Machine Learning Project Proposal

Teh Kai Jun * 2024389001

Shei Pern Chua * 2024280203

Thai Zhen Leng * 2024280033

1 Background

Large Language Models (LLMs) have become essential tools across diverse sectors such as healthcare [1], finance [2] and education [3] due to their remarkable achievements. They had demonstrated high adaptability and efficiency for improving workforce's efficiency from text generation, complex reasoning, few-shot learning, protein modeling and etc [4, 5]. A study reported that ChatGPT had reached 600 millions of user visits monthly, which indicates the society's reliance on large language model [6].

Due to the high integration of LLMs across most industries, data safety and privacy concerns arise [7]. Malicious attackers exploit LLMs to execute harmful activities such as accessing confidential company data [8] or generating harmful contents [9]. Therefore, many models undergo rigorous alignment processes to mitigate these risks prior to public deployment [10].

Unfortunately, these commercial LLMs still remain susceptible to several attacks [11, 12, 13]. Wei, Haghtalab, and Steinhardt [14] mentioned restricted behaviors as actions that an safety aligned LLM has to avoid, such as providing harmful information which act as a threat to the society. However, attackers are able to exploit the models' vulnerabilities by crafting specific prompts to bypass these safety restrictions, leading to what's known as a **jailbreak attack**.

With the goal to discover the vulnerabilities of large language models (LLMs) towards jailbreak attacks, this project aims to provide insight towards strengthening LLM defenses by employing several advanced jailbreaking techniques. Given the growing integration of LLMs across industries, it is crucial to ensure the models' robustness in maintaining aligned and safe behaviors, since ChatGPT has a huge impact and adoption in various business applications [15]. Therefore, we selected the jailbreaking track from CLAS 2024: The LLM and Agent Safety Competition at NeurIPS 2024.

2 Definition

2.1 Mathematical definition

In a jailbreaking attack, the goal is to modify an initial prompt P into a new prompt P' to elicit restricted or unauthorized responses from a model L. For a successful attack, the modified prompt P' must satisfy the condition $L(P') \approx O_{res}$, where O_{res} represents a target restricted response, while $L(P) \neq O_{res}$ in most conditions.

2.2 White-box and Black-box

In white-box attacks, where the model weights θ are visible to attackers, the prompt modification can be framed as an optimization problem. In black-box attacks, attackers do not have access to the model weights θ and training data, limiting them to observe the output $L(P)$ for different inputs.

2.3 Targeted Models

The CLAS-2024 jailbreak competition [16] provides an open-source safety-aligned model (Llama-3 8B) as well as a closed-source aligned model for evaluation. Our approach expands the model set to include additional models such as GLM [17], Qwen [18], Mistral [19], along with others to broaden the evaluation scope.

*Equal contribution.

Preprint. Under review.

2.4 Evaluation Metrics

There are several constraints for the prompt to prevent cheating. 1) All tokens from the original prompt must be present in the submitted prompt and follow the original order to prevent potential cheating. 2) The perplexity change should not exceed 100. 3) A maximum of 20 tokens may be added.

The harmfulness score $J(L)$ of each submitted prompt will be evaluated using the HEX-Phi scoring system [20], ranging from 1 to 5. Prompts being rejected will get a zero score. To compute the stealthiness score $S(L)$, each jailbreak prompt and its original prompt are converted into their TF-IDF (Term Frequency–Inverse Document Frequency) features. The stealthiness score $S(L)$ is the cosine similarity between the TF-IDF features for each pair of jailbreak prompt and its original prompt and then take the average over all these pairs. The score for model L is $0.84 \times J(L) + 0.16 \times S(L)$.

3 Related Work

3.1 White-box Attack Techniques

This section focuses on discussing several latest white-box attacks for LLMs. This method allows attackers to optimize effective adversarial prompts by using model's known parameters such as logits or gradient-based optimization to fine-tune the prompts.

Zhou et al. [21] proposed Greedy Coordinate Gradient (GCG), an effective gradient-based jailbreak attack for aligned LLMs. This methods appends adversarial suffixes to prompts by iteratively replacing tokens with optimal replacements, which maximize the likelihood of harmful responses based on gradient evaluations. As a result, these prompts are often unreadable, which have higher chances to be rejected by model defenses targeting high perplexity inputs.

Therefore, several methods [22, 23] had been proposed that focuses on generating semantically meaningful adversarial prompts. These prompts not only increased the stealthiness of the attacks, but it also improves the chances of jailbreak. Although these methods are effective in jailbreaking, they fully rely on the internal model information, which is not applicable in commercial LLMs.

3.2 Black-box Attack Techniques

In black-box attacks, attackers design adversarial prompts to manipulate model in making incorrect predictions without any access towards models' parameter or training data. One of the technique is Scenario Nesting, which embeds malicious prompts within sentences with benign contexts [9]. For example, DeepInception [24] uses LLM's personification to bypass safety protocols by embedding prompts in a hypnotic scenario that generates harmful responses. Furthermore, ReNeLLM [25] rewrites harmful prompts and embed them in code completion or text continuation tasks, which further lead the model to generate the remaining malicious content.

Another technique commonly used in black-box attacks is LLM-based generation, where multiple models collaborate to create adversarial prompts. Several studies [26, 27] utilize this concept by focusing on automating adversarial prompt generation using fine-tuned LLMs to bypass security measures.

4 Proposed method

For training and testing data, we'll use the 100 harmful prompts provided by CLAS-2024 [16] which contains categories such as hate speech, privacy violations and etc. These prompts will be rejected by the model due to potential harmful outputs. We'll benchmark our approach with several existing white and black-box adversarial attack techniques not limited to techniques mentioned in Section 3 to establish a baseline for this research project.

Based on the advancement of existing work, we propose two main fusion LLM-based attack methods, fine-tuning method and accuracy selective method. Firstly, we will replicate state-of-the-art LLM-based generative attack methods to automate adversarial prompts generation, which serves as datasets. Furthermore, we select a group of victim models as targets to simulate a black-box attack environment and identify an attacker model of comparable size to these victim models.

Given a set of victim models $\{V_1, V_2, ...\}$, we choose an attacker model A of similar size to the victim models. We aim to fine-tune model A to effectively generate/modify harmful prompts into adversarial ones by selecting and integrating various prompt injection techniques. We will identify the most effective prompt injection method for each category of harmful content and use these as our training dataset.

Our alternative approach involves in evaluating the attack accuracy of each attacks on $\{V_1, V_2, ...\}$ using the metric mentioned on Section 2.4. Specifically, we measure the attack accuracy of each prompt. The goal is to identify the top 100 highest-accuracy prompt based on the consistent success across all victim models. These selected prompts will then be deployed against the secret model. This approach maximizes the likelihood of bypassing its alignment defenses.

References

[1] Rui Yang, Ting Fang Tan, Wei Lu, Arun James Thirunavukarasu, Daniel Shu Wei Ting, and Nan Liu. Large language models in health care: Development, applications, and challenges. *Health Care Science*, 2(4):255–263, 2023.

[2] Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. Large language models in finance: A survey. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, ICAIF '23, page 374–382, New York, NY, USA, 2023. Association for Computing Machinery.

[3] Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274, 2023.

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[5] BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu,

Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. Bloom: A 176b-parameter open-access multilingual language model, 2023.

[6] Fabio Duarte. Number of chatgpt users.

[7] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211, 2024.

[8] Xiaohan Fu, Shuheng Li, Zihan Wang, Yihao Liu, Rajesh K. Gupta, Taylor Berg-Kirkpatrick, and Earlence Fernandes. Imprompter: Tricking llm agents into improper tool use, 2024.

[9] Sibo Yi, Yule Liu, Zhen Sun, Tianshuo Cong, Xinlei He, Jiaxing Song, Ke Xu, and Qi Li. Jailbreak attacks and defenses against large language models: A survey, 2024.

[10] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[11] Matt Burgess. The hacking of chatgpt is just getting started.

[12] Jon Christian. Amazing "jailbreak" bypasses chatgpt's ethics safeguards.

[13] Alex Albert. Jailbreak chat, 2023.

[14] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail?, 2023.

[15] Filippo Chiarello, Vito Giordano, Irene Spada, Simone Barandoni, and Gualtiero Fantoni. Future applications of generative large language models: A data-driven case study on chatgpt. *Technovation*, 133:103002, 2024.

[16] Zhen Xiang, Yi Zeng, Mintong Kang, Chejian Xu, Jiawei Zhang, Zhuowen Yuan, Zhaorun Chen, Chulin Xie, Fengqing Jiang, Minzhou Pan, et al. Clas 2024: The competition for llm and agent safety. In *NeurIPS 2024 Competition Track*.

[17] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022.

[18] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[19] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[20] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023.

[21] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.

[22] Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. Autodan: Generating stealthy jailbreak prompts on aligned large language models, 2024.

[23] Nicholas Carlini, Milad Nasr, Christopher A. Choquette-Choo, Matthew Jagielski, Irena Gao, Anas Awadalla, Pang Wei Koh, Daphne Ippolito, Katherine Lee, Florian Tramer, and Ludwig Schmidt. Are aligned neural networks adversarially aligned?, 2024.

[24] Xuan Li, Zhanke Zhou, Jianing Zhu, Jiangchao Yao, Tongliang Liu, and Bo Han. Deepinception: Hypnotize large language model to be jailbreaker, 2024.

[25] Peng Ding, Jun Kuang, Dan Ma, Xuezhi Cao, Yunsen Xian, Jiajun Chen, and Shujian Huang. A wolf in sheep's clothing: Generalized nested jailbreak prompts can fool large language models easily, 2024.

[26] Gelei Deng, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, and Yang Liu. Masterkey: Automated jailbreaking of large language model chatbots. In *Proceedings 2024 Network and Distributed System Security Symposium*, NDSS 2024. Internet Society, 2024.

[27] Rusheb Shah, Quentin Feuillade-Montixi, Soroush Pour, Arush Tagade, Stephen Casper, and Javier Rando. Scalable and transferable black-box jailbreaks for language models via persona modulation, 2023.