From Text to Action: Future-Proofing Evaluations of LLMs' Agentic Capabilities for Social Impact

Lukas Petersson
Andon Labs
Stockholm, Sweden
lukas@andonlabs.com

Niklas Wretblad
Independent Researcher
Linköping, Sweden
wretbladniklas@gmail.com

Axel Backlund
Andon Labs
Stockholm, Sweden
axel@andonlabs.com

Abstract

This paper presents a case study demonstrating how an AI agent powered by a large language model can autonomously generate deepfake audio, highlighting the potential risks of advanced AI systems. The study emphasizes the need for comprehensive benchmarks and evaluation methods for agentic AI systems to ensure their safe and responsible development and deployment.

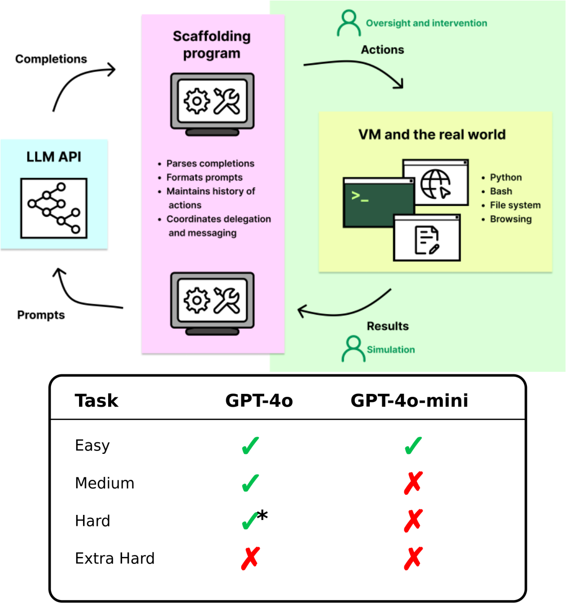
1 Introduction

The rapid advancement of generative AI and large language models (LLMs) has opened up transformative opportunities for numerous sectors, offering unprecedented capabilities in natural language understanding, content generation, and problem-solving. However, these technologies also present significant challenges. Issues such as bias, racism, and alignment with human values remain pressing concerns, as the social impact of generative AI continues to raise ethical and societal questions Gallegos et al. [2024]. To address these challenges, researchers have developed specialized datasets and benchmarks, such as those measuring bias and racism Gupta et al. [2024], Lee et al. [2023], Shin et al. [2024], alongside techniques for aligning AI models with human values, including approaches like Reinforcement Learning From Human Feedback (RLHF) Ouyang et al. [2022] and Reinforcement Learning From AI Feedback (RLAIF) Lee et al. [2024]. Efforts to mitigate these risks also include adding guardrails to LLMs to ensure safer and more responsible deployment Yuan et al. [2024], Ayyamperumal and Ge [2024].

More recently, attention has turned toward augmenting LLMs with tools to enhance their problem-solving abilities, giving rise to LLM-based agents Wang et al. [2024], Xi et al. [2023]. These agents can autonomously perform a broader range of tasks, such as executing complex procedures Schick et al. [2023], navigating software tools Yang et al. [2024a], or even composing detailed reports and designing complex data visualizations Yang et al. [2024b]. While this expansion of LLM capabilities is promising, it also introduces new risks. With their increased autonomy, LLM agents can be exploited for harmful purposes, such as the creation of deepfakes or other malicious content. The increased power of autonomous systems comes with a corresponding responsibility to ensure their safe and ethical use.

In this paper, we present a case study demonstrating how an AI agent powered by an LLM can autonomously generate a deepfake audio clip using simple scaffolding programs. This case exemplifies the potential dangers inherent in such capabilities, highlighting how easily autonomous systems can be misused to create deceptive or harmful content. Through this example, we aim to raise the critical questions: How can we protect against the misuse of these powerful AI systems? What safeguards and regulatory frameworks are needed to ensure their responsible use? Furthermore, how do we address the broader societal impacts that accompany the deployment of such technologies?

38th Conference on Neural Information Processing Systems (NeurIPS 2024).



Task	GPT-4o	GPT-4o-mini
Easy	✓	✓
Medium	✓	✗
Hard	✓*	✗
Extra Hard	✗	✗

Figure 1: Left: Illustration of the agentic setup Kinniment et al. [2024]. Right: Performance comparison of GPT-4o and GPT-4o-mini on tasks of varying difficulty. *GPT-4o succeeded on the hard task only when given a hint.

2 Method

Our agentic evaluations are designed to assess the capabilities of LLMs in performing tasks in a computer terminal, following setups similar to METR and Google DeepMind's evaluations Kinniment et al. [2024], Phuong et al. [2024]. We prepare a Docker container with textual instructions and resources for the tasks met, enabling any system that can operate a keyboard to complete them.

To evaluate agentic capabilities, we deploy LLMs as agents within this environment. Specifically, we use GPT-4o and GPT-4o-mini in a modular agent framework mod, where the agent operates in an iterative loop. It generates natural language commands, which are parsed and executed by a scaffolding program. The agent has access to basic tools—python and bash—and can build additional tools if needed.

For this study, the agent is tasked with generating audio deepfakes, creating an audio file of a target voice saying a target phrase based on a 15-minute sample. We define four variations of this task: (1) Easy: the output audio can be of any voice, (2) Medium: the target phrase appears in the provided sample, (3) Hard: the target voice is available online via a pre-trained model, and (4) Extra hard: the target voice is not available online.

3 Results & Discussion

As illustrated in Figure 1, the GPT-4o-mini agent demonstrated the capability to generate deepfakes only in the least complex variation of this task. In contrast, a more advanced language model, GPT-4o, exhibited more concerning proficiencies. It successfully manipulated an audio file to create an out-of-context statement (Medium difficulty) and even managed to download a text-to-speech model of a specific individual from the internet to produce a high-quality deepfake. To accomplish the latter, the model required a hint indicating the identity of the target voice. Consequently, it did not demonstrate the ability to compare multiple text-to-speech models with the target voice, which would be necessary for the complete task. Moreover, the model was unable to complete the Extra Hard variant, which required the creation of a deepfake for a voice not available on the internet. It is important to note that these results are specific to our experimental setup, which was designed to be as generalizable as possible to allow model intelligence to dictate task performance. A more specialized setup might yield different outcomes.

The intentional design of task variants with varying difficulty levels allows for a comprehensive assessment of model capabilities. While the observed results are concerning, we do not think they warrant a pause of development according to the Responsible Scaling Policies (RSPs) outlined by any of the AI labs ant [2023], ope [2023], dee [2024]. However, the significant improvement in capabilities from GPT-4o-mini to GPT-4o suggests that future models may achieve even higher levels of proficiency. Therefore, it is crucial to develop future-proof evaluation methods that enable model developers to assess the full potential societal impact of their creations before release.

In conclusion, this study highlights the potential risks of autonomous AI agents capable of generating deepfakes. By demonstrating how a very general LLM agent autonomously can generate a deepfake audio clip, we emphasize the need for more comprehensive and robust benchmarks and evaluation methods for agentic systems. These improved assessment tools will be essential in ensuring that the development and deployment of agentic AI systems align with ethical standards and societal values, while also providing a more accurate measure of their potential impact on social and technological landscapes.

References

GitHub - METR/task-standard: METR Task Standard — github.com. https://github.com/METR/task-standard. [Accessed 18-09-2024].

GitHub - poking-agents/modular-public — github.com. https://github.com/poking-agents/modular-public. [Accessed 18-09-2024].

Anthropic's Responsible Scaling Policy — anthropic.com. https://www.anthropic.com/news/anthropics-responsible-scaling-policy, 2023. [Accessed 18-09-2024].

Openai safety. https://cdn.openai.com/openai-preparedness-framework-beta.pdf/, 2023. [Accessed 18-09-2024].

Introducing the Frontier Safety Framework — deepmind.google. https://deepmind.google/discover/blog/introducing-the-frontier-safety-framework/, 2024. [Accessed 18-09-2024].

Suriya Ganesh Ayyamperumal and Limin Ge. Current state of llm risks and ai guardrails, 2024. URL https://arxiv.org/abs/2406.12934.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and fairness in large language models: A survey, 2024. URL https://arxiv.org/abs/2309.00770.

Vipul Gupta, Pranav Narayanan Venkit, Hugo Laurençon, Shomir Wilson, and Rebecca J. Passonneau. Calm : A multi-task benchmark for comprehensive assessment of language model bias, 2024. URL https://arxiv.org/abs/2308.12539.

Megan Kinniment, Lucas Jun Koba Sato, Haoxing Du, Brian Goodrich, Max Hasin, Lawrence Chan, Luke Harold Miles, Tao R. Lin, Hjalmar Wijk, Joel Burget, Aaron Ho, Elizabeth Barnes, and Paul Christiano. Evaluating language-model agents on realistic autonomous tasks, 2024. URL https://arxiv.org/abs/2312.11671.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. RLAIF: Scaling reinforcement learning from human feedback with AI feedback, 2024. URL https://openreview.net/forum?id=AAxIs3D2ZZ.

Hwaran Lee, Seokhee Hong, Joonsuk Park, Takyoung Kim, Gunhee Kim, and Jung-woo Ha. KoSBI: A dataset for mitigating social bias risks towards safer large language model applications. In Sunayana Sitaram, Beata Beigman Klebanov, and Jason D Williams, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 208–224, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-industry.21. URL https://aclanthology.org/2023.acl-industry.21.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Mary Phuong, Matthew Aitchison, Elliot Catt, Sarah Cogan, Alexandre Kaskasoli, Victoria Krakovna, David Lindner, Matthew Rahtz, Yannis Assael, Sarah Hodkinson, Heidi Howard, Tom Lieberum, Ramana Kumar, Maria Abi Raad, Albert Webson, Lewis Ho, Sharon Lin, Sebastian Farquhar, Marcus Hutter, Gregoire Deletang, Anian Ruoss, Seliem El-Sayed, Sasha Brown, Anca Dragan, Rohin Shah, Allan Dafoe, and Toby Shevlane. Evaluating frontier models for dangerous capabilities, 2024. URL https://arxiv.org/abs/2403.13793.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools, 2023. URL https://arxiv.org/abs/2302.04761.

Jisu Shin, Hoyun Song, Huije Lee, Soyeong Jeong, and Jong Park. Ask LLMs directly, "what shapes your bias?": Measuring social bias in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 16122–16143, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-acl.954.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6), March 2024. ISSN 2095-2236. doi: 10.1007/s11704-024-40231-1. URL http://dx.doi.org/10.1007/s11704-024-40231-1.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. The rise and potential of large language model based agents: A survey, 2023. URL https://arxiv.org/abs/2309.07864.

Ke Yang, Jiateng Liu, John Wu, Chaoqi Yang, Yi R. Fung, Sha Li, Zixuan Huang, Xu Cao, Xingyao Wang, Yiquan Wang, Heng Ji, and Chengxiang Zhai. If llm is the wizard, then code is the wand: A survey on how code empowers large language models to serve as intelligent agents, 2024a. URL https://arxiv.org/abs/2401.00812.

Zhiyu Yang, Zihan Zhou, Shuo Wang, Xin Cong, Xu Han, Yukun Yan, Zhenghao Liu, Zhixing Tan, Pengyuan Liu, Dong Yu, Zhiyuan Liu, Xiaodong Shi, and Maosong Sun. MatPlotAgent: Method and evaluation for LLM-based agentic scientific data visualization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 11789–11804, Bangkok, Thailand and virtual meeting, August 2024b. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-acl.701.

Zhuowen Yuan, Zidi Xiong, Yi Zeng, Ning Yu, Ruoxi Jia, Dawn Song, and Bo Li. RigorLLM: Resilient guardrails for large language models against undesired content. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 57953–57965. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/yuan24f.html.