SoMe: A Realistic Benchmark for LLM-based Social Media Agents

Dizhan Xue[1,2], **Jing Cui**[3], **Shengsheng Qian**[1,2*], **Chuanrui Hu**[4], **Changsheng Xu**[1,2,5]
[1]State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]School of Computer Science and Engineering, Tianjin University of Technology
[4]Nanjing University of Posts and Telecommunications
[5]Peng Cheng Laboratory
xuedizhan17@mails.ucas.ac.cn, cjtxdy@stud.tjut.edu.cn, shengsheng.qian@nlpr.ia.ac.cn, csxu@nlpr.ia.ac.cn

Abstract

Intelligent agents powered by large language models (LLMs) have recently demonstrated impressive capabilities and gained increasing popularity on social media platforms. While LLM agents are reshaping the ecology of social media, there exists a current gap in conducting a comprehensive evaluation of their ability to comprehend media content, understand user behaviors, and make intricate decisions. To address this challenge, we introduce SoMe, a pioneering benchmark designed to evaluate social media agents equipped with various agent tools for accessing and analyzing social media data. SoMe comprises a diverse collection of 8 social media agent tasks, 9,164,284 posts, 6,591 user profiles, and 25,686 reports from various social media platforms and external websites, with 17,869 meticulously annotated task queries. Compared with the existing datasets and benchmarks for social media tasks, SoMe is the first to provide a versatile and realistic platform for LLM-based social media agents to handle diverse social media tasks. By extensive quantitative and qualitative analysis, we provide the first overview insight into the performance of mainstream agentic LLMs in realistic social media environments and identify several limitations. Our evaluation reveals that both the current closed-source and open-source LLMs cannot handle social media agent tasks satisfactorily. SoMe provides a challenging yet meaningful testbed for future social media agents.

Code and Datasets — https://github.com/LivXue/SoMe
Extended version — https://arxiv.org/pdf/2512.14720

Introduction

Language agents, which integrate tools with large language models (LLMs), have attracted broad research interest and demonstrate promising applications in human-level intelligent tasks, including coding (Zhang et al. 2024; Islam, Ali, and Parvez 2024), web browsing (He et al. 2024; Li et al. 2025), and healthcare (Kim et al. 2024; Li et al. 2024). Meanwhile, the development of social media agents, aimed at comprehending vast social data and executing human-like behaviors, has been a longstanding goal in the field of artificial intelligence (Edwards et al. 2014; Carr and Hayes 2015). Recent progress in LLM-based agents has also led to unprecedented prosperity of social media agents (Mou, Wei, and Huang 2024; Zhang et al. 2025a), which can potentially take on various tasks such as social event analysis (Qian, Zhang, and Xu 2016), post and user recommendation (Tang, Tang, and Liu 2014), and social behavior simulation (Yang et al. 2024). These agents can alleviate the burden on users and organizations managing an active online presence and analyzing vast amounts of social media data. Moreover, these human-like agents are gradually active in the social networks of more and more human users.

While the increasing participation of social media agents is reshaping the ecology of social media, the incomplete understanding of their capability in multiplexed social media tasks has raised concerns. Previous evaluations of agents and LLMs in the area of social media typically focus on a single task, such as misinformation detection (Nakazato et al. 2024) or user behavior prediction (Jiang and Ferrara 2023). For example, BotSim (Qiao et al. 2025) designs a user behavior simulation framework for LLM-based agents, powered by tens of thousands of data collected from Reddit. TrendSim (Zhang et al. 2025b) creates a user simulation environment for social media agents that incorporates a time-aware interaction mechanism, based on data of 1,000 users collected from Weibo. These existing evaluations are also limited by insufficient data and the lack of ground truth (e.g., TrendSim utilizes LLMs to evaluate the rationality of agent behaviors without ground truth references). To sum up, these evaluations are insufficient for providing a full range of knowledge about social media agents. To further catalyze the research, a versatile benchmark with abundant real data and annotations is required for developing and deploying trustworthy social media agents.

To address the abovementioned challenges, we propose SoMe, the first versatile benchmark for social media agents, which aims at comprehensively evaluating the agentic capabilities of LLMs on social media tasks. SoMe assesses LLMs across 8 critical tasks: real-time event detection, streaming event summarization, misinformation detection, user behaviour prediction, user emotion analysis, user comment simulation, media content recommendation, and social media question-answering. To enable LLM-based agents to perform these tasks, we build a platform with 8 agent tools

*Corresponding Author.

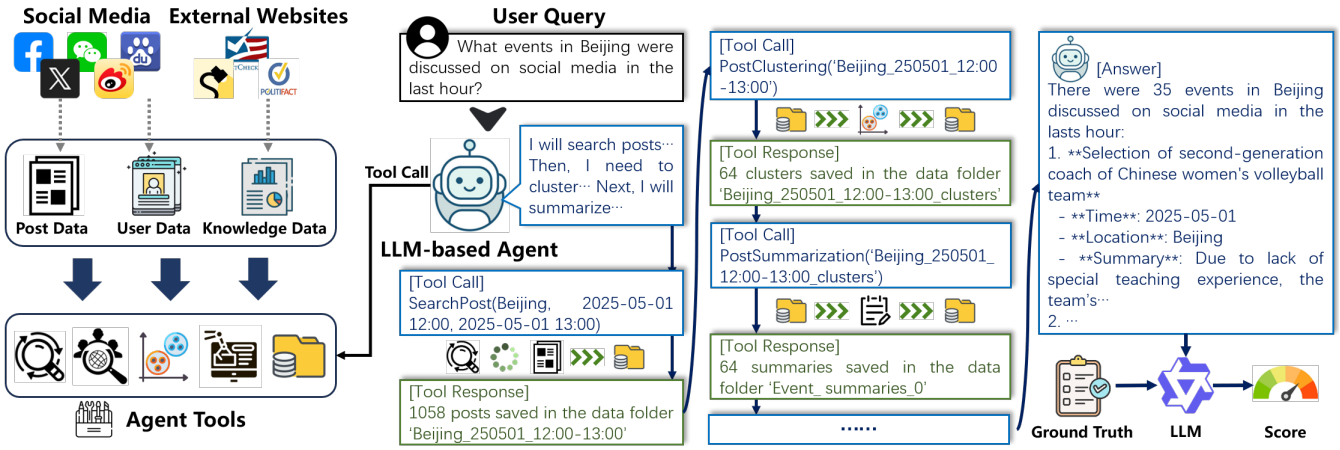


Figure 1: **Workflow diagram of agents in SoMe.** The social media agent interacts with tools for data acquisition, management, and analysis, in order to arrive at an answer for the user query. The answer is evaluated with the assistance of an LLM scorer.

for acquiring, managing, and analyzing data in and outside of social media. Moreover, we collect 9 million of real and public data from 32 social media platforms (e.g., X[1] and Weibo[2]) and external websites, aligning with the practical application of social media agents. These data cover a wide range of topics, events, and users, while remaining real and challenging attributions of social media data, which are noisy, temporal, and diverse. We leverage human-LLM interactive pipelines to annotate task queries and their ground truth for all tasks, elaborately verified by 10 professional annotators in total. In our benchmark, LLM-based agents are required to conduct step-by-step data processing and long-context reasoning with appropriate tool calls in a realistic environment, as exemplified in Figure 1. Brief statistics of data are demonstrated in Table 1.

Our contributions are summarized as follows:

- We propose SoMe, the first realistic evaluation benchmark for LLM-based social media agents. SoMe comprises 8 tasks and millions of real data, assessing capabilities across social data analysis, user personality understanding, and long-context knowledge reasoning.
- We establish a running platform for LLM-based social media agents with 8 tools for data acquisition, management, and analysis. The agents can perform complex tasks in social media by step-by-step data processing and reasoning with appropriate tool combinations.
- We provide insight into the performance of 13 mainstream agentic LLMs in realistic social media environments. Our findings reflect the bottleneck of existing social media agents, providing suggestions for future work.

Related Work

LLM-based Social Media Agents

Social media has emerged as one of the most popular technological innovations globally, attracting billions of users and fundamentally reshaping how individuals, organizations, and societies communicate, interact, and exchange information. Due to its tremendous social impact, the research on social media has become an active and long-standing area in artificial intelligence (Roy et al. 2012; Qian, Zhang, and Xu 2016; Qian et al. 2023; Xue et al. 2025). However, the open-world nature and sophisticated structure of social media significantly reduce the effectiveness of models with static inputs. Recently, in the pursuit of developing intelligent agents, there has been considerable focus on integrating LLMs with external tools (Wang et al. 2024b; Xue, Qian, and Xu 2024; He, Demartini, and Gadiraju 2025; Qian et al. 2024; Yang et al. 2025b). Specially, agentic LLMs (Hurst et al. 2024; Comanici et al. 2025; Yang et al. 2025a) are proposed to enhance the abilities of using external tools and handling agentic tasks. Based on agentic LLMs, these agents enable powerful capabilities in environment interaction, decision-making, and task execution. Especially, LLM-based social media agents exhibit unprecedented ability in multi-round information acquisition, analysis, and reasoning, boosting performance in tasks such as misinformation detection (Wan et al. 2024), user simulation (Gao et al. 2024), and event analysis (Wang et al. 2024a). For instance, Wang et al. (2024a) design an LLM-based agent to iteratively filter out irrelevant news and employ human-like reasoning to analyze complex social events. Zhang et al. (2025b) propose an LLM-based multi-agent system to simulate user interactions in trending topics on social media. Wei et al. (2024) devise anonymous opinion leader agents to simulate and predict the user emotional responses to different events on social media.

However, these social media agents specialize in different tasks for social media and cannot handle multiple tasks simultaneously. To thoroughly evaluate the capabilities of social media agents, we propose SoMe to measure the performance across 8 agentic social media tasks with a versatile agent platform. This platform is equipped with various tools enabling interactions within a realistic social media environment.

[1]https://x.com
[2]https://weibo.com/

Task	#Query	#Data	Data Type
Real-time Event Detection	568	476,611	Posts
Streaming Event Summarization	154	7,898,959	Posts
Misinformation Detection	1,451	27,137	Posts & Knowledge
User Behavior Prediction	3,000	840,200	Posts & Users
User Emotion Analysis	2,696	840,200	Posts & Users
User Comment Simulation	4,000	840,200	Posts & Users
Media Content Recommendation	4,000	840,200	Posts & Users
Social Media Question-answering	2,000	8,651,759	Posts & Users
Total	17,869	9,242,907	All

Table 1: Statistics of data in SoMe: **#Query** denotes the number of queries (with annotations) in the dataset, #Data denotes the number of accessible data in the database, and **Data Type** denotes the type of data in the database.

Benchmarks for LLM-based Agents

With the rise of LLM-based agents, the performance and trustworthiness of agents in various real-world tasks are of concern (Liu et al. 2024b; He, Demartini, and Gadiraju 2025). Numerous studies have been conducted to evaluate the tool-use and environment-interaction abilities of LLM-based agents (Xie et al. 2024; Skrynnik et al. 2025). For example, CharacterEval (Tu et al. 2024) evaluates the role-playing ability of agents in multi-turn role-playing dialogues. OSWorld (Xie et al. 2024) evaluates the real-world computer-using ability of multimodal agents in a real computer environment. VisualWebArena (Koh et al. 2024) assesses the instruction execution capability of multimodal web agents on realistic visually grounded tasks. While existing benchmarks have evaluated LLM-based agents in various application domains, a comprehensive benchmark for social media agents is still lacking, despite their prevalence in real-world applications.

Unlike environments with explicit guidance in the above benchmarks, social media is an extremely noisy environment, where information about a concept can usually be scarce or inundated with massive irrelevant information (Baldwin et al. 2013). Therefore, we establish SoMe, a novel and comprehensive benchmark that aims to address this gap by offering an elaborate and realistic evaluation framework for LLM-based social media agents.

SoMe Benchmark

Social Media Tasks

Evaluating the capabilities of social media agents necessitates a comprehensive and structured approach. We collect the 8 most-considered agentic tasks related to social media, categorized into 3 classes, as follows:

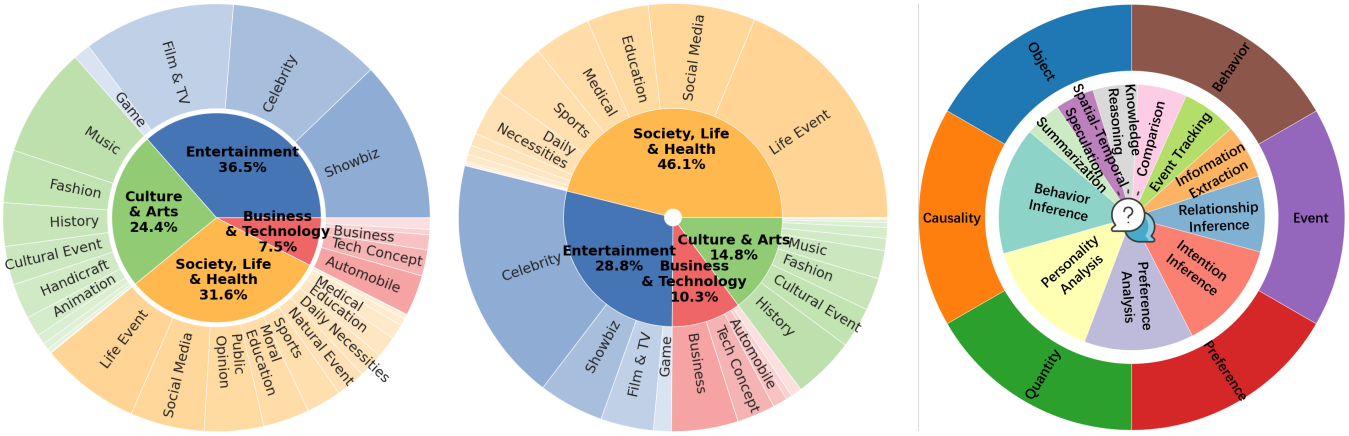
- **Post-centered Tasks:** This class of tasks requires multi-round data processing and interactions with social media posts, sometimes necessitating access to external knowledge bases. **Real-time Event Detection (RED)** aims to detect social events in real-time from a large volume of recent posts. **Streaming Event Summarization (SES)** aims to progressively summarize the details of a social event from continuously published posts. **MisInformation Detection (MID)** aims to identify false, misleading, or inaccurate information within posts, utilizing support from external knowledge.
- **User-centered Tasks:** This class of tasks involves understanding user preferences and behavior patterns through multi-round tool calls, thereby making personalized predictions. **User Behavior Prediction (UBP)** aims to predict the user interaction behaviors with specific posts. **User Emotion Analysis (UEA)** aims to predict the emotions that emerge from users towards particular posts. **User Comment Simulation (UCS)** aims to predict the comments that users will make on given posts.
- **Comprehensive Tasks**: This class of tasks needs comprehensive analyses of both extensive posts and users to deduce the answers. **Media Content Recommendation (MCR)** aims to recommend social media content that aligns with user preferences. **Social Media Question-answering (SMQ)** aims to answer questions regarding the public information available in posts and users.

Data Sources and Annotations

Data sources. We collect data from primarily 32 social media platforms, such as X and Weibo, with a predominant focus on English and Chinese content. Specifically, these data are from three sources: 1) We buy commercial data from the Xiaoying company[3], which are publicly visible data and crawled from diverse social media platforms. 2) We crawl data of top influencers and users in their social networks on Weibo. 3) We adopt open-source datasets (Yang et al. 2022), where the fact-checked reports from external websites are separated to form a knowledge base for misinformation detection. To sum up, there are 9,164,284 posts, 6,591 user profiles, and 25,686 reports in our database, which can be accessed through various agent tools.

Data annotations. While text is the major modality in social media, we convert images and videos in posts into captions (by Qwen2.5-VL-72B (Bai et al. 2025)), OCR texts (by Qwen2.5-VL-72B), and ASR texts (by GPT-4o (Hurst et al. 2024)). After that, we propose a semi-automated annotation pipeline for all tasks. Specifically, for UBP, UCS, and MCR, we automatically generate task queries about the collected users and convert the collected data into ground truth annotations based on templates. Since the results are already

[3]https://www.xiaoying.tv/



(a) User interest distribution (b) Post topic distribution (c) Task query distribution

Figure 2: Distributions of social media data in SoMe.

Tool Name(Parameters)	Description
DataFolder(folder_name, start_idx, end_idx)	Output data in the data folder of a given name, from a start index to an end index.
SearchPost(location, start_time, end_time)	Search posts about a location published in a time period, save them in a data folder.
SearchTopic(topic_name)	Search posts about a given topic and save them in a data folder.
SearchUser(uid)	Search a specific user, output the profile, and save the posts into a data folder.
RetrievePost(query, folder_name, topk)	Retrieve the top-k most relevant posts with the query in a data folder.
RetrieveKnowledge(query, topk)	Retrieve the top-k most relevant reports with the query in the knowledge base.
PostClustering(folder_name)	Cluster posts in a data folder and save clusters in another data folder.
PostSummarization(folder_name)	Summarize post clusters in a data folder and save summaries in another data folder.

Table 2: Agent tools in SoMe.

included in the collected data, no LLM- or human-involved edition of ground truth is involved for these tasks. For RED, SES, UEA, and SMQ, we leverage multi-round human-LLM interactive pipelines to annotate ground truth results for automatically generated task queries, based on Qwen3-32B. All ground truth results undergo manual filtering and verification by 10 professional annotators to guarantee the quality of annotations. For MID, we merge the open-source LIAR-RAW and RAWFC datasets (Yang et al. 2022), where the ground truth is already annotated.

Data Statistics

The brief statistics of the dataset are shown in Table 1. We conduct further investigation on the data diversity of SoMe.

User analysis. The user interests identified by Qwen3-32B based on the collected user profiles are presented in Figure 2(a), showcasing the outline and a broad spectrum of user interests. Specifically, there are 4 major interests, including Entertainment (36.5%), Society, Life, & Health (31.6%), Culture & Arts (24.4%), and Business & Technology (7.5%), and 33 subcategories. Among subcategories, Showbiz (12.2%), Celebrity (11.6%), Film & TV (11.5%), Music (8.4%), and Life Event (7.6%) are the most popular. Overall, the interests of users in our data are highly diverse, as the most popular one accounts for 12.2%, which improves the comprehensiveness and robustness of our benchmark.

Post analysis. The topic distribution of the collected posts, as identified by Qwen3-32B, is presented in Figure 2(b). The results reveal a diverse spread across four primary domains aligned with user interests: Society, Life & Health (46.1%), Entertainment (28.8%), Culture & Arts (14.8%), and Business & Technology (10.3%). Within these categories, Life Events (18.8%) and Celebrities (18.5%) emerged as the most prevalent subtopics. Significantly, all other identified subtopics individually accounted for less than 10% of the posts. This substantial heterogeneity within the collected social media posts underscores the breadth of content captured. Consequently, this diverse representation bolsters both the comprehensiveness and the robustness of our benchmark.

Query analysis. We further investigate the queries in all 8 tasks, as shown in Figure 2(c). Task queries in SoMe can be roughly classified into questions about behaviors, events, preferences, quantity, causality, and objects. Moreover, we classify all queries into 11 categories, of which the percentages are presented in the figure. Each category corresponds to the assessment for one specific skill of social media agents. The criteria for these categories are as follows: (1) *Behavior Inference*: Analyzing the user behavior patterns and predicting user behaviors. (2) *Personality Analysis*: Identifying the personality of users based on their social media activities. (3) *Preference Analysis*: Inferring and interpreting the preferences of users. (4) *Intent Inference*: Discerning the underlying intention behind actions and speech. (5) *Relationship Inference*: Discovering the relationships between people, objects, or events. (6) *Information*

Model	Size	RED	SES	MID	UBP	UEA	UCS	MCR	SMQ	Avg.
API-based										
GPT-4o	N/A	47.59	36.17	50.24	55.17	31.53	52.48	61.63	64.21	49.88
Gemini-2.5-Flash	N/A	**54.92**	**44.87**	45.62	57.50	41.94	56.00	62.75	71.01	**54.33**
Kimi-K2-Instruct	1T	50.40	38.57	47.83	51.50	**45.94**	**58.25**	57.00	77.38	53.36
DeepSeek-V3	671B	35.94	38.83	**51.00**	55.06	42.98	54.92	56.25	75.94	51.37
Open-source										
Llama-3.3-70B-Instruct	70B	34.84	33.77	50.24	55.83	32.77	53.00	61.40	64.83	48.34
Qwen3-32B	32B	44.25	41.04	47.42	**67.03**	33.53	54.98	**63.28**	**80.27**	53.98
GLM-4-32B-0414	32B	27.55	27.27	41.38	40.17	26.10	47.13	44.08	57.19	38.86
Devstral-Small-2507	24B	30.99	33.31	47.07	47.93	22.43	46.68	47.75	58.25	41.80
Qwen3-14B	14B	43.97	40.19	44.80	66.20	33.35	54.43	62.13	77.47	52.82
GLM-4-9B-0414	9B	14.42	29.68	33.63	48.73	28.63	44.10	49.05	51.82	37.51
Qwen3-8B	8B	40.38	36.69	45.21	61.73	33.03	53.33	60.55	76.18	50.89
DeepSeek-R1-0528-Qwen3-8B	8B	17.71	28.83	26.46	43.53	21.18	31.10	34.33	51.84	31.87
Llama-3.1-8B-Instruct	8B	3.37	20.78	40.45	34.23	37.11	33.98	47.40	31.05	31.65

Table 3: Performance of agentic LLMs evaluated on 8 tasks in SoMe. Avg. denotes the average scores over 8 tasks. The highest are highlighted in bold while the lowest are marked with underlines.

Extraction: Extracting key information from noisy data for specific queries. (7) *Event Tracking*: Tracking the processes and transitions of events within long contexts. (8) *Comparison*: Comparing the information about multiple objects or from multiple sources. (9) *Knowledge Reasoning*: Retrieving and applying external knowledge to conduct logical reasoning. (10) *Spatial-Temporal Speculation*: Understanding the spatial and temporal relationships within social media data. (11) *Summarization*: Integrating information from various sources to provide thorough and succinct reports.

Agent Tools

We construct a versatile framework for social media agents with 8 tools for data acquisition, management, and analysis. All tools are designed in accordance with Model Context Protocol (MCP) (Hou et al. 2025) to ensure broad compatibility with mainstream agentic LLMs. These tools are summarized in Table 2. Detailed descriptions of all tools and their parameters are provided to agents following MCP, ensuring a full understanding of the tools. Finally, upon receiving the task query, the constructed social media agents can interact with appropriate tools and conduct step-by-step reasoning to handle the task.

Evaluation

Experiment Settings

We evaluate 13 mainstream agentic LLMs on SoMe. For API-based models, we select GPT-4o (Hurst et al. 2024), Gemini-2.5-Flash (Comanici et al. 2025), Kimi-K2-Instruct (Team 2025), and DeepSeek-V3 (Liu et al. 2024a). For open-source models, we adopt Qwen3 series (Yang et al. 2025a), Llama-3 series (Dubey et al. 2024), GLM-4 series (GLM et al. 2024), Devstral-Small-2507 (Jiang et al. 2023), and DeepSeek-R1-0528-Qwen3-8B (Guo et al. 2025). Though Kimi-K2-Instruct and DeepSeek-V3 are also open-source, we opt for API calling due to the overly large size of these models. Experiments are conducted using 8 NVIDIA A800-SMX4-80GB GPUs with the vLLM framework (Kwon et al. 2023) for open-source model deployment. All LLMs are equipped with agent tools using the standard Model Context Protocol (MCP), while the native tool-calling formats of all LLMs are supported and correctly parsed in our agent framework.

Evaluation Metrics

We separately design appropriate metrics for 8 social media agent tasks. For RED, SES, and SMQ, we employ LLM-based metrics to compare the semantics of the generated results with the ground truth and assign corresponding scores. For MID, UBP, UEA, UCS, and MCR, we utilize LLMs to filter out redundant information and extract the key answers in the agent responses. Then, we compute the accuracy (ACC) of the answers based on the ground truth. Additionally, we compute the Task Completion Rate (TCR) for all tasks, which measures the percentage of task queries an agent successfully completes in a specific task. All metric scores are normalized to the range of $[0, 100]$, and we report the average scores over test samples for all tasks.

Overall Results

The experimental results of different agentic LLMs in social media agent tasks are demonstrated in Tables 3-4.

Analysis of performance. Based on Table 3, we have the following observations:

- Both the current closed-source and open-source LLMs cannot handle social media agent tasks satisfactorily. In most cases, current agentic LLMs typically fail to attain an evaluation score exceeding 70 in most tasks. This is particularly evident in tasks such as RED, SES, and MID, characterized by a high degree of openness in the information involved, where the majority of LLMs receive scores below 50. These results reveal that building social media agents is nontrivial, and more work is required to improve the performance and trustworthiness of social

Model	Size	RED	SES	MID	UBP	UEA	UCS	MCR	SMQ	Avg.
API-based										
GPT-4o	N/A	94.54	90.91	95.11	98.00	98.70	96.45	97.35	87.50	94.82
Gemini-2.5-Flash	N/A	99.23	94.81	88.63	95.50	99.04	99.75	99.25	96.63	96.61
Kimi-K2-Instruct	1T	98.86	85.06	94.21	97.17	98.87	99.75	99.25	98.03	96.40
DeepSeek-V3	671B	99.36	92.86	98.62	99.78	99.95	100.00	100.00	98.25	98.60
Open-source										
Llama-3.3-70B-Instruct	70B	88.91	90.26	98.07	97.15	98.86	97.53	97.85	93.72	95.29
Qwen3-32B	32B	99.47	94.16	97.73	98.11	100.00	99.88	98.60	99.68	98.45
GLM-4-32B-0414	32B	88.67	68.18	87.72	67.67	85.52	86.58	79.40	83.73	80.93
Devstral-Small-2507	24B	82.92	86.36	93.45	79.73	70.10	86.78	78.70	82.53	82.57
Qwen3-14B	14B	96.23	94.81	99.45	99.39	99.00	99.80	99.80	98.62	98.39
GLM-4-9B-0414	9B	76.06	79.87	78.08	88.90	93.35	83.63	91.50	79.15	83.82
Qwen3-8B	8B	91.37	92.86	99.38	98.07	98.68	99.40	97.33	98.73	96.98
DeepSeek-R1-0528-Qwen3-8B	8B	67.43	70.78	64.58	82.93	67.88	62.86	64.88	76.51	69.73
Llama-3.1-8B-Instruct	8B	29.93	56.49	90.63	67.59	85.98	75.86	86.90	44.39	67.20

Table 4: Task Completion Rate (TCR) of agentic LLMs on 8 tasks in SoMe. Avg. denotes the average TCR over 8 tasks. The values lower than 80 are highlighted in red and the values higher than 99 are highlighted in green.

media agents. Our proposed SoMe provides a challenging yet meaningful testbed for future social media agents.

- Among API-based agentic LLMs, Gemini-2.5-Flash achieves the highest average performance score of 54.33. Furthermore, it demonstrates superior performance on individual tasks, including RED and SES with the scores of 54.92 and 44.87, respectively. Within the open-source agentic LLM category, Qwen3-32B attains the highest average score of 53.98. Moreover, it leads on key tasks such as UBP, MCR, and SMQ.
- Comparing the serial LLMs with different sizes, we could find that larger models typically outperform smaller models in social media agent tasks. For example, considering the average evaluation scores, Qwen3-32B outperforms Qwen3-14B by 2.2%, Qwen3-14B outperforms Qwen3-8B by 3.8%, and Llama-3.3-70B-Instruct outperforms Llama-3.1-8B-Instruct by 52.7%. However, larger models typically have lower inference efficiency and face challenges regarding compatibility with local development. Especially for social media agent tasks that involve a large number of posts, user profiles, and even external knowledge, developing efficient and cost-effective agents is particularly crucial.
- Interestingly, while DeepSeek-R1-0528-Qwen3-8B significantly outperforms Qwen3-8B in reasoning tasks (Guo et al. 2025), its capability declines when utilized as a social media agent. Compared to Qwen3-8B, DeepSeek-R1-0528-Qwen3-8B consistently experiences a decline in performance across 8 tasks by 56.14%, 21.42%, 41.07%, 29.48%, 35.88%, 41.68%, 43.30%, and 31.95%, respectively. These results indicate that agentic ability in social media tasks does not necessarily result from enhancing the reasoning skills of LLMs. Therefore, improving both the reasoning and agentic capabilities is crucial in the development of more powerful LLMs.

Analysis of Task Completion Rate (TCR). Based on Table 4, we have the following observations:

- Since tasks in SoMe involve calling tools in correct orders, completing the task query is nontrivial. Some open-source LLMs, such as DeepSeek-R1-0528-Qwen3-8B and Llama-3.1-8B-Instruct, achieve relatively low TCR in handling social media agent tasks. Especially, Llama-3.1-8B-Instruct can only complete task queries with an average TCR of 65.76%. These results indicate that improving the multi-round reasoning and tool-calling capability is still a challenge for small open-source LLMs.
- Among all evaluated LLMs, Gemini-2.5-Flash, DeepSeek-V3, Qwen3-32B, and Qwen3-14B can handle tasks with extremely high average TCRs of 96.61%, 98.60%, 98.45%, 98.39%, respectively. These results show the outstanding agentic ability of these models, which can correctly understand and conduct tool-calling. Correspondingly, these models also achieve better performance on SoMe, as demonstrated in Table 3. This implies that the ability to call tools and reason step-by-step is fundamental for social media agents.

Challenge Analysis

We conduct in-depth analyses of the typical errors to identify the major challenges for future research.

Errors in Planning Tool-chains. In complex social media tasks, the agents need to call tools multiple times in appropriate sequences to handle the task query. However, agents may encounter challenges in calling the tool-chains step-by-step. Therefore, we further investigate the probabilities of social media agents successfully calling the tool-chain in a single attempt. We report the One-attempt Success Rates (OSRs) in calling tool-chains on 100 samples of the real-time event detection task in Figure 3. In the figure, we also outline the rates of different cases where the agents fail to call the next tool correctly, i.e., failing to generate the correct formats of tool-calling, choose the right tool, or provide the correct parameters for the tool. Interestingly, we can observe that smaller models exhibit notably lower OSRs compared to

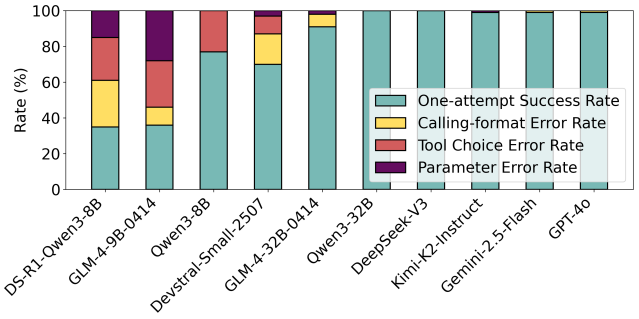


Figure 3: One-attempt success rates and error rates of different agentic LLMs in tool-chain planning.

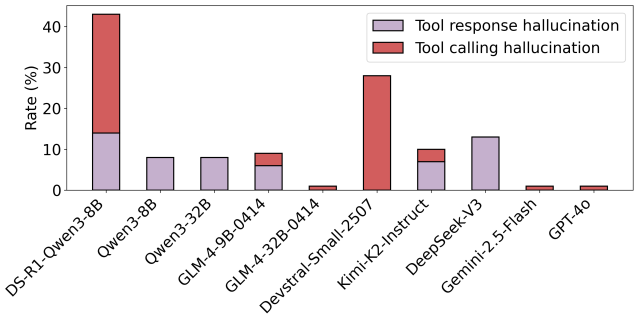
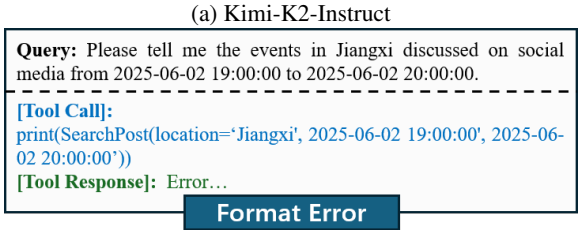


Figure 4: Hallucination rates in tool-use of different agentic LLMs.

larger models. Qwen3-32B, DeepSeek-V3 (with 671B parameters), Kimi-K2-Instruct (with 1T parameters), Gemini-2.5-Flash, and GPT-4o achieve nearly 100% success rates in accurately planning the tool-chains in one attempt. GLM4-32B-0414 also significantly improves OSR compared to GLM4-9B-0414. Small models generally struggle more in inferring the subsequent tool-calling based on past interactions. Therefore, the tool-planning capability of smaller LLMs should be a focal point in future work.

Hallucination in Tool-use In experiments, a common error scenario is observed where the agent fails to call the tool correctly and imagines the tool response or tool-calling format. For example, in real-time event detection, the agent sometimes fails to call the tools and imagine the events satisfying the task query. To further investigate the hallucination in tool-use, we design an experiment by intentionally altering the response of tools as errors and observing whether the agents will feign completing the task queries. The experiments are conducted on 100 samples of the real-time event detection task. As shown in Figure 4, most open-source LLMs (including DeepSeek-R1-0528-Qwen3-8B, Qwen3-8B, Qwen3-32B, GLM-4-9B-0414, Kimi-K2-Instruct, and DeepSeek-V3) suffer from the tool response hallucination, with the hallucination rates of 14%, 8%, 8%, 6%, 7%, and 13% individually. Interestingly, this issue persists regardless of the model size increase, as even Kimi-K2-Instruct (with 1T parameters) and DeepSeek-V3 (with 671B parameters) are affected by the hallucination. Moreover, the tool calling hallucination rates (which represent the rate of failing to call any tools due to the wrong format) of some open-source LLMs are also significantly high. Specifically, DeepSeek-R1-0528-Qwen3-8B and Devstral-Small-2507 imagine the wrong format of tool-calling and fail to call any tools with a percentage of 29% and 28% individually. However, it is noticeable that GLM-4-32B-0414 appears to address this prevalent limitation of open-source LLMs. These results indicate that overcoming hallucination in tool-use is a significant challenge for current open-source models.

(a) Kimi-K2-Instruct



(b) Gemini-2.5-Flash

Figure 5: Typical errors made by Kimi-K2-Instruct and Gemini-2.5-Flash. Blue texts are generated by agents.

Case Study We further recognize some common errors made by top agentic LLMs in Figure 5. As shown in (a), Kimi-K2-Instruct sometimes fails to repeat the keywords in tool response or query, leading to tool parameter errors. As shown in (b), a common error made by Gemini-2.5-Flash is adding *print(·)* outside the tool function. These errors may reflect inherent flaws in training these state-of-the-art agentic LLMs, which should be considered in future work.

Conclusion

This paper introduces SoMe, a novel benchmark for evaluating LLM-based social media agents. SoMe comprises a diverse collection of 8 social media agent tasks, 9,164,284 posts, 6,591 user profiles, and 25,686 reports from 32 social media platforms and external websites, with 17,869 meticulously annotated task queries. Our experiments reveal that both the current closed-source and open-source agentic LLMs cannot handle social media agent tasks satisfactorily. Moreover, we identify several challenges encountered by the current agents through in-depth experimental analysis, in order to inspire future work. By providing a challenging benchmark, we hope to stimulate the development of advanced models capable of tackling the complexities of social media agent tasks.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (No.2023YFC3310700), the Beijing Natural Science Foundation (JQ23018, L252032), and the National Natural Science Foundation of China (No.62276257).

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Baldwin, T.; Cook, P.; Lui, M.; MacKinlay, A.; and Wang, L. 2013. How noisy social media text, how diffrnt social media sources? In *Proceedings of the sixth international joint conference on natural language processing*, 356–364.

Carr, C. T.; and Hayes, R. A. 2015. Social media: Defining, developing, and divining. *Atlantic journal of communication*, 23(1): 46–65.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints*, arXiv–2407.

Edwards, C.; Edwards, A.; Spence, P. R.; and Shelton, A. K. 2014. Is that a bot running the social media feed? Testing the differences in perceptions of communication quality for a human agent and a bot agent on Twitter. *Computers in Human Behavior*, 33: 372–376.

Gao, C.; Xu, F.; Chen, X.; Wang, X.; He, X.; and Li, Y. 2024. Simulating human society with large language model agents: City, social media, and economic system. In *Companion Proceedings of the ACM Web Conference 2024*, 1290–1293.

GLM, T.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Zhang, D.; Rojas, D.; Feng, G.; Zhao, H.; et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

He, G.; Demartini, G.; and Gadiraju, U. 2025. Plan-then-execute: An empirical study of user trust and team performance when using llm agents as a daily assistant. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–22.

He, H.; Yao, W.; Ma, K.; Yu, W.; Dai, Y.; Zhang, H.; Lan, Z.; and Yu, D. 2024. WebVoyager: Building an End-to-End Web Agent with Large Multimodal Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6864–6890.

Hou, X.; Zhao, Y.; Wang, S.; and Wang, H. 2025. Model context protocol (mcp): Landscape, security threats, and future research directions. *arXiv preprint arXiv:2503.23278*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Islam, M. A.; Ali, M. E.; and Parvez, M. R. 2024. MapCoder: Multi-Agent Code Generation for Competitive Problem Solving. In *Annual Meeting of the Association of Computational Linguistics 2024*, 4912–4944. Association for Computational Linguistics (ACL).

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. arXiv:2310.06825.

Jiang, J.; and Ferrara, E. 2023. Social-llm: Modeling user behavior at scale using language models and social network data. *arXiv preprint arXiv:2401.00893*.

Kim, Y.; Park, C.; Jeong, H.; Chan, Y. S.; Xu, X.; McDuff, D.; Lee, H.; Ghassemi, M.; Breazeal, C.; and Park, H. W. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37: 79410–79452.

Koh, J. Y.; Lo, R.; Jang, L.; Duvvur, V.; Lim, M.; Huang, P.-Y.; Neubig, G.; Zhou, S.; Salakhutdinov, R.; and Fried, D. 2024. VisualWebArena: Evaluating Multimodal Agents on Realistic Visual Web Tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 881–905. Bangkok, Thailand: Association for Computational Linguistics.

Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C. H.; Gonzalez, J. E.; Zhang, H.; and Stoica, I. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Li, J.; Lai, Y.; Li, W.; Ren, J.; Zhang, M.; Kang, X.; Wang, S.; Li, P.; Zhang, Y.-Q.; Ma, W.; et al. 2024. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.

Li, K.; Zhang, Z.; Yin, H.; Zhang, L.; Ou, L.; Wu, J.; Yin, W.; Li, B.; Tao, Z.; Wang, X.; et al. 2025. WebSailor: Navigating Super-human Reasoning for Web Agent. *arXiv preprint arXiv:2507.02592*.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Liu, X.; Yu, H.; Zhang, H.; Xu, Y.; Lei, X.; Lai, H.; Gu, Y.; Ding, H.; Men, K.; Yang, K.; et al. 2024b. AgentBench: Evaluating LLMs as Agents. In *The Twelfth International Conference on Learning Representations*.

Mou, X.; Wei, Z.; and Huang, X.-J. 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*, 4789–4809.

Nakazato, T.; Onishi, M.; Suzuki, H.; and Shibuya, Y. 2024. JSocialFact: a misinformation dataset from social media for benchmarking LLM safety. In *2024 IEEE International Conference on Big Data (BigData)*, 3017–3025. IEEE.

Qian, S.; Chen, H.; Xue, D.; Fang, Q.; and Xu, C. 2023. Open-world social event classification. In *Proceedings of the ACM web conference 2023*, 1562–1571.

Qian, S.; Zhang, T.; and Xu, C. 2016. Multi-modal multi-view topic-opinion mining for social event analysis. In *Proceedings of the 24th ACM international conference on Multimedia*, 2–11.

Qian, S.; Zhou, Z.; Xue, D.; Wang, B.; and Xu, C. 2024. From linguistic giants to sensory maestros: A survey on cross-modal reasoning with large language models. *arXiv preprint arXiv:2409.18996*.

Qiao, B.; Li, K.; Zhou, W.; Li, S.; Lu, Q.; and Hu, S. 2025. BotSim: LLM-Powered Malicious Social Botnet Simulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 14377–14385.

Roy, S. D.; Mei, T.; Zeng, W.; and Li, S. 2012. Socialtransfer: cross-domain transfer learning from social streams for media applications. In *Proceedings of the 20th ACM international conference on Multimedia*, 649–658.

Skrynnik, A.; Andreychuk, A.; Borzilov, A.; Chernyavskiy, A.; Yakovlev, K.; and Panov, A. 2025. POGEMA: A Benchmark Platform for Cooperative Multi-Agent Pathfinding. In *The Thirteenth International Conference on Learning Representations*.

Tang, J.; Tang, J.; and Liu, H. 2014. Recommendation in social media: recent advances and new frontiers. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1977–1977.

Team, K. 2025. Kimi K2: Open Agentic Intelligence. Technical report, Moonshot AI.

Tu, Q.; Fan, S.; Tian, Z.; Shen, T.; Shang, S.; Gao, X.; and Yan, R. 2024. CharacterEval: A Chinese Benchmark for Role-Playing Conversational Agent Evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11836–11850. Bangkok, Thailand: Association for Computational Linguistics.

Wan, H.; Feng, S.; Tan, Z.; Wang, H.; Tsvetkov, Y.; and Luo, M. 2024. DELL: Generating Reactions and Explanations for LLM-Based Misinformation Detection. In *Findings of the Association for Computational Linguistics ACL 2024*, 2637–2667.

Wang, X.; Feng, M.; Qiu, J.; Gu, J.; and Zhao, J. 2024a. From news to forecast: Integrating event analysis in llm-based time series forecasting with reflection. *Advances in Neural Information Processing Systems*, 37: 58118–58153.

Wang, Y.; Xue, D.; Zhang, S.; and Qian, S. 2024b. BadAgent: Inserting and Activating Backdoor Attacks in LLM Agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9811–9827.

Wei, Q.; Xue, R.; Wang, Y.; Xiao, H.; Wang, Y.; and Duan, X. 2024. Mimicking the mavens: agent-based opinion synthesis and emotion prediction for social media influencers. *arXiv preprint arXiv:2407.20668*.

Xie, T.; Zhang, D.; Chen, J.; Li, X.; Zhao, S.; Cao, R.; Hua, T. J.; Cheng, Z.; Shin, D.; Lei, F.; et al. 2024. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37: 52040–52094.

Xue, D.; Qian, S.; Hu, C.; and Xu, C. 2025. Short-video Propagation Influence Rating: A New Real-world Dataset and A New Large Graph Model. *arXiv preprint arXiv:2503.23746*.

Xue, D.; Qian, S.; and Xu, C. 2024. Few-shot multimodal explanation for visual question answering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 1875–1884.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yang, Z.; Hu, Y.; Du, Z.; Xue, D.; Qian, S.; Wu, J.; Yang, F.; Dong, W.; and Xu, C. 2025b. SVBench: A Benchmark with Temporal Multi-Turn Dialogues for Streaming Video Understanding. In *The Thirteenth International Conference on Learning Representations*.

Yang, Z.; Ma, J.; Chen, H.; Lin, H.; Luo, Z.; and Chang, Y. 2022. A Coarse-to-fine Cascaded Evidence-Distillation Neural Network for Explainable Fake News Detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, 2608–2621.

Yang, Z.; Zhang, Z.; Zheng, Z.; Jiang, Y.; Gan, Z.; Wang, Z.; Ling, Z.; Chen, J.; Ma, M.; Dong, B.; et al. 2024. Oasis: Open agent social interaction simulations with one million agents. *arXiv preprint arXiv:2411.11581*.

Zhang, C.; Yang, Z.; Liu, J.; Li, Y.; Han, Y.; Chen, X.; Huang, Z.; Fu, B.; and Yu, G. 2025a. Appagent: Multimodal agents as smartphone users. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–20.

Zhang, K.; Li, J.; Li, G.; Shi, X.; and Jin, Z. 2024. CodeAgent: Enhancing Code Generation with Tool-Integrated Agent Systems for Real-World Repo-level Coding Challenges. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13643–13658.

Zhang, Z.; Lian, J.; Ma, C.; Qu, Y.; Luo, Y.; Wang, L.; Li, R.; Chen, X.; Lin, Y.; Wu, L.; et al. 2025b. TrendSim: Simulating Trending Topics in Social Media Under Poisoning Attacks with LLM-based Multi-agent System. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 2930–2949.