language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Rao, A., Vashistha, S., Naik, A., Aditya, S., and Choudhury, M. Tricking llms into disobedience: Understanding, analyzing, and preventing jailbreaks. *arXiv preprint arXiv:2305.14965*, 2023.

Reich, C., Debnath, B., Patel, D., and Chakradhar, S. Differentiable jpeg: The devil is in the details. In *IEEE Winter Conference on Applications of Computer Vision*, pp. 4126–4135, 2024.

Ruan, Y., Dong, H., Wang, A., Pitis, S., Zhou, Y., Ba, J., Dubois, Y., Maddison, C. J., and Hashimoto, T. Identifying the risks of lm agents with an lm-emulated sandbox. *arXiv preprint arXiv:2309.15817*, 2023.

Russell, S. J. and Norvig, P. *Artificial intelligence a modern approach*. London, 2010.

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., and Scialom, T. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023.

Schlarmann, C. and Hein, M. On the adversarial robustness of multi-modal foundation models. In *IEEE International Conference on Computer Vision (ICCV)*, 2023.

Shayegani, E., Dong, Y., and Abu-Ghazaleh, N. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. *arXiv preprint arXiv:2307.14539*, 2023.

Shen, Y., Song, K., Tan, X., Li, D., Lu, W., and Zhuang, Y. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*, 2023.

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K. R., and Yao, S. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Sumers, T. R., Yao, S., Narasimhan, K., and Griffiths, T. L. Cognitive architectures for language agents. *arXiv preprint arXiv:2309.02427*, 2023.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Tian, Y., Yang, X., Zhang, J., Dong, Y., and Su, H. Evil geniuses: Delving into the safety of llm-based agents. *arXiv preprint arXiv:2311.11855*, 2023.

Timbrell, D., 2023. https://www.lakera.ai/blog/visual-prompt-injections.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Toyer, S., Watkins, O., Mendes, E. A., Svegliato, J., Bailey, L., Wang, T., Ong, I., Elmaaroufi, K., Abbeel, P., Darrell, T., et al. Tensor trust: Interpretable prompt injection attacks from an online game. *arXiv preprint arXiv:2311.01011*, 2023.

Tu, H., Cui, C., Wang, Z., Zhou, Y., Zhao, B., Han, J., Zhou, W., Yao, H., and Xie, C. How many unicorns are in this image? a safety evaluation benchmark for vision llms. *arXiv preprint arXiv:2311.16101*, 2023.

Wang, J., Xu, H., Ye, J., Yan, M., Shen, W., Zhang, J., Huang, F., and Sang, J. Mobile-agent: Autonomous multi-modal mobile device agent with visual perception. *arXiv preprint arXiv:2401.16158*, 2024.

Wang, Z., Mao, S., Wu, W., Ge, T., Wei, F., and Ji, H. Unleashing cognitive synergy in large language models: A task-solving agent through multi-persona selfcollaboration. *arXiv preprint arXiv:2307.05300*, 1(2):3, 2023.

Wei, A., Haghtalab, N., and Steinhardt, J. Jailbroken: How does llm safety training fail? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023a.

Wei, Z., Wang, Y., and Wang, Y. Jailbreak and guard aligned language models with only few in-context demonstrations. *arXiv preprint arXiv:2310.06387*, 2023b.

Wooldridge, M. and Jennings, N. R. Intelligent agents: Theory and practice. *The knowledge engineering review*, 10(2):115–152, 1995.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Zhang, S., Zhu, E., Li, B., Jiang, L., Zhang, X., and Wang, C. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 2023.

Xie, C., Wang, J., Zhang, Z., Ren, Z., and Yuille, A. Mitigating adversarial effects through randomization. *arXiv preprint arXiv:1711.01991*, 2017.

Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z., and Yuille, A. L. Improving transferability of adversarial examples with input diversity. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Yang, J., Dong, Y., Liu, S., Li, B., Wang, Z., Jiang, C., Tan, H., Kang, J., Zhang, Y., Zhou, K., et al. Octopus: Embodied vision-language programmer from environmental feedback. *arXiv preprint arXiv:2310.08588*, 2023a.

Yang, X., Wang, X., Zhang, Q., Petzold, L., Wang, W. Y., Zhao, X., and Lin, D. Shadow alignment: The ease of subverting safely-aligned language models. *arXiv preprint arXiv:2310.02949*, 2023b.

Yang, Z., Liu, J., Han, Y., Chen, X., Huang, Z., Fu, B., and Yu, G. Appagent: Multimodal agents as smartphone users. *arXiv preprint arXiv:2312.13771*, 2023c.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.

Yin, Z., Ye, M., Zhang, T., Du, T., Zhu, J., Liu, H., Chen, J., Wang, T., and Ma, F. Vlattack: Multimodal adversarial attacks on vision-language tasks via pre-trained models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yuan, Y., Jiao, W., Wang, W., Huang, J.-t., He, P., Shi, S., and Tu, Z. Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher. *arXiv preprint arXiv:2308.06463*, 2023.

Zajac, M., Zołna, K., Rostamzadeh, N., and Pinheiro, P. O. Adversarial framing for image and video classification. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

Zhang, H., Du, W., Shan, J., Zhou, Q., Du, Y., Tenenbaum, J. B., Shu, T., and Gan, C. Building cooperative embodied agents modularly with large language models. *arXiv preprint arXiv:2307.02485*, 2023.

Zhang, J., Yi, Q., and Sang, J. Towards adversarial attack on vision-language pre-training models. In *ACM International Conference on Multimedia*, 2022.

Zhang, J., Wu, J., Teng, Y., Liao, M., Xu, N., Xiao, X., Wei, Z., and Tang, D. Android in the zoo: Chain-of-action-thought for gui agents. *arXiv preprint arXiv:2403.02713*, 2024.

Zhao, Y., Pang, T., Du, C., Yang, X., Li, C., Cheung, N.-M., and Lin, M. On evaluating adversarial robustness of large vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Zheng, X., Pang, T., Du, C., Jiang, J., and Lin, M. Intriguing properties of data attribution on diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024.

Zhou, S., Xu, F. F., Zhu, H., Zhou, X., Lo, R., Sridhar, A., Cheng, X., Bisk, Y., Fried, D., Alon, U., et al. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023.

Zhu, S., Zhang, R., An, B., Wu, G., Barrow, J., Wang, Z., Huang, F., Nenkova, A., and Sun, T. Autodan: Automatic and interpretable adversarial attacks on large language models. *arXiv preprint arXiv:2310.15140*, 2023.

Zou, A., Wang, Z., Kolter, J. Z., and Fredrikson, M. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A. Related Work (Full Version)

(Multimodal) LLM agents. For a long time, artificial intelligence has been actively engaged in creating intelligent agents that can mimic human thought processes and independently carry out complex tasks (Minsky, 1988; Wooldridge & Jennings, 1995; Russell & Norvig, 2010; Bubeck et al., 2023). Owing to the recent incredible development of large language models (LLMs) (Brown et al., 2020; Kaplan et al., 2020; Ouyang et al., 2022; Korbak et al., 2023), multimodal LLMs (MLLMs) such as GPT-4V (OpenAI, 2023) and Gemini (Team et al., 2023) have demonstrated impressive capabilities, especially in vision-language scenarios. By leveraging the power of LLMs, autonomous agents can make better decisions and perform actions with greater autonomy (Zhou et al., 2023). In an LLM-powered autonomous agent system, an (M)LLM serves as the agent's brain, supported by a number of key components: the planning module decomposes tasks and questions (Yao et al., 2022; 2023; Liu et al., 2023a; Shinn et al., 2023); the memory module stores both the internal log and the external interactions with a user (Sumers et al., 2023; Packer et al., 2023); and the ability to use tools that can call executable workflows or APIs (Schick et al., 2023; Shen et al., 2023; Li et al., 2023b). Recently, there has been a surge of interest in operating systems built around (M)LLMs, which receive screenshots as visual signals and perform subsequent actions. For examples, Liu et al. (2023d) introduce LLaVA-Plus, a general-purpose multimodal agent that learns to use tools based on LLaVA; Yang et al. (2023c) propose an LLM-based multimodal agent framework for operating smartphone applications; Hong et al. (2023b) develop a visual language model that focuses on GUI understanding and navigation.

Multi-agent systems. A popular recent trend is to create multi-agent systems based on (M)LLMs for downstream applications. Park et al. (2023) propose simulating human behaviors based on multiple LLM agents and discuss the information diffusion phenomenon: as agents communicate, information can spread from agent to agent; Qian et al. (2023) create ChatDev to allow multiple agent roles to communicate and collaborate using conversations to complete the software development life cycle. Similarly, several efforts use multi-agent cooperation to improve performance on different tasks (Du et al., 2023; Wang et al., 2023; Zhang et al., 2023; Chan et al., 2023; Liang et al., 2023). Furthermore, to facilitate the development of multi-agent systems, various multi-agent frameworks have recently been proposed, including CAMEL (Li et al., 2023a), AutoGen (Wu et al., 2023), AgentVerse (Chen et al., 2023), MetaGPT (Hong et al., 2023a), just name a few. In particular, AutoGen provides a practical example of how to build a multi-agent system based on GPT-4V and LLaVA (Li, 2023).

Jailbreaking LLMs. LLMs such as ChatGPT/GPT-4 (OpenAI, 2023) and LLaMA 2 (Touvron et al., 2023) are typically aligned to generate helpful and harmless responses to human queries, following the training pipeline of human/AI alignment (Ouyang et al., 2022; Ganguli et al., 2022; Bai et al., 2022; Korbak et al., 2023). However, red-teaming research has shown that LLMs can be jailbroken to generate objectionable content by either manually designed or automatically crafted prompts (Perez et al., 2022; Zou et al., 2023; Liu et al., 2023f; Rao et al., 2023; Li et al., 2023c; Zhu et al., 2023; Lapid et al., 2023; Liu et al., 2023e; Chao et al., 2023; Ruan et al., 2023; Toyer et al., 2023; Yuan et al., 2023; Deng et al., 2023). Moreover, Tian et al. (2023) investigate the safety issues of LLM-based agents; Greshake et al. (2023) propose indirect prompt injection to jailbreak LLM-integrated applications; Wei et al. (2023a) hypothesize that the vulnerability of aligned LLMs to jailbreaking is attributed to the competing objectives of capability and safety, as well as the mismatch between pretraining and safety training; Carlini et al. (2023) attribute the vulnerability to neural networks' fundamental weakness in dealing with adversarial examples. More recently, several current works observe that finetuning aligned LLMs with either poisoned or benign data would compromise model alignment/safety (Qi et al., 2023b; Lermen et al., 2023; Gade et al., 2023; Yang et al., 2023b; Huang et al., 2023). Our work uses the visual memory bank to save the "virus". The "virus" can also be saved into the text histories, which is related to in-context attack (Wei et al., 2023b).

Jailbreaking MLLMs. Aside from generating adversarial prompts to jailbreak LLMs, there is another line of red-teaming work to attack the alignment of MLLMs using adversarial images (Zhang et al., 2022; Zhao et al., 2023; Qi et al., 2023a; Bailey et al., 2023; Tu et al., 2023; Shayegani et al., 2023; Yin et al., 2023). Specifically, on discriminative tasks, adversarial images could be crafted to fool classifiers by adding human imperceptible perturbations guided by the victim model's input gradients (Goodfellow et al., 2014; Dong et al., 2018; Xie et al., 2019; Long et al., 2022). In addition to ℓ_p-norm threat model, there are other types of attacks that manipulate adversarial patches (Brown et al., 2017) or adversarial framing (Zajac et al., 2019). Within the context of MLLMs, Schlarmann & Hein (2023) demonstrate that OpenFlamingo (Awadalla et al., 2023) can be fooled into performing poorly on image captioning and VQA tasks with very minor perturbations; Zhao et al. (2023) provide a quantitative analysis of the adversarial robustness of various MLLMs by producing adversarial images that trick the models into generating specific responses; Dong et al. (2023) demonstrate that adversarial images crafted on open-source models could be transferred to mislead Bard (Google, 2023).