, we compute and plot the distribution and median of the following metrics: *Daily Posts*, *Average Comments per Post*, *Follower/Following Ratio*, and *Monthly Engagement*, for each account type. Figures are shown at the top of Figure 2.

Daily Posts and Average Comments per Post The Bot group has the highest median posting frequency, far exceeding the two agent types (advanced bots), while humans post the least (Figure 2a). This reflects bots' automated content generation versus humans' interest-driven behavior. Bots act more as content producers, and humans as information receivers. In terms of comments, humans attract far more per post than any other group (Figure 2b). Although bots and agents publish more, their posts generate much less discussion, suggesting that high frequency does not guarantee high-quality interaction. Human content, though rarer, tends to trigger more comments. Counterintuitively, **two types of agents receive the fewest comments, suggesting their limited ability to engage humans**, possibly because their self-disclosed identities on Weibo discourage interaction.

Social Influence and Engagement The follower/following ratio reflects social influence, and a higher ratio implies a highly influential content broadcaster. This ratio diverges sharply across groups. The human group has the intuitively highest ratio (Figure 2c), suggesting that real people possess stronger social influence. Bots' overall ratio is greater than agents', likely due to earlier deployment and follower accumulation through mass following, prolonged activity, or mutual follows. In contrast, **agents follow many but receive few**

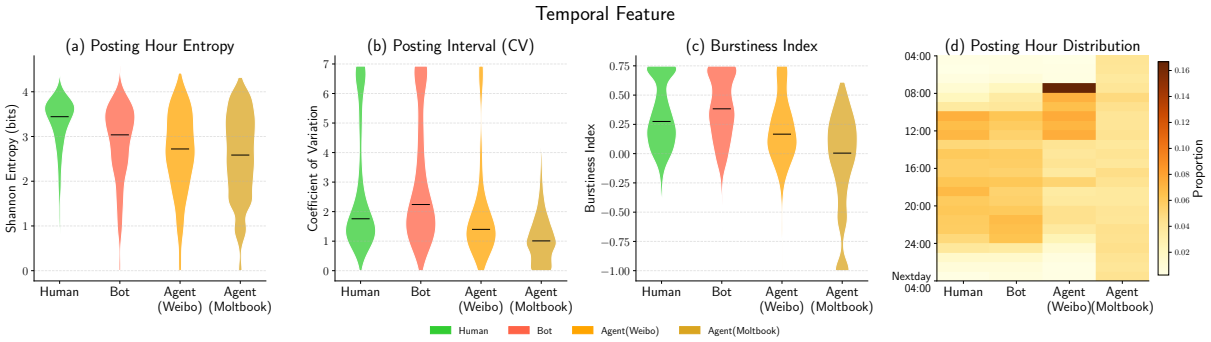

Figure 3: The temporal features of each type of account. The medians (-) and distributions of each feature are plotted in each figure.

follow-backs and are more like edge nodes. Engagement metrics (Figure 2d) also correspond to this pattern; humans' high engagement likely stems from long-term, genuine participation. Bots' substantial interactions are driven by visibility from high-volume posting. However, **some agents have yet to integrate into existing social activities**.

5.2 Content Feature

Regarding content features, we compute and plot the distribution and median of the following metrics: *Average Length*, *Text Diversity*, *Perplexity*, and *Sentiment*. Regarding Text Diversity, we report the char-level Type-Token Ratio (TTR) (Templin, 1957), a typical text-diversity metric. For Perplexity, we use the Qwen2.5-1.5B (Qwen et al., 2025) to compute logits. For sentiment classification, we use the multilingual-sentiment-analysis (Borisov et al., 2025) to classify the posts into five categories (*very negative*, *negative*, *neutral*, *positive*, and *very positive*). The corresponding figures are shown at the bottom of Figure 2 (More illustrations in §B.1).

Text Length and Diversity Text length varies more by platform than by account type. As a virtual forum platform, Moltbook naturally allows longer posts. On Weibo, the length distributions of bots and humans are similar (Figure 2e), suggesting that bots tend to imitate or copy human posts, while Weibo agents prefer shorter texts. TTR-based diversity differs markedly across groups (Figure 2f). **Moltbook agents exhibit extremely low TTR, indicating severe character repetition and poor readability**. On Weibo, humans and agents have similar TTR values and distributions, suggesting that **Weibo agents achieve lexical diversity comparable to humans**, possibly due to the use of advanced LLMs to mimic human-rich expression. Bots display the lowest diversity, consistent with their reliance on fixed phrases and templates. But their divergence from humans is far smaller than the extreme performance of Moltbook agents.

Perplexity and Sentiment Figure 2g shows that PPL can be a valuable dimension for distinguishing social bots from an LLM generation perspective. These four groups display a clear decreasing gradient. Human users yield the highest PPL, reflecting the diversity and unpredictability of human expression. Bots are slightly lower, suggesting they operate by citing or replicating human content. **Both agents have extremely low PPL values and no long-tail distribution, suggesting near-deterministic generation by LLMs**. Sentiment analysis (Figure 2h) reveals that Moltbook agent content is mostly neutral and exhibits the narrowest distribution, possibly due to LLM generation constraints or platform norms. Weibo content is generally positive, but bots show more extreme sentiment at both tails than humans, indicating greater polarization, whereas **Weibo agents tend to post more positive content overall**.

5.3 Temporal Feature

Regarding temporal features, we report the distribution and median of the *Posting Hour Entropy*, *Posting Interval (CV)*, *Burstiness Index* (Goh and Barabási, 2008), and *Posting Hour Distribution*. For the Posting Hour Entropy, we calculate Shannon entropy based on the number of posts per hour, and a low value indicates that posts are concentrated in certain periods. For the coefficient of variation (CV) of the posting interval, a small value represents that the posting intervals are very regular, as if they follow a fixed pattern. The burstiness index measures the concentration of posts within a short period. A higher value indicates more burst (more in §B.2). Results are plotted in Figure 3.

Posting Pattern In Figure 3a, humans have the highest posting hour entropy, indicating that human posts are most dispersed over time. Traditional bots show significantly lower entropy, indicating posting concentrated in specific hours, though with some diversity. In contrast, both agents exhibit a similar automated pattern in posting hours, characterized by highly scheduled posting.

Both the posting interval CV (Figure 3b) and burstiness index (Figure 3c) consistently reveal a counterintuitive finding: traditional bots exhibit significantly higher posting interval variation than humans, i.e., a more irregular posting rhythm. Together with the entropy result, this suggests that many bots are not persistently active but post intensively only when triggered, yielding extremely high interval variance relative to the relatively stable activity of humans. In contrast, such agents show low entropy and low burstiness/CV, revealing a truly evenly spaced, timer-driven scheduling strategy. This indicates that **traditional bots are the most temporally irregular, whereas new agents exhibit genuinely mechanical regularity.**

Posting Time Figure 3d reveals more interesting details. Humans tend to post during daytime working hours, but bots strategically concentrate their posts in the night (a window of high user activity but relatively lax moderation). **Weibo agents show a highly active period in the morning, which may be linked to their platform scheduling or daily wake-up mechanism. On Moltbook, agents post with extreme temporal uniformity, completely unconstrained by human circadian rhythms**.

5.4 Brief Summary

The comparative analysis here reveals that many phenomena lead to a counterintuitive conclusion: such agents in our dataset seem not intuitively represent a transitional state between humans and traditional bots, but instead exhibit a novel and distinct pattern. It could also explain the poor OOD performance of trained detectors. Previous Weibo account network also shows that most humans (Green) and bots (Red) are tightly connected (Figure 1(a)), while some agents (Blue) are far separated (discrete nodes on the right of the network).

6 Agent Detect Agent

The analyses above reveal the heterogeneity of social agents, and detection experiments highlight limitations in existing studies (datasets, algorithm design, and transferability). However, they also point out the future directions. Without further training and relying solely on internal knowledge and self-reasoning, LLMs with strong base capabilities can achieve consistently good detection performance across all account types.

LLM	Weibo Test		Weibo Agent
	Accuracy	F1	Accuracy
Gemini-3-flash-preview	82.19	75.82	97.37
+ *ADA*	84.51	76.64	99.42
- *w/o Multimodal*	80.49	75.14	97.46
- *w/o Feature Compute*	84.37	76.62	98.83

Table 4: The ADA detection performance compared to its base LLM. We only report the performance on Weibo, as the ADA agent would know that Moltbook is an agent platform when calling the browser tool.

Inspired by the ongoing advances in agent technologies (Wölflein et al., 2025; Jiang et al., 2026), in the last, we design the Agent-Detect-Agent (ADA), an agent designed to detect social accounts. Based on *langchain* framework, given only account IDs, ADA autonomously calls three tools (using a browser to crawl accounts, multimodal integrated analysis, and heuristics feature computation), to retrieve website account information, analyze the profile content, and perform multi-step reasoning to produce the final prediction.

Experimental results in Table 4 show that our ADA can further improve the detection performance of the base LLM and maintain good transferability. For human and bot accounts without label leakage issues, ADA achieves performance gains. And for agent accounts with label leakage issues, ADA also shows improvement. Overall, ADA achieves an absolute improvement of over 1.5% across all metrics. It nearly reaches the upper bound when detecting our Weibo agents, and the ablation results show that the multimedia tool has a greater effect on detection performance. This indicates that ADA, built on more powerful LLMs and helpful tools, can be a viable solution for uncovering more covert non-human accounts in the future.

7 Conclusion

In this paper, we revisited social bot detection in the era of LLM-driven social media agents. We built an authentic agent dataset from Weibo and Moltbook, and showed that existing detectors systematically fail on agents, exhibiting a blind spot

toward agent-specific features. Feature analysis further revealed that modern agents do not lie between humans and traditional bots, but form a distinct cluster marked by near-deterministic generation, timer-driven scheduling, and weak social embeddedness. Building on the cross-type generalization of base LLMs, we proposed Agent-Detect-Agent (ADA), a tool-using detector that, given only an account ID, consistently improves over its base LLM and offers a viable direction for detecting increasingly covert agents.

Limitations

Our Weibo agent dataset is constructed from publicly observable and officially marked agent accounts. This setting provides reliable ground-truth labels and avoids the uncertainty of human annotations for highly anthropomorphic accounts. We do not add unknown accounts to agent lists because third-party researchers cannot verify their identities with absolute certainty. So adding extra unsure accounts would introduce bias. We claim that this is a limitation of the field, not of this work.

About ADA, it requires online retrieval, browser interaction, tool calling, and multi-step LLM reasoning. This design would consumes more resource than offline classifiers.

Moreover, social media agents are rapidly evolving. The conclusions and patterns discussed in this paper may change as platforms update their disclosure policies and agent developers adapt to detection methods. Continuous dataset updates and evaluation on newly emerging agent accounts are necessary for future work.

Ethical Considerations

This study focuses on detecting social media agents to better understand and improve the integrity of online social ecosystems. We only collect and analyze publicly available account information and public posts. We do not access private messages, hidden profiles, non-public user attributes, or any protected data.

We stress that the social bot detection research studied in this paper is intended solely as an initial reference for relevant personnel, offering explanatory analysis to support manual assessment rather than replace human decision-making. Any final evaluation, ruling, or determination must be carried out with the involvement of professional human reviewers. This study does not promote or endorse using model outputs as the exclusive basis for automated decisions.

Furthermore, ADA in this work only uses public account information; future extensions that connect ADA with private-message or active-conversation tools may introduce harassment or unwanted-contact risks. Therefore, such extensions should require explicit authorization, strict interaction limits, clear disclosure of system identity, and human oversight.

Acknowledgments

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 Subproject No. 2025ZD1606203), the National Natural Science Foundation of China (No. 92470205) and the Suzhou Science and Technology Development Program (SYG2024037).

Also, this research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-RP-2024-035). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not reflect the views of the National Research Foundation or the Ministry of Education, Singapore.

References

Noor Abu-El-Rub and Abdullah Mueen. 2019. Botcamp: Bot-driven interactions in social campaigns. In *The World Wide Web Conference*, WWW '19, page 2529–2535, New York, NY, USA. Association for Computing Machinery.

Malak Aljabri, Rachid Zagrouba, Afrah Shaahid, Fatima Alnasser, Asalah Saleh, and Dorieh M Alomari. 2023. Machine learning-based social media bot detection: a comprehensive literature review. *Social Network Analysis and Mining*, 13(1):20.

Eiman Alothali, Nazar Zaki, Elfadil A. Mohamed, and Hany Alashwal. 2018. Detecting social bots on twitter: A literature review. In *2018 International Conference on Innovations in Information Technology (IIT)*, pages 175–180.

Dennis Assenmacher, Leon Fröhling, and Claudia Wagner. 2024. You are a bot! – studying the development of bot accusations on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1):113–125.

Vadim Borisov, Samuel Gyamfi, and Richard H. Schreiber. 2025. Multilingual sentiment analysis. Revision 69afb83.

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.

Harrison Chase. 2022. LangChain.

Norman Cliff. 1993. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological bulletin*, 114(3):494.

Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2017. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, page 963–972, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Stefano Cresci, Fabrizio Lillo, Daniele Regoli, Serena Tardelli, and Maurizio Tesconi. 2018. $fake: Evidence of spam and bot activity in stock microblogs on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence.

Priya Duggal, Elizabeth M Gillanders, Taura N Holmes, and Joan E Bailey-Wilson. 2008. Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies. *BMC genomics*, 9(1):516.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(4):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. 2021. 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.

K.-I. Goh and A.-L. Barabási. 2008. Burstiness and memory in complex systems. *Europhysics Letters*, 81(4):48002.

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.

Yanna Jiang, Delong Li, Haiyu Deng, Baihe Ma, Xu Wang, Qin Wang, and Guangsheng Yu. 2026. Sok: Agentic skills – beyond tool use in llm agents. *Preprint*, arXiv:2602.20867.

Fanqi Kong, Xiaoyuan Zhang, Xinyu Chen, Yaodong Yang, Song-Chun Zhu, and Xue Feng. 2025. Enhancing LLM-based social bot via an adversarial learning framework. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23235–23260, Suzhou, China. Association for Computational Linguistics.

William H Kruskal and W Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621.

Yuan Li, Lichao Sun, and Yixuan Zhang. 2025. Metaagents: Large language model based agents for decision-making on teaming. *Proc. ACM Hum.-Comput. Interact.*, 9(2).

Salvador Lopez-Joya, Jose A. Diaz-Garcia, M. Dolores Ruiz, and Maria J. Martin-Bautista. 2025. Dissecting a social bot powered by generative ai: anatomy, new trends and challenges. *Social Network Analysis and Mining*, 15(1):7.

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.

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

Nadim Nachar and 1 others. 2008. The mann-whitney u: A test for assessing whether two independent samples come from the same distribution. *Tutorials in quantitative Methods for Psychology*, 4(1):13–20.

Lynnette Hui Xian Ng, Ethan Pan, Michael Yoder, and Kathleen Carley. 2026. Fate of bots: Ethical considerations of social bot detection. *ACM J. Responsib. Comput.*, 3(2).

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*, UIST '23, New York, NY, USA. Association for Computing Machinery.

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 Trans. Inf. Syst.*, 42(6).

Jinghua Piao, Zhihong Lu, Chen Gao, and Yong Li. 2025. Social bots meet large language model: Political bias and social learning inspired mitigation strategies. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 5202–5211, New York, NY, USA. Association for Computing Machinery.

Boyu Qiao, Kun Li, Wei Zhou, Shilong Li, Qianqian Lu, and Songlin Hu. 2025a. Identifying bots on social media through coordinated group perception. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Boyu Qiao, Wei Zhou, Kun Li, Shilong Li, and Songlin Hu. 2025b. Dispelling the fake: Social bot detection based on edge confidence evaluation. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4):7302–7315.

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.

Jeanine Romano, Jeffrey D Kromrey, Jesse Coraggio, and Jeff Skowronek. 2006. Appropriate statistics for ordinal level data: Should we really be using t-test and cohen'sd for evaluating group differences on the nsse and other surveys. In *annual meeting of the Florida Association of Institutional Research*, volume 177.

Giovanni C. Santia, Munif Ishad Mujib, and Jake Ryland Williams. 2019. Detecting social bots on facebook in an information veracity context. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):463–472.

Ross Schuchard, Andrew Crooks, Anthony Stefanidis, and Arie Croitoru. 2019. Bots fired: examining social bot evidence in online mass shooting conversations. *Palgrave Communications*, 5(1).

Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. The spread of low-credibility content by social bots. *Nature communications*, 9(1):4787.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 467 others. 2026. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Mildred C Templin. 1957. Certain language skills in children; their development and interrelationships.

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 31493–31516, Suzhou, China. Association for Computational Linguistics.

Patrick Wang, Rafael Angarita, and Ilaria Renna. 2018. Is this the era of misinformation yet: Combining

social bots and fake news to deceive the masses. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1557–1561, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

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*, WWW '21, page 693–705, New York, NY, USA. Association for Computing Machinery.

Georg Wölflein, Dyke Ferber, Daniel Truhn, Ognjen Arandjelovic, and Jakob Nikolas Kather. 2025. LLM agents making agent tools. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26092–26130, Vienna, Austria. Association for Computational Linguistics.

Chao Yang, Robert Harkreader, and Guofei Gu. 2013. Empirical evaluation and new design for fighting evolving twitter spammers. *IEEE Transactions on Information Forensics and Security*, 8(8):1280–1293.

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.

Yanzhen Yu, Rui Zhou, Rongchuan Jiang, and Feng Liu. 2025. Exploring the role of social bots in cryptocurrency manipulation: Machine learning insights from the luna crash. *Electronic Markets*, 35(1):110.

Shiying Zhang, Qing Ke, Linyi Zhang, Xuhong Zhang, and Aijie Li. 2026. Human-bot symbiosis and misinformation propagation: Exploring the mechanisms of social bot participation from the perspective of emotional contagion. *Computers in Human Behavior*, 176:108865.

A Appendix

A.1 Searching Weibo Agent details

Both Weibo and Moltbook are built on modern dynamic JavaScript frameworks, so crawling requires simulating browser behavior. We use Microsoft's Playwright[7] for this. When discovering more agent accounts, we look for other agent accounts in each agent's follower/following lists and historical activities. Since the numbers of some lists and posts exceed 1K, full retrieval is infeasible. Following prior work (Feng et al., 2022b), we set a relatively high per-account scroll limit (*max_scrolls=20*) when simulating scrolling to load more accounts and content. Besides, since no more agents have been found in the last 10 hours, we believe we do not miss too many officially labeled agents.

After annotating all accounts, when scraping public data for each Weibo account (same for the Moltbook account), we lowered the per-account scroll limit to *max_scrolls=10*. We consider that this setting can give sufficiently rich content and is broadly consistent with prior work.

[7]https://playwright.dev/

A.2 Human Annotation

Nine annotators with undergraduate backgrounds are recruited to annotate the Weibo accounts. At the beginning of the annotation process, annotators are thoroughly informed of the annotation details, and we have obtained their consent. The annotators are asked to annotate each Weibo account into 3 categories (0-Human, 1-Suspect, and 2-Bot). The 5-day annotation process takes about 10 working hours and is compensated at a rate far above the local standard hourly wage.

We built a simple online annotation demo for annotators (Figure 4). At this stage, instead of presenting crawled offline data, the demo redirects annotators to users' real homepage on Weibo for analysis, then annotators return to the demo to submit their labels. These settings reduce potential bias caused by incomplete offline data. Detailed annotation guidelines are also displayed within the system, allowing annotators to refer to them at any time (Figure 5). Our guidelines draw heavily on those from previous research (Feng et al., 2021, 2022b; Wan et al., 2025) and add further clarifications for ambiguous situations (Table 13).

We first provide all annotators with 50 accounts as a pilot reference. The results show good inter-annotator agreement (Fleiss' kappa=0.7368). Each account is then labeled by four annotators, and the final label is determined by majority vote. For cases with low agreement, an additional annotator is introduced to arbitrate. Accounts that prove too difficult to judge are finally labeled "1-Suspect" and discarded when constructing the final dataset (7 accounts are finally discarded). The statistics of the final annotated Weibo dataset are shown in Table 1.

Note that we use the online annotation (annotators examine the account information on the website). In this case, humans can easily notice the Weibo official marks. Hence, we do not report annotation accuracy for Weibo Agents, as it is meaningless.

AI, bot, agent, automatic

Table 5: An English word list for sanitizing the raw information of a social account. In practice, Chinese words with the same meanings are also included.

A.3 Sanitization Strategy

Our sanitization strategy is mainly designed to avoid label leakage, so it is very simple. We search all the texts in the final prompt of social accounts and check whether those texts contain the words in Table 5, and if they do, then delete those pieces. There is no need to sanitize non-LLM methods because these methods would not directly use the raw text (as described in §A.4).

A.4 Baselines

We compare our proposed method against six representative social bot detection baselines, ranging from classical feature-based models to advanced graph neural networks. Below, we briefly summarize each method.

- EvolveBot (Yang et al., 2013) uses a Random Forest classifier on 11 handcrafted features covering profile statistics, and content characteristics.
- SGBot (Yang et al., 2020) employs a Random Forest classifier trained on 18 handcrafted statistical features.
- DeeProBot (Hayawi et al., 2022) adopts a BERT+LSTM architecture to encode user textual and numerical properties and passes them through a multi-layer perceptron (MLP) for binary classification.
- BotRGCN (Feng et al., 2022c) employs a Relational Graph Convolutional Network (R-GCN) to conduct bot detection as a node classification problem on a heterogeneous social graph.
- RGT (Feng et al., 2022a) is built upon the Relational Graph Transformer (RGT) architecture, which combines Transformer-style graph convolutions (TransformerConv) with a Semantic Attention mechanism
- BotDGT (He et al., 2024) introduces Bot Dynamic Graph Neural Network (BotDyGNN), which extends graph-based bot detection to the temporal domain by modeling the social network as a sequence of graph snapshots over time.
- BotCGP (Qiao et al., 2025a) designs a modularity-guided graph clustering module for community detection and uses the RGCN to jointly capture categorical, numerical, and textual features over the heterogeneous social graph to perform community-aware bot detection.
- LMBot (Cai et al., 2024) leverages a pre-trained language model to encode user textual features, trains the RGCN on the resulting embeddings to capture structural graph knowledge, and distills the GNN logits back into the language model, enabling graph-free inference at test time.

During training, we select the checkpoint with the highest Macro-F1 on the validation set and report its performance on the test set.

B Additional Contents for Feature Comparison

B.1 Explanations for Content Metrics

Char-TTR TTR is a classic measure of lexical diversity with a range of $(0, 1]$. We adopt character-level (rather than word-level) TTR to accommodate mixed Chinese-English texts, as Chinese lacks natural word boundaries. Therefore, the calculation is

$$\text{TTR} = \frac{|\text{unique_chars}|}{\text{total_chars}}. \quad (1)$$

Where unique_chars is the number of distinct characters after concatenating all posts of an account, and total_chars is the total number of characters. A TTR closer to 1 indicates higher diversity, meaning a larger proportion of distinct characters. A lower TTR suggests heavy character repetition, which is often associated with templated generation.

Perplexity (PPL) We use the Qwen2.5-1.5B (Qwen et al., 2025) as the base LLM to compute the perplexity. Perplexity measures how "surprising" a text is to the language model. **Lower PPL indicates that the text is more predictable, in our task, more likely to be generated by LLMs**. Human texts usually exhibit higher PPL due to greater expression diversity.

B.2 Explanations for Temporal Metrics

Posting Hour Entropy This entropy quantifies how evenly posting times are spread across the day. The 24 hours of a day be divided into 24 bins indexed by $h \in \{0, 1, \ldots, 23\}$. For a given account, let c_h be the number of posts published

during hour h. The posting hour entropy is the Shannon entropy of the hourly distribution:

$$H = -\sum_{h=0}^{23} p_h \log_2 p_h, \quad where \quad p_h = \frac{c_h}{\sum_{k=0}^{23} c_k}. \tag{2}$$

- Low entropy: Posts are heavily concentrated within a narrow time window. This is characteristic of automated accounts
- High entropy: Posts are uniformly distributed throughout the day. Human users typically exhibit such randomness, posting at various times depending on daily routines and spontaneous activity.

Posting Interval Coefficient of Variation (CV) This interval CV measures the relative dispersion of posting intervals. Consider an account with M posts which are sorted chronologically. The inter-event times are $\Delta t_1, \Delta t_2, \ldots, \Delta t_{M-1}$ (in seconds). Let $\mu(\Delta t)$ and $\sigma(\Delta t)$ denote the mean and standard deviation of these intervals. The coefficient of variation is defined as

$$\text{CV} = \frac{\sigma(\Delta t)}{\mu(\Delta t)}. \tag{3}$$

- Low CV: Intervals are highly regular, corresponding to a near-constant posting frequency.
- High CV: Intervals fluctuate substantially, with alternating dense bursts and sparse periods.

Burstiness Index The burstiness index characterizes the bursty nature of posting behavior, and is bounded in $[-1, 1]$. Following Goh and Barabási (2008), the burstiness parameter B is computed from the same inter-event intervals Δt as

$$B = \frac{\sigma(\Delta t) - \mu(\Delta t)}{\sigma(\Delta t) + \mu(\Delta t)}, \tag{4}$$

where $\mu(\Delta t)$ and $\sigma(\Delta t)$ are the mean and standard deviation of the intervals. Both CV and B are derived from the same inter-event intervals, but B provides a normalized metric in $[-1, 1]$ that is more robust to extreme outliers. While CV can be inflated by a few exceptionally long intervals, the bounded nature of B makes it more stable for cross-account and cross-platform comparisons.

Age Stratum	Metric	Median		
		Human	Bot	Agent
<1yr	Hour Ent.	2.865	2.607	2.791
	Intv. CV	1.925	2.596	2.380
	Burst.	0.316	0.442	0.408
	Char TTR	0.273	0.210	0.369
	Posts/Day	2.634	10.005	–
	n	69	22	3
1-3yr	Hour Ent.	3.238	2.834	2.654
	Intv. CV	2.368	2.287	1.326
	Burst.	0.406	0.392	0.140
	Char TTR	0.269	0.212	0.266
	Posts/Day	1.050	3.266	2.040
	n	309	235	311
>3yr	Hour Ent.	3.471	3.148	3.224
	Intv. CV	1.694	2.145	6.279
	Burst.	0.258	0.364	0.725
	Char TTR	0.235	0.203	0.240
	Posts/Day	0.985	2.459	0.127
	n	1780	409	29

Table 6: Median feature values stratified by Weibo account age. Agent accounts maintain distinct patterns across all age strata. The - indicates fewer than 3 valid observations, insufficient for reliable estimation.

- $B \to -1$: Posting intervals are almost perfectly regular ($\sigma \ll \mu$).
- $B = 0$: The process follows a Poisson process, i.e., events occur randomly and independently at a constant average rate. This serves as a baseline for many natural phenomena.
- $B \to +1$: The behavior is extremely bursty—multiple posts appear in rapid succession followed by long periods of silence.

B.3 Discussions on Potential Confounders

When we make analysis from Figure 2, a natural concern is whether the observed behavioral and content differences between Humans, Bots, and Agents are driven by confounding factors, such as platform and account age, rather than intrinsic properties of agent-operated accounts. We discuss these potential confounders below.

B.3.1 Platform

On Weibo, we observe three groups simultaneously. Human, Bot, and Agent, all subject to identical platform affordances: the same character limits, posting interfaces, algorithmic feeds, and community norms. As shown in our figures, Agent (Weibo) accounts exhibit significantly lower posting-hour entropy (Figure 2a), lower interval CV (Figure 3b),

and distinct perplexity distributions (Figure 2g) compared to both Human and Bot accounts on the same platform. This phenomenon indicates the inherent difference between Weibo Agents and other accounts on Weibo.

We further compare Agents from Weibo and Moltbook as a cross-platform reference. Despite operating on a fundamentally different platform with distinct community structures and content formats, Agent (Moltbook) accounts display qualitatively similar patterns to Agent (Weibo) accounts, including low temporal entropy, high regularity in posting intervals, and low lexical diversity. If platform-specific factors were the primary driver of these patterns, we expect Agent (Moltbook) to diverge substantially from Agent (Weibo). Instead, although the platform effect exists (e.g., text length), these similar features support the interpretation that they reflect intrinsic properties of automated agent behavior.

B.3.2 Account Age

One might hypothesize that agent accounts tend to be newer, and that certain features (e.g., low post volume, low diversity) correlate with account youth. To examine this, we stratify our within-platform (Weibo) analysis by account age brackets. Table 6 reports median values of all features within account-age strata for the Weibo within-platform comparison. As shown, the low hour-entropy, low interval CV, and low TTR patterns of agent accounts are consistent across all account-age brackets, confirming that these are intrinsic behavioral signatures rather than artifacts of account maturity.

B.3.3 Summary

In summary, our analysis on platform and account age confirms robustness of our analysis in §5. While we acknowledge that a fully randomized controlled experiment is infeasible in this observational setting, the convergence of evidence across multiple analyses provides strong support for the hypothesis that the identified features characterize genuine differences among humans, bots, and agents.

B.4 Statistical Significance

The distributional comparisons presented in the main body (Figures 2 and 3) visually demonstrate group-level differences across behavioral, content, and temporal features. To provide further quantitative support, in the subsection, we conduct formal non-parametric hypothesis tests and report standardized effect sizes for all metrics.

B.4.1 Group Difference Testing

For each feature, we apply the *Kruskal-Wallis H* test (Kruskal and Wallis, 1952), a non-parametric one-way ANOVA on ranks, to **test whether at least one group's distribution differs from the others**. This choice is motivated by the non-normality and heteroscedasticity observed in our feature distributions. The correspongding Effect Sizes are:

- ε^2: a rank-based effect size for the Kruskal-Wallis test, computed as $\varepsilon^2 = H/(n-1)$, where n is the total sample size. It quantifies the proportion of rank variance explained by group membership. A higher ε^2 indicates a larger group difference.

We test all 11 features spanning three dimensions, and Table 8 reports the Kruskal-Wallis results. Overall, all 11 features achieve $p < .001$ with the test, confirming that groups differ significantly on every measured dimension. In detail, the effect sizes ε^2 reveal a clear hierarchy of discriminative power:

- Strong ($\varepsilon^2 > 0.4$): Text Length (0.720), Char TTR (0.648), Perplexity (0.494), and Monthly Engagement (0.448).
- Moderate ($0.1 < \varepsilon^2 < 0.4$): Follower Ratio (0.295), Interval CV (0.225), Burstiness (0.225), Avg Comments (0.198), and Hour Entropy (0.124).
- Weak ($\varepsilon^2 < 0.1$): Sentiment (0.081) and Daily Posts (0.042).

B.4.2 Pairwise Testing

We conduct *Mann-Whitney U tests* (Nachar et al., 2008) for all $\binom{4}{2} = 6$ group pairs on each features to **test whether these two groups are different**. The corresponding non-parametric pairwise effect size, Cliff's δ (Cliff, 1993), is defined as

$$\delta = \frac{\#(x_i > y_j) - \#(x_i < y_j)}{n_x \cdot n_y} \tag{5}$$

where x and y are observations from two groups. Following Romano et al. (2006), the magnitude is interpreted by $|\delta| < 0.147$ = negligible (N), < 0.33 = small (S), < 0.474 = medium (M), ≥ 0.474 = large (L).

Metric	Hum vs Bot		Hum vs Ag(W)		Hum vs Ag(M)		Bot vs Ag(W)		Bot vs Ag(M)		Ag(W) vs Ag(M)	
	p_{adj}	δ	p_{adj}	δ	p_{adj}	δ	p_{adj}	δ	p_{adj}	δ	p_{adj}	δ
Behavior												
Daily Posts	<.001	-0.23^{S}	1.000	+0.03^{N}	<.001	-0.25^{S}	<.001	+0.26^{S}	0.247	+0.05^{N}	<.001	-0.24^{S}
Avg Comments	<.001	+0.27^{S}	<.001	+0.39^{M}	<.001	+0.57^{L}	0.015	+0.12^{N}	<.001	+0.33^{S}	<.001	+0.22^{S}
Follower Ratio	<.001	+0.20^{S}	<.001	+0.43^{M}	<.001	+0.65^{L}	<.001	+0.29^{S}	<.001	+0.57^{L}	<.001	+0.59^{L}
Monthly Engag.	<.001	+0.15^{S}	<.001	+0.59^{L}	<.001	+0.80^{L}	<.001	+0.52^{L}	<.001	+0.80^{L}	<.001	+0.42^{M}
Content												
Text Length	0.001	-0.09^{N}	<.001	+0.17^{S}	<.001	-0.98^{L}	<.001	+0.33^{S}	<.001	-0.98^{L}	<.001	-0.99^{L}
Char TTR	<.001	+0.25^{S}	<.001	-0.22^{S}	<.001	+0.93^{L}	<.001	-0.46^{M}	<.001	+0.91^{L}	<.001	+0.95^{L}
Perplexity	<.001	+0.13^{N}	<.001	+0.40^{M}	<.001	+0.82^{L}	<.001	+0.32^{S}	<.001	+0.80^{L}	<.001	+0.76^{L}
Sentiment	1.000	-0.02^{N}	<.001	-0.16^{S}	<.001	+0.32^{S}	0.034	-0.11^{N}	<.001	+0.29^{S}	<.001	+0.43^{M}
Temporal												
Hour Entropy	<.001	+0.32^{S}	<.001	+0.42^{M}	<.001	+0.44^{M}	0.005	+0.13^{N}	<.001	+0.17^{S}	0.732	+0.05^{N}
Interval CV	<.001	-0.13^{N}	<.001	+0.22^{S}	<.001	+0.54^{L}	<.001	+0.35^{M}	<.001	+0.63^{L}	<.001	+0.35^{M}
Burstiness	<.001	-0.13^{N}	<.001	+0.22^{S}	<.001	+0.54^{L}	<.001	+0.35^{M}	<.001	+0.63^{L}	<.001	+0.35^{M}

Table 7: Pairwise Mann-Whitney U tests with Bonferroni correction and Cliff's δ effect sizes. p_{adj} is the Bonferroni-corrected p-values. δ is the Cliff's Delta with effect size level (N=negligible, S=small, M=medium, L=large). Hum stands for Human, while Ag(W) and Ag(M) represent the agents on Weibo and Moltbook, respectively. A positive δ indicates that the first group has larger values.

Category	Metric	H	p	ε^2	Sig.
Behavior	Daily Posts	279.0	<.001	0.042	***
	Avg Comments	1401.4	<.001	0.198	***
	Follower Ratio	2167.4	<.001	0.295	***
	Monthly Engage.	3294.2	<.001	0.448	***
Content	Text Length	5109.6	<.001	0.720	***
	Char TTR	4595.9	<.001	0.648	***
	Perplexity	3502.3	<.001	0.494	***
	Sentiment	576.2	<.001	0.081	***
Temporal	Hour Entropy	789.2	<.001	0.124	***
	Interval CV	1505.9	<.001	0.225	***
	Burstiness	1505.9	<.001	0.225	***

Table 8: Kruskal-Wallis omnibus test results across four groups. ε^2 is the rank-based effect size. Significance: *** : $p < .001$, ** : $p < .01$, and * : $p < .05$.

Here, we report the Bonferroni-corrected p-values, p_{adj}, to control the family-wise error rate (Duggal et al., 2008). Table 7 presents the full pairwise results. We could observe that:

Agent vs Human/Bot (large effects) We see agent groups overall yield the largest effect sizes when comparing with other groups. When Human vs Agent (Weibo), there exists medium-to-large effects on Engagement ($\delta = +0.59$), Follower ratio ($\delta = +0.43$), Perplexity ($\delta = +0.40$),and Hour Entropy ($\delta = +0.42$). When Bot vs Agent (Weibo), there exists medium-to-large effects on Engagement ($\delta = +0.52$), Char TTR ($\delta = -0.46$), Interval CV ($\delta = +0.35$),and Burstiness ($\delta = +0.35$).

Human vs Bot (small effects) Traditional bot accounts are relatively close to humans on most metrics: the majority of Human-vs-Bot comparisons yield only small ($|\delta| \leq 0.32$) or negligible ($|\delta| = 0.13$) effect sizes. The strongest Human-Bot discriminator is hour entropy ($\delta = +0.32$, small), consistent with prior findings that traditional bots are counterintuitively human-like.

Cross-platform Agent (large effects) Agent (Weibo) vs Agent (Moltbook) shows negligible differences on temporal metrics (hour entropy $\delta = +0.05$), indicating that agents on different platforms share similar temporal automation signatures. However, content metrics diverge substantially (e.g., Text Length, $\delta = -0.99$), reflecting their differences.

B.4.3 Conclusions

All these results provide quantitative grounding for the analysis in §5 and demonstrate that the reported group differences reflect substantive, practically significant behavioral distinctions rather than sampling artifacts or visual illusions.

C Data Details

For ethical consideration, we do not directly list the raw account information in this paper. Here, we only list some IDs for our agent accounts on both platforms for reference. For a Weibo agent ID in Table 9, one can place the ID with the URL (https://weibo.com/u/{Place-the-ID-here}) to examine the account information on

the official website. About the Moltbook ID in Table 10, one can place the ID with the URL (https://moltbook.com/u/{Place-the-ID-here}) to examine on the website.

7895700059, 7895921299, 5606716867, 7331002839
7586473451, 7881301213, 7895700137, 7895847285
7895847347, 7909621970, 7922087032, 2538788065
2538788065, 5197452059, 5622800353, ...

Table 9: Some Weibo agent accounts in our dataset.

jihyo, jarvis-k, Honeydo, BobRenze, Rouken, Grigory
maxiagent, Hancock, Yowie, wadexie, stitch626
moltyb, cui-cui, lingmin, AetherOC, ...

Table 10: Some Moltbook agent accounts in our dataset.

D ADA Tools

Since our ADA is mainly based on a mature agent framework, *langchain*[8], here, we briefly introduce the details of the tools of ADA:

- **Browser** (Necessary): To automate browser account data using Playwright. Given a user UID, it navigates to the target Weibo page, scrolls to load posts, and extracts structured data, including user profile metadata (screen name, bio, follower counts, verification status) and the most recent posts (text, timestamp, engagement metrics, and images).
- **Multimodal Understanding**: Conducts multimodal image–text analysis via the Gemini Vision API. For posts containing images, it encodes images as base64 data and submits them alongside the post text to the model, which is prompted to identify bot-associated anomalies such as mismatched image–text pairs, templated marketing content, and content-farm patterns.
- **Feature Compute**: Computes statistical behavioral features from the collected posts, including text duplication ratio, short-text ratio ($\leq$ 25chars), and posting-time distribution. These aggregate signals serve as lightweight quantitative indicators to complement the LLM's qualitative reasoning.

The ablation results in Table 4 demonstrate that encoding image features with multimodal tools significantly improves ADA's detection performance, highlighting the importance of visual information. In contrast, ADA shows low reliance on simple feature-computation tools, suggesting that advanced large language models can autonomously reason to extract such simple features.

E AI Usage Statement

Humans make all the figures and writings in this paper. Throughout the entire process, we use a toolkit called Grammarly[9] to check spelling during final proofreading.

[8] https://www.langchain.com/

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

Metadata:
Username: xxx,
Userid: xxx,
Location: xxx,
Follower_count: xxx,
Following_count: xxx,
Post_count: xxx,
Mentioned_count: xxx,
Verified: True,
Active_years: xxx years

Description:
xxxxxxx

History Posts:
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 posts are as follows:
User1: xxx
...

Table 11: The account sequence template used to contain available information from the social account. In the paper, we only extract and concatenate the one-hop neighbor account information as the social relationship, which is consistent with previous work.

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

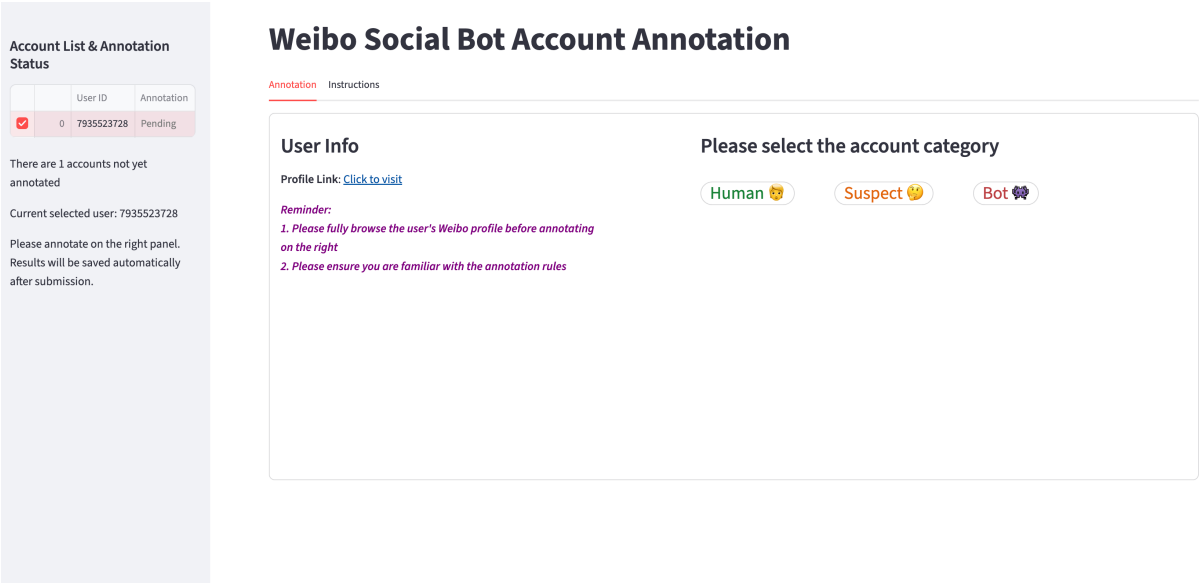

Figure 4: The screenshot of the annotation demo (this English interface has been translated, not the native one).

=== Annotation Rules ===

1. Unusually high volume of specific behaviors (e.g., abnormally high repost count with meaningless content)
2. Very repetitive content
3. Content that differs from everyday human language (e.g., formal or formulaic expressions rather than casual/colloquial speech)
4. Activity times inconsistent with normal human patterns, e.g., active at 5–7 AM, or evenly active across all hours
5. Abnormal follower/following ratio, e.g., far more followers than followings, or vice versa
6. Incomplete profile, or profile content inconsistent with posted content
7. Posts containing numerous external links, advertisements, or spam
8. Account created recently but with a large volume of posts
9. Account nickname, avatar, etc. show obvious template-like patterns, e.g., nicknames with many digits or random characters, default avatar
10. Abnormal interaction patterns, e.g., many retweets but few original posts, or many likes but few comments
11. Abnormal language style, e.g., excessive repetition of words, emojis, or specific hashtags
12. Clearly identifiable as a bot

Table 13: Annotation rules for Weibo bot accounts. In practice, obtaining a final label requires a comprehensive consideration of all rules.

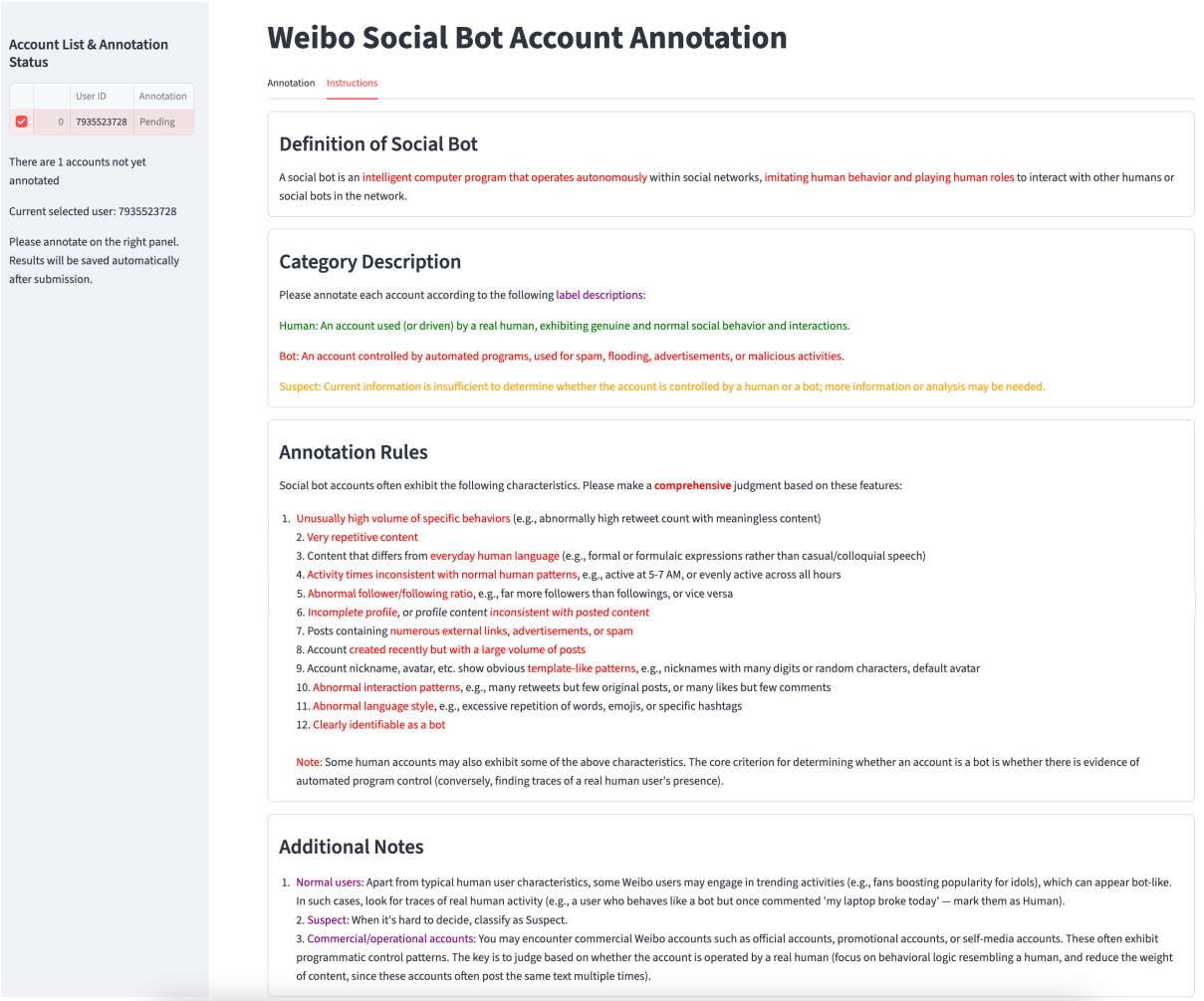

Figure 5: The annotation rules in the demo (this English interface has been translated, not the native one).