A Comprehensive Study on LLM Agent Challenges

Palash Ingle[1], Mithun Parab[2], Pranay Lendave[3], Amisha Bhanushali[1], Pavan Kumar B N [4]

[1]Sejong University, Seoul, South Korea
[2]R.J. College, Mumbai, India
[3]K J Somaiya College of Engineering, Mumbai, India
[4]Indian Institute of Information Technology, Sri city, India
palashngl@gmail.com, {mithun.sp, pranay.lendave}@somaiya.edu, bhanushali.amisha29@gmail.com, pavanbn8@gmail.com

Abstract

This paper intricately examines the manifold challenges and inherent issues associated with Large Language Models (LLMs), both for the models themselves and the human context. It systematically investigates adversarial vulnerability and reliability in LLM agents, emphasizing the intricate surroundings of these potent language generation systems. Beyond these fundamental concerns, the paper delves into the exploitation of LLM agents, highlighting the significant problem of misuse, and addresses the formidable challenges linked to the scaling of LLM agents. It also scrutinizes open problems, focusing on migrating AI agents from virtual simulations to physical realities and exploring collaborative cognition in LLM agents. Through this comprehensive analysis, the paper contributes to a holistic understanding of the hurdles faced by LLM agents, fostering the ongoing discourse on responsible AI development.

Introduction

As humans continuously seek autonomy to alleviate challenging tasks, the term "agent" in AI research signifies entities with intelligent behavior, autonomy, reactivity, proactiveness, and social ability. LLM agents, being personalized and user-aligned, are pivotal in the future landscape, addressing the longstanding issue of agency by seamlessly integrating intelligent capabilities. LLM agents are poised to revolutionize tasks, providing efficient and tailored solutions, making them essential technological drivers in the evolving realm of human-AI interactions. This paper traverses through diverse sections, commencing with an exploration of adversarial vulnerability and reliability in LLM agents, progressing to the risks associated with LLM agent exploitation and the challenges of scaling these models. It then delves into the nuanced transition from virtual to physical environments and concludes with an examination of the emerging realm of collective cognitive fusion in LLM agents.

Adversarial Vulnerability

Persistent challenges in deep learning (Madry et al. 2019; Zheng et al. 2023; Zhiheng, Rui, and Tao 2023) encompass adversarial attacks across computer vision (Madry et al. 2019; Akhtar and Mian 2018), natural language processing (Wang, Wang, and Yang 2022; Li et al. 2019; Zhu et al. 2020; Xi et al. 2022), and reinforcement learning (Pinto et al. 2017; Rigter, Lacerda, and Hawes 2022; Panaganti et al. 2022) domains. Studies indicate vulnerabilities in pre-trained language models (PLMs) (Jin et al. 2020; Li et al. 2019; Ren et al. 2019), affecting large language models (LLMs) and impeding LLM-based agents (Zhu et al. 2023; Chen et al. 2023a). Various attack vectors, such as dataset poisoning, backdoor attacks (Chen et al. 2021; Li et al. 2021), and prompt-specific attacks (Shi et al. 2022; Perez and Ribeiro 2022), induce harmful content(Liang et al. 2023; Gururangan et al. 2022; Liu et al. 2023). Adversarial attacks on LLM-based agents (Xi et al. 2023), carry societal risks, compelling potentially destructive actions. When perception model encounters challenges when subjected to adversarial inputs, such as perturbed images (Akhtar and Mian 2018), audio signals (Carlini and Wagner 2018), or text input. This susceptibility within LLM agents can result in actions that deviate from the intended course. Traditional techniques like adversarial training (Madry et al. 2019; Zhu et al. 2020), data augmentation (Morris et al. 2020; Si et al. 2021) , and sample detection (Yoo et al. 2022; Le, Park, and Lee 2021) enhance the resilience of LLM-based agents. However, securing all modules while preserving utility poses a complex challenge (Tsipras et al. 2019; Zhang et al. 2019). The human-in-the-loop strategy entails human supervision to oversee agent behavior and provide feedback (Kenton et al. 2021; Du et al. 2022; Cai, Chang, and Han 2023). The Sleeper agent (Souri et al. 2022) paper reveals a safety threat wherein attackers distribute specialized text online with trigger phrases to poison base models. This sophisticated attack, using obfuscation techniques, resembles zero-day vulnerability markets, underscoring the limitations of safety fine-tuning in ensuring uniform LLM safety.

Reliability in LLM agents

Establishing trust in deep learning (Wong, Wang, and Hryniowski 2021; Huang et al. 2020, 2023) is challenging due to the obscured factors contributing to the success of deep neural networks (Brown et al. 2020; Devlin et al. 2019). Large Language Models (LLMs), akin to neural networks, struggle to express prediction certainty (Huang et al. 2023; Chen et al. 2023b), leading to calibration issues in practical

AAAI 2024 Spring Symposium on User-Aligned Assessment of Adaptive AI Systems. Stanford University, Stanford, CA, USA.

applications. Dynamic interactions pose a risk of agent outputs deviating from human intentions, and biases from training data may introduce partial content in LLM applications, with potential societal implications (Bolukbasi et al. 2016; Caliskan, Bryson, and Narayanan 2017). Language models display noticeable hallucination problems, diminishing reliability by generating inaccuracies (Ji et al. 2023; Mündler et al. 2023). The current emphasis underscores the need for intelligent agents characterized by honesty and reliability. Ongoing research strives to boost credibility by elucidating thought processes(Wei et al. 2023; Kojima et al. 2023) in models and integrating external knowledge bases.

LLM Agent Exploitation Risk

Capitalizing on their intricate capabilities, LLM-based agents exhibit a vast range of functionalities (Yang, Yue, and He 2023; Chase 2022). Yet, individuals with nefarious motives can wield these agents as instruments, posing substantial threats to both individuals and society (Brundage et al. 2018). This misuse potential extends to malevolent manipulation of public opinion, dissemination of false information, compromise of cybersecurity, fraudulent activities, and even orchestration of terrorism. Consequently, the imperative arises to institute robust regulatory frameworks governing the ethical deployment of LLM-based agents (Bai et al. 2022; Wang et al. 2021). Fostering responsible use necessitates technological companies to fortify the security architecture of these systems. Emphasizing the need for vigilance, these agents should undergo training to adeptly discern and reject requests indicative of malicious intent during their learning phase.

Scaling LLM agents

While showcasing superior performance in task-oriented applications and the simulation of diverse social phenomena, LLMs have primarily been investigated with a limited number of agents in current research, neglecting efforts to scale up for more complex systems or larger societal simulations (Zhuge et al. 2023; Bai, Zhang, and Chen 2023). Expanding the agent count holds promise for introducing greater specialization, enhancing efficiency in handling intricate tasks like software development or government policy formulation (Qian et al. 2023). Augmenting agents in social simulations boosts credibility and realism (Park et al. 2023), offering insights into societal functioning, breakdowns, and potential risks. This expanded scope allows for tailored interventions in societal operations, facilitating observations of specific conditions, such as the occurrence of black swan events, and their impact on societal states.

Expanding the agent count can enhance task efficiency and the authenticity of social simulations (Park et al. 2023; Qian et al. 2023; Williams et al. 2023), but challenges loom. Managing the computational load posed by numerous AI agents necessitates refined architectural design and computational optimization. Increased agent population heightens communication complexities, hindering efficient message propagation and elevating the risk of biased information dissemination in multi-agent systems. As numbers grow, the threat of unreliable communication and distorted information exchange intensifies.

Virtual vs Physical Environment

A substantial disparity exists between virtual simulation settings and the tangible physical realm. Virtual environments are confined to specific scenes and task-oriented, engaging in simulated interactions (Zhou et al. 2023; Shridhar et al. 2021), whereas real-world settings encompass diverse tasks and involve physical interactions. Consequently, agents must confront challenges arising from external influences and their intrinsic capabilities, necessitating adeptness in navigating the intricate dynamics of the physical world.

Deploying agents in the physical realm poses a significant challenge, demanding adaptable hardware support. While a simulated setting ensures reliable outcomes, transitioning to reality risks hardware inadequacies impacting task efficiency.

To successfully transition into the real physical world, an agent must demonstrate improved environmental generalization skills. It must comprehend and deduce meanings from ambiguous instructions with implied nuances and exhibit the capacity to learn and adapt to new skills in a flexible manner (Wang et al. 2023; Colas et al. 2023). Grappling with the complexities of an infinite and open world presents challenges due to the agent's constrained context (Bertsch et al. 2023; Chowdhury and Caragea 2023). The agent's effectiveness hinges on its ability to manage copious amounts of information from the world and navigate seamlessly within this expansive and dynamic environment.

Collective Cognitive Fusion

Collective intelligence, rooted in shared or group intellect, plays a crucial role in decision-making, drawing insights from diverse opinions. This phenomenon, observed in various domains, relies on effective coordination to avoid issues like "groupthink" and individual biases, fostering cooperation. In the context of LLM agents, achieving equilibrium is essential for optimizing collective intelligence and ensuring stability. This analysis highlights the challenges in decision-making, emphasizing the need for a balance where changes benefit an agent without disrupting the overall equilibrium. Understanding and addressing these dynamics is vital for achieving the optimal state in multi LLM agents systems.

Conclusion and Discussion

In the preceding sections, we provided insights into key challenges and risks associated with LLM agents, encompassing adversarial vulnerability, reliability, misuse, and collective cognition. However, a comprehensive exploration reveals additional critical concerns. The looming threat of unemployment, potential risks to human well-being, the ongoing debate on the trajectory towards AGI, and the emerging paradigm of LLM-based Agent as a Service demand thorough examination. These multifaceted dimensions underscore the complex landscape of LLM technologies, necessitating a holistic understanding and proactive measures for their responsible development and deployment.

References

Akhtar, N.; and Mian, A. 2018. Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. arXiv:1801.00553.

Bai, J.; Zhang, S.; and Chen, Z. 2023. Is There Any Social Principle for LLM-Based Agents? arXiv:2308.11136.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; Chen, C.; Olsson, C.; Olah, C.; Hernandez, D.; Drain, D.; Ganguli, D.; Li, D.; Tran-Johnson, E.; Perez, E.; Kerr, J.; Mueller, J.; Ladish, J.; Landau, J.; Ndousse, K.; Lukosuite, K.; Lovitt, L.; Sellitto, M.; Elhage, N.; Schiefer, N.; Mercado, N.; DasSarma, N.; Lasenby, R.; Larson, R.; Ringer, S.; Johnston, S.; Kravec, S.; Showk, S. E.; Fort, S.; Lanham, T.; Telleen-Lawton, T.; Conerly, T.; Henighan, T.; Hume, T.; Bowman, S. R.; Hatfield-Dodds, Z.; Mann, B.; Amodei, D.; Joseph, N.; McCandlish, S.; Brown, T.; and Kaplan, J. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.

Bertsch, A.; Alon, U.; Neubig, G.; and Gormley, M. R. 2023. Unlimiformer: Long-Range Transformers with Unlimited Length Input. arXiv:2305.01625.

Bolukbasi, T.; Chang, K.-W.; Zou, J.; Saligrama, V.; and Kalai, A. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. arXiv:1607.06520.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165.

Brundage, M.; Avin, S.; Clark, J.; Toner, H.; Eckersley, P.; Garfinkel, B.; Dafoe, A.; Scharre, P.; Zeitzoff, T.; Filar, B.; Anderson, H.; Roff, H.; Allen, G.; Steinhardt, J.; Flynn, C.; hÉigeartaigh, S.; Beard, S.; Belfield, H.; Farquhar, S.; and Amodei, D. 2018. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation.

Cai, Z.; Chang, B.; and Han, W. 2023. Human-in-the-Loop through Chain-of-Thought. arXiv:2306.07932.

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183–186.

Carlini, N.; and Wagner, D. 2018. Audio Adversarial Examples: Targeted Attacks on Speech-to-Text. arXiv:1801.01944.

Chase, H. 2022. LangChain.

Chen, X.; Salem, A.; Chen, D.; Backes, M.; Ma, S.; Shen, Q.; Wu, Z.; and Zhang, Y. 2021. BadNL: Backdoor Attacks against NLP Models with Semantic-preserving Improvements. In *Annual Computer Security Applications Conference*, ACSAC '21. ACM.

Chen, X.; Ye, J.; Zu, C.; Xu, N.; Zheng, R.; Peng, M.; Zhou, J.; Gui, T.; Zhang, Q.; and Huang, X. 2023a. How Robust is GPT-3.5 to Predecessors? A Comprehensive Study on Language Understanding Tasks. arXiv:2303.00293.

Chen, Y.; Yuan, L.; Cui, G.; Liu, Z.; and Ji, H. 2023b. A Close Look into the Calibration of Pre-trained Language Models. arXiv:2211.00151.

Chowdhury, J. R.; and Caragea, C. 2023. Monotonic Location Attention for Length Generalization. arXiv:2305.20019.

Colas, C.; Teodorescu, L.; Oudeyer, P.-Y.; Yuan, X.; and Côté, M.-A. 2023. Augmenting Autotelic Agents with Large Language Models. arXiv:2305.12487.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

Du, W.; Kim, Z. M.; Raheja, V.; Kumar, D.; and Kang, D. 2022. Read, Revise, Repeat: A System Demonstration for Human-in-the-loop Iterative Text Revision. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*. Association for Computational Linguistics.

Gururangan, S.; Card, D.; Dreier, S.; Gade, E.; Wang, L.; Wang, Z.; Zettlemoyer, L.; and Smith, N. A. 2022. Whose Language Counts as High Quality? Measuring Language Ideologies in Text Data Selection. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2562–2580. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Huang, X.; Kroening, D.; Ruan, W.; Sharp, J.; Sun, Y.; Thamo, E.; Wu, M.; and Yi, X. 2020. A Survey of Safety and Trustworthiness of Deep Neural Networks: Verification, Testing, Adversarial Attack and Defence, and Interpretability. arXiv:1812.08342.

Huang, X.; Ruan, W.; Huang, W.; Jin, G.; Dong, Y.; Wu, C.; Bensalem, S.; Mu, R.; Qi, Y.; Zhao, X.; Cai, K.; Zhang, Y.; Wu, S.; Xu, P.; Wu, D.; Freitas, A.; and Mustafa, M. A. 2023. A Survey of Safety and Trustworthiness of Large Language Models through the Lens of Verification and Validation. arXiv:2305.11391.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12): 1–38.

Jin, D.; Jin, Z.; Zhou, J. T.; and Szolovits, P. 2020. Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment. arXiv:1907.11932.

Kenton, Z.; Everitt, T.; Weidinger, L.; Gabriel, I.; Mikulik, V.; and Irving, G. 2021. Alignment of Language Agents. arXiv:2103.14659.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2023. Large Language Models are Zero-Shot Reasoners. arXiv:2205.11916.

Le, T.; Park, N.; and Lee, D. 2021. A Sweet Rabbit Hole by DARCY: Using Honeypots to Detect Universal Trigger's

Adversarial Attacks. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3831–3844. Online: Association for Computational Linguistics.

Li, J.; Ji, S.; Du, T.; Li, B.; and Wang, T. 2019. TextBugger: Generating Adversarial Text Against Real-world Applications. In *Proceedings 2019 Network and Distributed System Security Symposium*, NDSS 2019. Internet Society.

Li, Z.; Mekala, D.; Dong, C.; and Shang, J. 2021. BFClass: A Backdoor-free Text Classification Framework. arXiv:2109.10855.

Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; Newman, B.; Yuan, B.; Yan, B.; Zhang, C.; Cosgrove, C.; Manning, C. D.; Ré, C.; Acosta-Navas, D.; Hudson, D. A.; Zelikman, E.; Durmus, E.; Ladhak, F.; Rong, F.; Ren, H.; Yao, H.; Wang, J.; Santhanam, K.; Orr, L.; Zheng, L.; Yuksekgonul, M.; Suzgun, M.; Kim, N.; Guha, N.; Chatterji, N.; Khattab, O.; Henderson, P.; Huang, Q.; Chi, R.; Xie, S. M.; Santurkar, S.; Ganguli, S.; Hashimoto, T.; Icard, T.; Zhang, T.; Chaudhary, V.; Wang, W.; Li, X.; Mai, Y.; Zhang, Y.; and Koreeda, Y. 2023. Holistic Evaluation of Language Models. arXiv:2211.09110.

Liu, Y.; Deng, G.; Li, Y.; Wang, K.; Zhang, T.; Liu, Y.; Wang, H.; Zheng, Y.; and Liu, Y. 2023. Prompt Injection attack against LLM-integrated Applications. arXiv:2306.05499.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2019. Towards Deep Learning Models Resistant to Adversarial Attacks. arXiv:1706.06083.

Morris, J. X.; Lifland, E.; Yoo, J. Y.; Grigsby, J.; Jin, D.; and Qi, Y. 2020. TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. arXiv:2005.05909.

Mündler, N.; He, J.; Jenko, S.; and Vechev, M. 2023. Self-contradictory Hallucinations of Large Language Models: Evaluation, Detection and Mitigation. arXiv:2305.15852.

Panaganti, K.; Xu, Z.; Kalathil, D.; and Ghavamzadeh, M. 2022. Robust Reinforcement Learning using Offline Data. arXiv:2208.05129.

Park, J. S.; O'Brien, J. C.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior. arXiv:2304.03442.

Perez, F.; and Ribeiro, I. 2022. Ignore Previous Prompt: Attack Techniques For Language Models. arXiv:2211.09527.

Pinto, L.; Davidson, J.; Sukthankar, R.; and Gupta, A. 2017. Robust Adversarial Reinforcement Learning. arXiv:1703.02702.

Qian, C.; Cong, X.; Liu, W.; Yang, C.; Chen, W.; Su, Y.; Dang, Y.; Li, J.; Xu, J.; Li, D.; Liu, Z.; and Sun, M. 2023. Communicative Agents for Software Development. arXiv:2307.07924.

Ren, S.; Deng, Y.; He, K.; and Che, W. 2019. Generating Natural Language Adversarial Examples through Probability Weighted Word Saliency. In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1085–1097. Florence, Italy: Association for Computational Linguistics.

Rigter, M.; Lacerda, B.; and Hawes, N. 2022. RAMBO-RL: Robust Adversarial Model-Based Offline Reinforcement Learning. arXiv:2204.12581.

Shi, Y.; Li, P.; Yin, C.; Han, Z.; Zhou, L.; and Liu, Z. 2022. PromptAttack: Prompt-based Attack for Language Models via Gradient Search. arXiv:2209.01882.

Shridhar, M.; Yuan, X.; Côté, M.-A.; Bisk, Y.; Trischler, A.; and Hausknecht, M. 2021. ALFWorld: Aligning Text and Embodied Environments for Interactive Learning. arXiv:2010.03768.

Si, C.; Zhang, Z.; Qi, F.; Liu, Z.; Wang, Y.; Liu, Q.; and Sun, M. 2021. Better Robustness by More Coverage: Adversarial Training with Mixup Augmentation for Robust Fine-tuning. arXiv:2012.15699.

Souri, H.; Fowl, L.; Chellappa, R.; Goldblum, M.; and Goldstein, T. 2022. Sleeper Agent: Scalable Hidden Trigger Backdoors for Neural Networks Trained from Scratch. arXiv:2106.08970.

Tsipras, D.; Santurkar, S.; Engstrom, L.; Turner, A.; and Madry, A. 2019. Robustness May Be at Odds with Accuracy. arXiv:1805.12152.

Wang, G.; Xie, Y.; Jiang, Y.; Mandlekar, A.; Xiao, C.; Zhu, Y.; Fan, L.; and Anandkumar, A. 2023. Voyager: An Open-Ended Embodied Agent with Large Language Models. arXiv:2305.16291.

Wang, X.; Wang, H.; and Yang, D. 2022. Measure and Improve Robustness in NLP Models: A Survey. arXiv:2112.08313.

Wang, Z. J.; Choi, D.; Xu, S.; and Yang, D. 2021. Putting Humans in the Natural Language Processing Loop: A Survey. arXiv:2103.04044.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903.

Williams, R.; Hosseinichimeh, N.; Majumdar, A.; and Ghaffarzadegan, N. 2023. Epidemic Modeling with Generative Agents. arXiv:2307.04986.

Wong, A.; Wang, X. Y.; and Hryniowski, A. 2021. How Much Can We Really Trust You? Towards Simple, Interpretable Trust Quantification Metrics for Deep Neural Networks. arXiv:2009.05835.

Xi, Z.; Chen, W.; Guo, X.; He, W.; Ding, Y.; Hong, B.; Zhang, M.; Wang, J.; Jin, S.; Zhou, E.; Zheng, R.; Fan, X.; Wang, X.; Xiong, L.; Zhou, Y.; Wang, W.; Jiang, C.; Zou, Y.; Liu, X.; Yin, Z.; Dou, S.; Weng, R.; Cheng, W.; Zhang, Q.; Qin, W.; Zheng, Y.; Qiu, X.; Huang, X.; and Gui, T. 2023. The Rise and Potential of Large Language Model Based Agents: A Survey. arXiv:2309.07864.

Xi, Z.; Zheng, R.; Gui, T.; Zhang, Q.; and Huang, X. 2022. Efficient Adversarial Training with Robust Early-Bird Tickets. arXiv:2211.07263.

Yang, H.; Yue, S.; and He, Y. 2023. Auto-GPT for Online Decision Making: Benchmarks and Additional Opinions. arXiv:2306.02224.

Yoo, K.; Kim, J.; Jang, J.; and Kwak, N. 2022. Detection of Adversarial Examples in Text Classification: Benchmark and Baseline via Robust Density Estimation. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Findings of the Association for Computational Linguistics: ACL 2022*, 3656–3672. Dublin, Ireland: Association for Computational Linguistics.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E. P.; Ghaoui, L. E.; and Jordan, M. I. 2019. Theoretically Principled Trade-off between Robustness and Accuracy. arXiv:1901.08573.

Zheng, R.; Xi, Z.; Liu, Q.; Lai, W.; Gui, T.; Zhang, Q.; Huang, X.; Ma, J.; Shan, Y.; and Ge, W. 2023. Characterizing the Impacts of Instances on Robustness. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 2314–2332. Toronto, Canada: Association for Computational Linguistics.

Zhiheng, X.; Rui, Z.; and Tao, G. 2023. Safety and ethical concerns of large language models. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 4: Tutorial Abstracts)*, 9–16.

Zhou, S.; Xu, F. F.; Zhu, H.; Zhou, X.; Lo, R.; Sridhar, A.; Cheng, X.; Ou, T.; Bisk, Y.; Fried, D.; Alon, U.; and Neubig, G. 2023. WebArena: A Realistic Web Environment for Building Autonomous Agents. arXiv:2307.13854.

Zhu, C.; Cheng, Y.; Gan, Z.; Sun, S.; Goldstein, T.; and Liu, J. 2020. FreeLB: Enhanced Adversarial Training for Natural Language Understanding. arXiv:1909.11764.

Zhu, K.; Wang, J.; Zhou, J.; Wang, Z.; Chen, H.; Wang, Y.; Yang, L.; Ye, W.; Zhang, Y.; Gong, N. Z.; and Xie, X. 2023. PromptBench: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts. arXiv:2306.04528.

Zhuge, M.; Liu, H.; Faccio, F.; Ashley, D. R.; Csordás, R.; Gopalakrishnan, A.; Hamdi, A.; Hammoud, H. A. A. K.; Herrmann, V.; Irie, K.; Kirsch, L.; Li, B.; Li, G.; Liu, S.; Mai, J.; Piekos, P.; Ramesh, A.; Schlag, I.; Shi, W.; Stanić, A.; Wang, W.; Wang, Y.; Xu, M.; Fan, D.-P.; Ghanem, B.; and Schmidhuber, J. 2023. Mindstorms in Natural Language-Based Societies of Mind. arXiv:2305.17066.