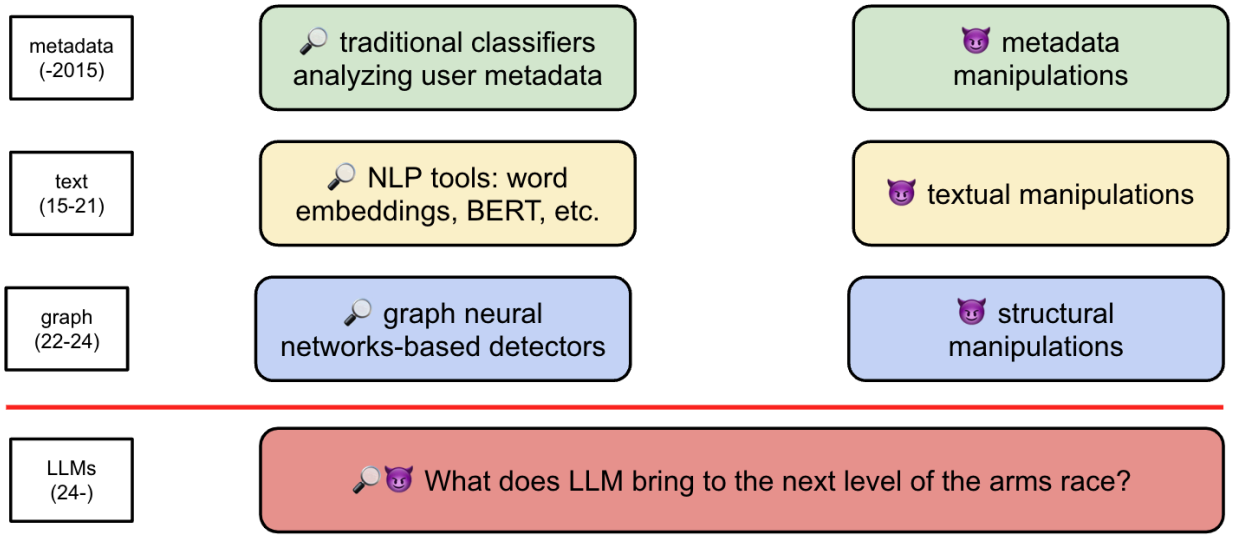
LLM-Powered Social Bot Detectors

Social media bot accounts are behind many online perils such as misinformation, election interference, and conspiracy theories. Research on detecting social media bots has always been an arms race [1]: early methods focus on analyzing user metadata with machine learning classifiers, while bot operators manipulate user features to evade detection; later approaches employed word embeddings and encoder-based language models to characterize user texts, while bot operators re-post genuine content to dilute malicious content and appear innocuous; recent models tap into the network information of user interactions with graph neural networks, while advanced bots strategically follow and unfollow users to appear out-of-distribution.

Recent advances in language technologies brought up large language models: powerful tools that excel in generating texts and following instructions. Still, they also come with risks and biases that could cause real-world harm [2]. We argue that there is significant risk in potential LLM-powered social media bot accounts, which necessitate equally strong LLM-powered bot detectors to counter.



We develop LLM-based bot detectors by proposing a mixture-of-heterogeneous-experts framework to tackle the diverse user information. Specifically, different user information modalities are separately analyzed with LLMs while majority voting is conducted to ensemble uni-modality predictions. LLMs are fine-tuned-based instructions to analyze user texts, metadata, and structural information, while these modality-specific LLM experts participate in prediction ensemble or multi-agent discussions to make a final decision about social bots.

We conduct a preliminary analysis of the tool to demonstrate its effectiveness. LLM-based detectors achieve state-of-the-art performance on two bot detection datasets [3], outperforming the second-best by 2.6% and 9.1% in classification accuracy. We also find that these approaches are stronger when the underlying LLM is better: GPT-4 is better than LLaMA2,

indicating that the gap between open and closed models persists in the task of social bot detections.

Aside from that, we also study the adversarial nature of LLMs and bots: what if LLM-based detectors face off against LLM-powered bots? We found that existing bot detectors suffer from quality drops when encountering LLM-powered bots, evident in the 7-11% performance drops in classification accuracy. However, with LLM-powered bots we only see a drop of 2.3% on average, indicating that LLMs might be able to identify artifacts generated by themselves [4] and hence enhance bot detection against LLM-powered bots. We should have hope: even if LLM-powered bots are more evasive, we find that LLMs themselves could be useful tools against their own creation if adapted to the right task.

If a model like this were to be further improved and deployed on real-world social media platforms, we would expect to see a drop in bot and AI-generated content, which could further need less polarization, more genuine conversation, and a strengthened democratic process.

LLM-powered social bot detectors also come with ethical considerations: Language models have been extensively documented to have inherent social biases and such biases could have an impact on downstream tasks [5]: social bot detection would be no exception. We hypothesize that LLM-based bot detectors might underserve certain users and communities, potentially informed by LLMs' internal biases, stereotypes, and spurious correlations. We argue that the decisions of LLM-based bot detectors should be interpreted as an initial screening of malicious accounts, while content moderation decisions should be made with humans in the loop. Future work could also investigate the fairness implications of social media bot detectors based on LLMs and other machine learning models.

We highlight that actionable items for the general public to counter LLM-powered bots could be: 1) be vigilant about AI-generated content online. Not everything said or shared online is genuine, so be extra careful when you interact with shady personas; 2) report suspicious content. This would help the model developed to gain more raw data, human feedback, and real-world labels to work with, greatly enhancing their capability to stay up to date about current threat modes and develop quick counter solutions.

[1] Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2017, April). 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 (pp. 963-972).

[2] Kumar, S., Balachandran, V., Njoo, L., Anastasopoulos, A., & Tsvetkov, Y. (2023, May). Language Generation Models Can Cause Harm: So What Can We Do About It? An Actionable Survey. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (pp. 3299-3321).

[3] Feng, S., Tan, Z., Wan, H., Wang, N., Chen, Z., Zhang, B., ... & Luo, M. (2022). Twibot-22: Towards graph-based twitter bot detection. Advances in Neural Information Processing Systems, 35, 35254-35269.

[4] Pu, X., Zhang, J., Han, X., Tsvetkov, Y., & He, T. (2023, December). On the Zero-Shot Generalization of Machine-Generated Text Detectors. In Findings of the Association for Computational Linguistics: EMNLP 2023 (pp. 4799-4808).

[5] Feng, S., Park, C. Y., Liu, Y., & Tsvetkov, Y. (2023, July). From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 11737-11762).