

# BotCourt: Towards Explainable Social Bot Detection via Collective Intelligence

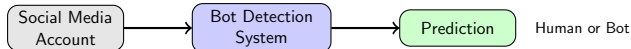
DING Bo

Department of Electrical Engineering, CityU

November 4, 2025

# Research Background

- ▶ Social bots pose significant threats to online social ecosystems
- ▶ Malicious purposes include:
  - ▶ Manipulating public discourse
  - ▶ Spreading disinformation
  - ▶ Interfering in elections and public health crises
- ▶ Need for social bot detection methodologies that are:
  - ▶ Accurate
  - ▶ Explainable
  - ▶ ...



**Social Bot Detection**

# Existing works for Social Bot Detection

## ▶ **Feature-based approaches**

- ▶ User metadata
- ▶ E.g., active days, follower/following count, posting frequency

## ▶ **Content-based approaches**

- ▶ User text
- ▶ E.g., tweet content, description, name

## ▶ **Graph-based approaches**

- ▶ Social structure, including user metadata, text, and neighbors
- ▶ Graph node (embedding based on metadata/text/...)
- ▶ Graph edge (relation based on follower/following/...)

# Challenges and Opportunities

## ▶ **Insufficient Explainability:**

- ▶ Existing methods focus on accuracy and often fail to provide explanations for predictions [1]
- ▶ Such as confidence or evidence for the decision

## ▶ **Threats Posed by LLM:**

- ▶ Social bots powered by LLM are more difficult to distinguish from human accounts in terms of content [2, 3]
- ▶ Increasing the decision-making risk of detection systems

## ▶ **Opportunities for LLM-based approaches:**

- ▶ LLMs is trained by lots of data, which can provide rich knowledge for social media bot detection
- ▶ LLMs can be used to provide explanations for decision-making, which is more interpretable for humans

# Preliminary Experiment: LLM for Bot Detection

**Key Insights:** The LLMs without fine-tuning showed competitive results compared to supervised learning baselines on the Twibot-22 dataset [4].

Model	Approach	Acc.	F1	Prec.	Rec.
SGBot	Baseline	0.623	0.395	<b>1.000</b>	0.247
BOTPERCENT		<u>0.731</u>	<b>0.726</b>	0.738	0.714
ROBERTA		0.633	0.432	0.955	0.280
BOTOMETER		<b>0.755</b>	0.585	0.440	<b>0.873</b>
BOTBUSTER		0.627	<u>0.439</u>	0.882	0.292
LOBO		0.552	0.198	0.944	0.110
RGT		0.509	0.509	0.323	<u>0.854</u>
Gemma-7b	tweet	0.515	0.525	0.514	0.535
	metadata	0.444	0.415	0.438	0.394
	description	0.521	0.519	0.521	0.518
	meta + desc	0.509	0.480	0.510	0.453
	structure	0.559	0.556	0.560	0.553
Mistral-v0.1-7b	tweet	0.447	0.580	0.468	0.765
	metadata	0.497	0.190	0.488	0.118
	description	0.521	0.546	0.519	0.577
	meta + desc	0.409	0.236	0.333	0.182
	structure	0.547	0.267	<b>0.700</b>	0.165
Mistral-v0.3-7b	tweet	0.553	0.487	0.571	0.424
	metadata	0.644	0.686	0.614	<b>0.777</b>
	description	0.485	0.249	0.460	0.171
	meta + desc	<b>0.668</b>	<b>0.700</b>	0.638	<b>0.777</b>
	structure	0.538	0.511	0.543	0.482
Qwen2.5-7b	tweet	0.532	0.413	0.555	0.329
	metadata	0.650	0.659	0.643	0.677
	description	<b>0.677</b>	0.686	<b>0.667</b>	0.706
	meta + desc	0.656	0.686	0.631	0.753
	structure	0.621	<b>0.706</b>	0.576	<b>0.912</b>

**Bold:** max value in Baseline; Underlined: 2nd max value in Baseline.

**Red:** max value among four LLMs; **Blue:** 2nd max value among four LLMs;

# Motivations

**Core Idea:** For social platforms, a social bot detection system should be an explainable decision-making support tool rather than a black box that can only output prediction results.

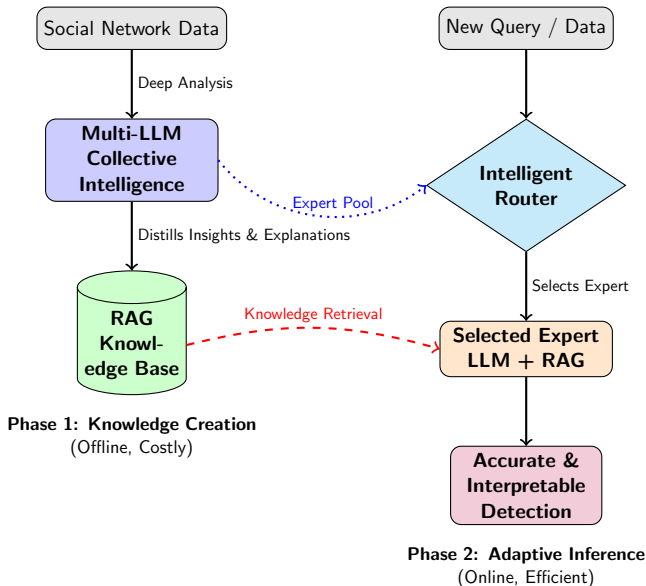
## 1. Collective Intelligence for Interpretable Evidence Mining

- ▶ Leverages a multi-LLM system where LLMs debate and review to dissect sophisticated bot behaviors [5, 6].
- ▶ Delivers high-quality, human-readable explanations for each detection, transforming the decision-making black box into an interpretable reasoning process.

## 2. Adaptive Inference via Expert Routing and RAG

- ▶ Implements an intelligent router to select the optimal expert LLM from the multi-LLM system based on the input sample.
- ▶ The selected expert then performs RAG-augmented inference [4], retrieving tailored reasoning patterns to ensure high accuracy and interpretability.

# Proposed Framework



# Core: Multi-LLM Collective Intelligence

**Question:** How to leverage collective intelligence for detection?

**Answer:** LLM-as-a-judge, Propose → Rebuttal → Judge




## Plan 1: Multi-LLM System [5]

- ▶ **Multiple specialized LLMs**
  - ▶ Each LLM has a distinct perspective/knowledge
  - ▶ E.g., different LLM, different pretrained dataset
- ▶ **Collective decision-making**
  - ▶ Multi-LLM debate and review mechanism

## Plan 2: Single LLM with Multiple Roles [7]

- ▶ **Single LLM with role-playing**
  - ▶ Same LLM, different role-based prompts
  - ▶ E.g., proposer, rebutter, judge roles
- ▶ **Virtual multi-llm debate**
  - ▶ Role-based single-LLM debate and review

# References I

-  Q. Wu, Y. Yang, H. Peng, B. He, Y. Xia, Y. Liao, *et al.*, “Certainly bot or not? trustworthy social bot detection via robust multi-modal neural processes,” *arXiv preprint arXiv:2503.09626*, 2025.
-  F. Kong, X. Zhang, X. Chen, Y. Yang, S.-C. Zhu, and X. Feng, “Enhancing llm-based social bot via an adversarial learning framework,” *arXiv preprint arXiv:2508.17711*, 2025.
-  B. Qiao, K. Li, W. Zhou, S. Li, Q. Lu, and S. Hu, “Botsim: Llm-powered malicious social botnet simulation,” in *AAAI*, 2025.
-  S. Feng, H. Wan, N. Wang, Z. Tan, M. Luo, and Y. Tsvetkov, “What does the bot say? opportunities and risks of large language models in social media bot detection,” in *ACL*, 2024.

## References II

-  Y. Jiang, W. Ding, S. Feng, G. Durrett, and Y. Tsvetkov, “Sparta alignment: Collectively aligning multiple language models through combat,” *NeurIPS*, 2025.
-  S. Feng, Z. Wang, P. Goyal, Y. Wang, W. Shi, H. Xia, H. Palangi, L. Zettlemoyer, Y. Tsvetkov, C.-Y. Lee, *et al.*, “Heterogeneous swarms: Jointly optimizing model roles and weights for multi-llm systems,” *NeurIPS*, 2025.
-  Y. Chen, Y. Wang, S. Zhu, H. Yu, T. Feng, M. Zhang, M. Patwary, and J. You, “Multi-agent evolve: Llm self-improve through co-evolution,” 2025.