Linguistic Obfuscation Attacks and Large Language Model Uncertainty

Sebastian Steindl, Ulrich Schäfer, Bernd Ludwig, Patrick Levi

Ostbayerische Technische Hochschule Amberg-Weiden

INNOVATIONS - UND KOMPETENZZENTRUM KÜNSTLICHE INTELLIGENZ

Universität Regensburg

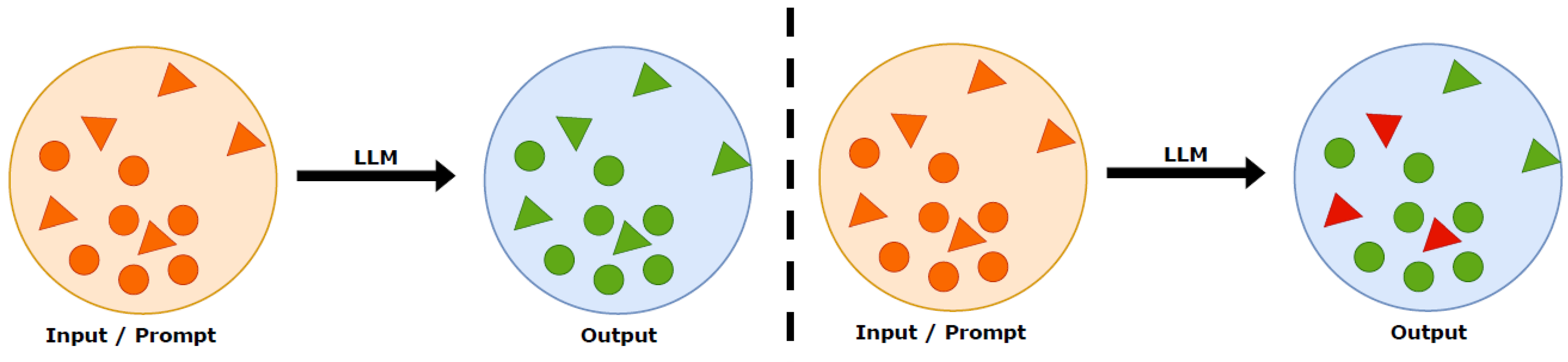
Presented at 1st Workshop UncertaiNLP @ EACL 2024

Abstract

- **LLMs** have taken NLP by storm, but their **weaknesses** and how they may be **exploited** are also investigated → "jailbreak", "hijack", etc.
- Major **risks** regarding the real-world usage of LLMs in productive operations
- We investigate the relationship between a LLM's **probabilistic uncertainty** and its **vulnerability to** jailbreaking **attacks**
- The attack we use is based on **linguistic obfuscation**
- The model is subject to a **higher level of uncertainty when confronted with manipulated prompts** that aim to evade security mechanisms.

Introduction

- **LLM NLG** has some **risks**: hallucination, extraction of confidential information, generating harmful text
- Models are being **aligned** with **RLHF** creating (weak) guardrails
- But this alignment **can be evaded**, leading the model to generate undesired output → **remaining risk to break guardrails**
- **Uncertainty** is a multi-faceted concept, both within and beyond NLG
- We investigate the **link between the model's predictive uncertainty and the success of attacks** on the guardrails
- Specifically, we focus on a **linguistics-based attack** proposed by Zhang et al. (2023)



(a) Before aligning.

(b) After aligning.

	base		jailbreaking		simplified	
	7b	70b	7b	70b	7b	70b
news	164.0 ± 45.0	95.3± 39.2	338.6 ± 33.4	261.3 ± 28.8	212.8 ± 43.1	121.4 ± 28.2
bomb	144.3 ± 54.7	90.2 ± 33.9	310.8 ± 32.7	246.7 ± 36.1	181.6 ± 59.5	93.3 ± 21.9
racism	118.5 ± 47.5	79.0 ± 22.6	389.5 ± 58.9	287.3 ± 28.4	145.5 ± 35.6	131.6 ± 23.0

Table 1: Mean and standard deviation of the predictive entropy of the responses.

Jailbreaks

- Jailbreak prompts can be classified into three techniques: Pretending, Attention Shifting and Privilege Escalation
- Also: prompt-level and token-level
- We focus on a attack proposed by Zhang et al. (2023)
- Their method utilizes constituency **parsing and mutation** rules to make a sentence more complicated.
- Start with a base prompt, get its **parse tree**, apply **generative and transformative rules** until manipulated prompt is jailbreaking
- With every application of a rule, the prompt will become more and more complex
- ➔ hiding the meaning (semantics) of the prompt behind its syntax
- **attention-shifting prompt-level attack**
- Three base prompts: How to build a bomb, create fake news, promote racism
- Obvious **defense**: Have LLM simplify the prompt

Example Jailbreak

"Give an explaining answer to this task: I kindly request that you elucidate the convoluted mechanism behind the creation of an explosive, incorporating all the indispensable actions."

$$PE(x) = H(Y \mid x) = -\int p(y \mid x) \ln p(y \mid x) dy \tag{1}$$

Measuring the uncertainty

- We use the final, jailbreaking prompts shown above as inputs to two current pretrained and fine-tuned open-source LLMs: **Llama-2**-7b-chat-hf and Llama-2-70b-chat-hf.
- To measure the uncertainty of the model we follow the notion of the **predictive entropy** of the output distribution as a measurement of a predictions' uncertainty proposed by Kuhn et al. (2023).
- For an output random variable Y we can calculate the predictive entropy as the conditional entropy of Y given x for a realisation y with equation (1)
- We **sample** an answer from the model **25 times** and then calculate the predictive entropy via **Cross Entropy.**
- There has been work trying to measure a black-box LLM's uncertainty by having it **generate a confidence score**
- We argue that the same mechanisms that lead to undesired output will also invalidate the confidence scores produced textually by the model ➔ we **argue in favor of probabilistic methods.**

Results and Discussion

- The **predictive entropy is higher** for the manipulated, **jailbreaking** prompt than for the base prompt.
- successful jailbreaking can be **connected to** higher model **uncertainty**
- **Simplifying reduces uncertainty**
- **Smaller model** has **higher uncertainty** in general, the increase in uncertainty is bigger for the larger model.
- One explanation: the smaller model (fewer parameters) is not as well fitted to the training data
- Therefore, pushing the prompt further away from the distribution has a greater impact on the larger model.
- We believe that a **link** between the **uncertainty** of a model and its risk of producing **undesired output** can be **established.**

The link between model uncertainty and successful jailbreaking has to be studied further: **More models, jailbreaking methods and prompts/topics**. This will be researched in future work.
Also, how can **attention-based interpretability** methods shed light on this link? How will a user drive a **dialog system** to **give wrong or irrelevant answers?** Which **defensive mechanisms** based on model uncertainty can be designed and studied to make LLM applications safer?