Early-Stopping LLM Generation with Token-wise Advance Toxicity Prediction

GenAI & Modern DL (EECS6694): Project Proposal

V Sai Vignesh, Dhruv Srikanth, Likhith Ayinala

October 2024

1 Problem Statement

Our problem statement can be summarized as follows: Is it possible to predict the trajectory of an LLM's response based on features extracted from the initially generated tokens?
We will conduct a thorough analysis to determine the feasibility of this prediction, explore how early-on in generation we can accurately predict the output direction and identify optimal policies for implementation. Additionally, we will explore potential applications of this approach.

2 Proposed Work & Novelties

LLMs are known to plan ahead for future tokens, as shown in [3]. Our goal is to determine whether meaningful information can be extracted during the generation of the initial tokens to predict the direction the model's output will take. This would enhance our understanding of LLM behavior and enable practical strategies, such as early stopping of toxic responses or replacing generic "I cannot answer this question" outputs with template responses. This could ultimately help us reduce inference-time computation.

3 Data Sources

We will use the BeaverTails prompts from [1] on our chosen model to create a model-specific dataset. This dataset will contain features extracted from the model's initial tokens along with the corresponding response type: safe, unsafe, or refusal. We will use this dataset to train our proposed network and evaluate its performance on a held-out section of the same dataset.
To streamline the process, we can automate classification by feeding the chosen model's outputs into a state-of-the-art reference model. This approach will allow us to generate a large dataset with minimal manual effort.

4 Expected Methodology

1. Identify a dataset that triggers unsafe responses in LLMs, and annotate the responses as safe, unsafe, or refusal. This dataset will be used to train our proposed pipeline. We have identified the BeaverTails dataset [1] for this purpose. It contains over 300000 QA pairs across various categories such as animal safety, discrimination, and more, which could potentially elicit unsafe responses.

2. Use a model capable of generating unsafe outputs with minimal safeguards to generate more unsafe prompts and leverage existing safety datasets. From our preliminary analysis, state-of-the-art LLMs have numerous safeguards preventing toxic responses. Additionally, many are fine-tuned on the BeaverTails dataset, which limits the generation of unsafe outputs. Open-source models without safeguards may lack refusal responses ("I cannot answer this question"), reducing their utility. To address these challenges, we plan to use older LLMs or minimally safeguarded open-source models. As a fallback, we will explore purposeful degradation of LLMs using techniques such as those described in [2].

3. Select a neural network architecture and train it on features extracted from the first few tokens. A classifier will assess whether the response aligns with expectations. Suitable models for this task include Qwen 2.5 or other smaller-scale LLMs with lower inference costs.

4. After obtaining concrete results, we will experiment with and optimize the model. This includes determining the minimum number of tokens required for accurate estimation and the optimal frequency for running the neural network.

5. Develop a pipeline to apply the model across different use cases. A key application would be preventing the model from generating undesired responses that deviate from expected instructions.

5 Outcomes

Our expected outcomes are as follows:

1. Determine meta-information about the direction of LLM responses from features obtained in the initial passes.

2. Utilize this information to save compute in the 'toxic' case.

Additionally, this approach could be extended to other tasks, such as predicting the likelihood of an LLM successfully following instructions.

References

[1] Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Chi Zhang, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset, 2023.

[2] Domenic Rosati, Giles Edkins, Harsh Raj, David Atanasov, Subhabrata Majumdar, Janarthanan Rajendran, Frank Rudzicz, and Hassan Sajjad. Defending against reverse preference attacks is difficult, 2024.

[3] Wilson Wu, John Xavier Morris, and Lionel Levine. Do language models plan ahead for future tokens? In *First Conference on Language Modeling*, 2024.

[4] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey, 2023.