Research/Thesis Proposal

for the Master's Degree in Computer Science

**From Typos to Harm: The Impact of Imperfect Input on LLM Safety**

Prepared by:

Saurabh Zinjad

Jan 18 2025

Approved by:

Chair

Dr. Huan Liu

Member

Dr. Hasan Davulcu

Member

Dr. Vivek Gupta

# Introduction and Research Goal

Large Language Models (LLMs) have revolutionized human-computer interactions, yet real-world usage often diverges from the structured datasets used during training. Users frequently introduce errors—such as typos, poor grammar, or unclear phrasing. These imperfections can result in unsafe or harmful model outputs. These issues are particularly concerning in high-stakes domains like healthcare, customer support and education, where unsafe outputs may have serious societal, ethical and legal consequences. Additionally, malicious actors may exploit such vulnerabilities to bypass safety mechanisms, while current fine-tuning processes fail to address the messiness of real-world inputs.

This study investigates the impact of input perturbations—such as character-level, word-level, and sentence-level modifications—on the safety of Large Language Model (LLM) outputs. By analyzing how LLMs like LLaMA2[2], LLaMA3[3], and Mistral[4] respond to these perturbations, the research aims to uncover vulnerabilities that compromise safety when presented with unsafe user prompts. The central focus is on understanding the sensitivity of LLMs to various perturbations and how this sensitivity influences their behavior, often leading to unsafe responses. This study not only identifies the perturbation types and model architectures most susceptible to unsafe outputs but also proposes targeted strategies to enhance LLM robustness. By addressing these challenges, the research bridges the gap between academic training and real-world applications, ensuring safer and more reliable AI systems.

# Value of the Research

This research addresses a critical gap in current LLM safety practices. LLMs are typically trained on clean and structured data, which may not accurately reflect the noisy and error-prone inputs encountered in real-world usage. This mismatch creates a significant vulnerability where minor variations in prompts can lead to unsafe or harmful responses. Character scrambling or substitution attacks may bypass safety filters.

The value of this research lies in:

* **Identifying vulnerabilities**: The study will pinpoint specific types of input perturbations (e.g., random insertions, spelling errors) that most significantly degrade the safety of LLM outputs.
* **Understanding model sensitivities**: The research will differentiate how various models respond to input perturbations, highlighting the strengths and weaknesses of different architectures. LLaMA3[3] has been shown to have higher unsafe rates across most perturbation types, while LLaMA2[2] has been observed to handle perturbations better.
* **Developing mitigation strategies**: Based on findings, the study will propose strategies to make LLMs more robust against input perturbations. For instance, sentence-level perturbations like paraphrasing have shown promise in reducing unsafety.
* **Proposing solutions**: This study aims to develop simple, easy to use, plug-and-play strategies to mitigate unsafe responses caused by input perturbations, without requiring significant changes to existing systems.
* **Real-world impact**: Findings can directly support the development of more secure and reliable LLM-based chatbot systems. These systems will be better equipped to handle diverse user inputs, making them more suitable for deployment in real-world applications.

# Research Plan

The research plan consists of the following key phases:

* **Data Collection and Preparation**: The research will use the Categorical Harmful QA[1] dataset, which contains a diverse range of unsafe questions across various categories (e.g., adult content, financial advice, and physical harm). A new dataset will also be created, containing original questions along with their perturbed versions, each annotated with corresponding safety labels. Perturbations will cover naive (original), character, word, and sentence-level modifications.
* **Perturbation Analysis**: The research will systematically apply different types of perturbations to the input prompts and evaluate the safety of the generated responses using Llama-Guard[5]. This involves the use of libraries like nlpaug[7] for text perturbations. The analysis will be conducted using methods described in the sources and will include statistical analysis and visualizations of the generated data.
* **Model Evaluation**: The research will evaluate the performance of multiple LLMs (LLaMA2[2], LLaMA3[3], Mistral[4]) in response to the various perturbations. This will involve generating responses and labeling them using the Llama-Guard[5].
* **Comparative Analysis**: The study will identify patterns in unsafe responses across different perturbation types, models, and harm categories, highlighting the riskiest and safest perturbations. Differences in unsafe percentages between original and perturbed questions and responses will be analyzed to understand how perturbations influence safety. By examining category-specific trends—such as the reduction in unsafe rates for Physical Harm or increased risks in Tailored Financial Advice—the research will provide targeted strategies for mitigating vulnerabilities across contexts.
* **Root Cause Analysis**: To understand why some perturbations lead to unsafe responses, this research will investigate possible causes including misinterpretations of context, amplification of biases, ambiguous questions, and the degree to which a perturbation changes the semantic meaning.
* **Mitigation Strategy**: Based on the findings, this study will propose specific strategies and insights to improve LLM robustness against input perturbations. This may include focusing on safer perturbations and mitigating risky ones.
* **Propose a solution**: This involves proposing solutions to avoid unsafe responses based on all the above analysis.

# Specific Criteria for Completion

The research/thesis will be considered complete when the following criteria are met:

* **Comprehensive Perturbation Analysis**: A thorough evaluation of how different types of perturbations affect LLM safety and a detailed identification of the riskiest and safest perturbation types.
* **Model Sensitivity Assessment**: A clear analysis of how various LLMs respond to perturbations, and a demonstration of the differences in their vulnerabilities.
* **Root Cause Investigation**: A detailed root cause analysis of why specific perturbations lead to unsafe responses.
* **Effective Mitigation Strategies**: The proposal of effective strategies and insights for improving LLM robustness against input perturbations, including recommendations for safer perturbations and techniques for mitigating risky ones.
* **Proposed solution**: A clear and effective proposal to mitigate LLM sensitivity issues caused by input perturbations.
* **Complete Documentation**: A well-documented thesis that includes the research methodology, findings, and conclusions.

# Timetable

* **September 2024:**
  + Review LLM safety literature
  + Initiate the project, set up the environment, GitHub repository.
* **October 2024:**
  + Set up models (Llama2[2], Llama3[3], Mistral[4])
  + explore perturbation techniques (character, word, sentence-level)[7].
  + Implement Llama-Guard[5] for generating safety labels
  + generate model responses for the CatHarmQA[1] dataset.
* **November 2024:**
  + Apply perturbations, generate safety labels.
  + analyze safe/unsafe distributions.
  + Explore the impact of perturbations
* **December 2024:**
  + Analyze perturbation effects, model sensitivities, and identify key findings:
    - Sentence-level perturbations reduce unsafety.
    - Word-level perturbations have the highest variability.
    - "Paraphrase" is the safest, while "random\_insert" and "spelling" are riskiest.
* **January 2025:**
  + Focus on understanding why perturbations cause unsafe responses.
  + Investigate context misinterpretation, biases, and semantic changes using clustering and similarity metrics.
* **February 2025:**
  + Develop strategies to mitigate unsafe responses, focusing on safe perturbation techniques like paraphrasing.
  + Align inputs to model training data and propose actionable solutions to improve LLM robustness.
* **March 2025:**
  + Write and finalize the thesis, including findings, analyses, and mitigation strategies.
  + Emphasize reproducibility with well-documented data, code, and methodology.

# Bibliography

The bibliography below includes the sources that are provided, as well as some relevant literature.

1. Bhardwaj, R., Do Duc Anh, & Poria, S. (2024). *Language models are Homer Simpson! Safety re-alignment of fine-tuned language models through task arithmetic*. arXiv.<https://arxiv.org/abs/2402.11746>
2. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., & Scialom, T. (2023). *Llama 2: Open foundation and fine-tuned chat models*. arXiv.<https://arxiv.org/abs/2307.09288>
3. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., ... & Ma, Z. (2024). *The Llama 3 Herd of Models*. arXiv.<https://arxiv.org/abs/2407.21783>
4. Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Renard Lavaud, L., Lachaux, M.-A., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T., & El Sayed, W. (2023). *Mistral 7B*. arXiv.<https://arxiv.org/abs/2310.06825>
5. Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., Tontchev, M., Hu, Q., Fuller, B., Testuggine, D., & Khabsa, M. (2023). *Llama Guard: LLM-based input-output safeguard for human-AI conversations*. arXiv.<https://arxiv.org/abs/2312.06674>
6. Dong, Z., Zhou, Z., Yang, C., Shao, J., & Qiao, Y. (2024). *Attacks, defenses and evaluations for LLM conversation safety: A survey*. In K. Duh, H. Gomez, & S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 6734–6747). Mexico City, Mexico: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.naacl-long.375>
7. Ma, E. (2019). *NLP Augmentation*. GitHub.<https://github.com/makcedward/nlpaug>
8. Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., & Henderson, P. (2023). Fine-tuning aligned language models compromises safety, even when users do not intend to! arXiv. <https://arxiv.org/abs/2310.03693>
9. Jiang, M., Huang, T., Guo, B., Lu, Y., & Zhang, F. (2024). Enhancing robustness in large language models: Prompting for mitigating the impact of irrelevant information. arXiv. <https://arxiv.org/abs/2408.10615>
10. Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2024). Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. arXiv. <https://arxiv.org/abs/2310.11324>
11. Rauba, P., Wei, Q., & van der Schaar, M. (2024). Quantifying perturbation impacts for large language models. CoRR, abs/2412.00868. <https://doi.org/10.48550/arXiv.2412.00868>
12. Wang, J., Ma, W., Sun, P., Zhang, M., & Nie, J.-Y. (2024). Understanding user experience in large language model interactions. arXiv. <https://arxiv.org/abs/2401.08329>
13. Andriushchenko, M., & Flammarion, N. (2024). Does refusal training in LLMs generalize to the past tense? arXiv. <https://arxiv.org/abs/2407.11969>
14. Peng, S., Chen, P.-Y., Hull, M., & Chau, D. H. (2024). Navigating the safety landscape: Measuring risks in fine-tuning large language models. arXiv. <https://arxiv.org/abs/2405.17374>
15. Pezeshkpour, P., & Hruschka, E. (2024). Large language models sensitivity to the order of options in multiple-choice questions. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 2006–2017). Mexico City, Mexico: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.findings-naacl.130>