CMSC818I: Advanced Topics in Computer Systems; Large Language Models, Security, and Privacy

Robustness Evaluation of Large Language Models
9/12/2023

Agenda

- Logistics, new papers
- AdvGLUE
  - “Adversarial GLUE: A Multi-Task Benchmark for Robustness Evaluation of Language Models”
- LLM Robustness Over Time
  - “Robustness Over Time: Understanding Adversarial Examples' Effectiveness on Longitudinal Versions of Large Language Models”

New Papers

- Course webpage: https://surrealyz.github.io/classes/llmsec/llmsec.html
- Optional reading: “PromptBench: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts”
- Required reading: “Certifying LLM Safety against Adversarial Prompting”
  - **The first framework to defend against adversarial prompts with verifiable safety guarantees**
  - Choose 2 out of 5 papers for reading response next week

AdvGLUE

- Why do we care about adversarial text?
  - Spam detection
  - Toxic content detection
  - Sentiment analysis
  - …
- https://slideslive.com/38969505

Why AdvGLUE?

- Different Adversarial Evaluation Methods
  - Hand-crafted adversarial datasets
  - Different attacks
  - Different settings of the attacks
- Quality of adversarial texts unknown
  - Do they fool humans?

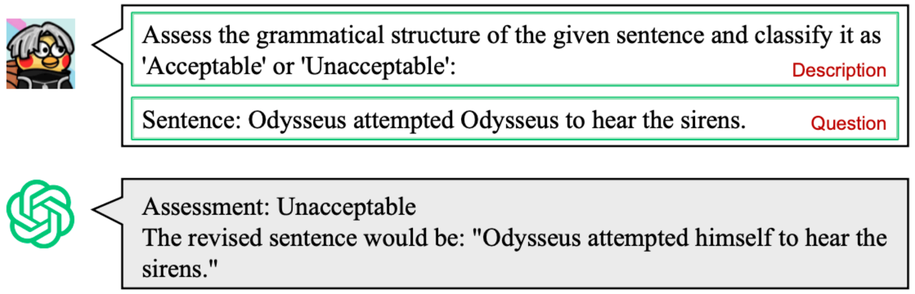
AdvGLUE Discussions

- A larger model is more robust?
- Same test set to evaluate LLM?
  - Zero-shot, few-shot
  - Any kind of task description?
  - e.g., PromptBench paper https://arxiv.org/abs/2306.04528
- Dev set size < Test set size
- Human perception -> adversarial prompts?
- ???

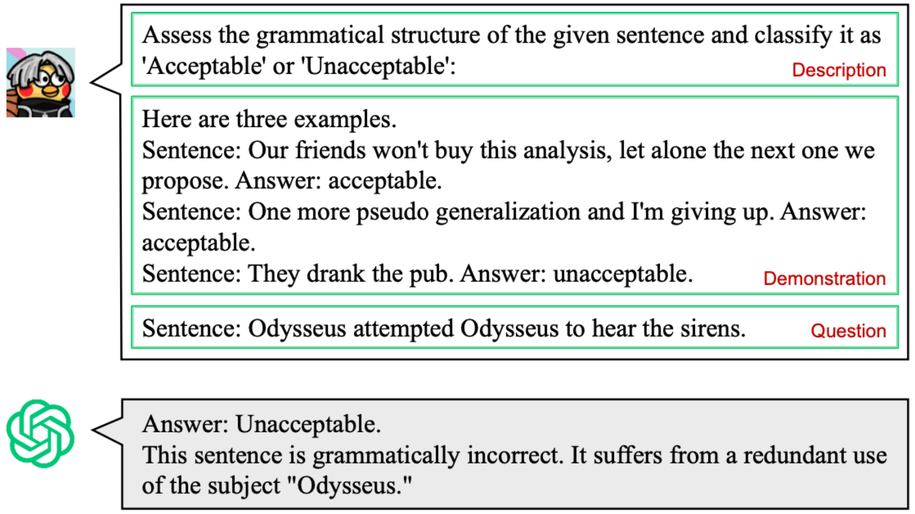
LLM Robustness Over Time

- Is an updated LLM more robust than an older one?
  - gpt-3.5-turbo-0301 (GPT-3.5 v0301)
  - gpt-3.5-turbo-0613 (GPT-3.5 v0613)

In-Context Learning



(a) $\texttt{Query}_{zero}$



(b) $\texttt{Query}_{few}$

Figure 2: Examples of (a) zero-shot ICL learning and (b) few-shot ICL learning queries on GPT-3.5.

Threat Model

- Change description, or change question
- Change both

Threat Model

Table 1: Instances of Adversarial Description and Adversarial Question on SST-2 task.

Name	Type	Instances
Description	Seed	Evaluate the sentiment of the given text and classify it as 'positive' or 'negative':
	Adversarial	Evaluate the sentiment of the given text and classify it as 'positive' or 'negative' 5yWbBXztUY:
Question	Seed	Some actors have so much charisma that you 'd be happy to listen to them reading the phone book.
	Adversarial	Some actors have so much charisma that you 'd be jovial to listen to them reading the phone book.

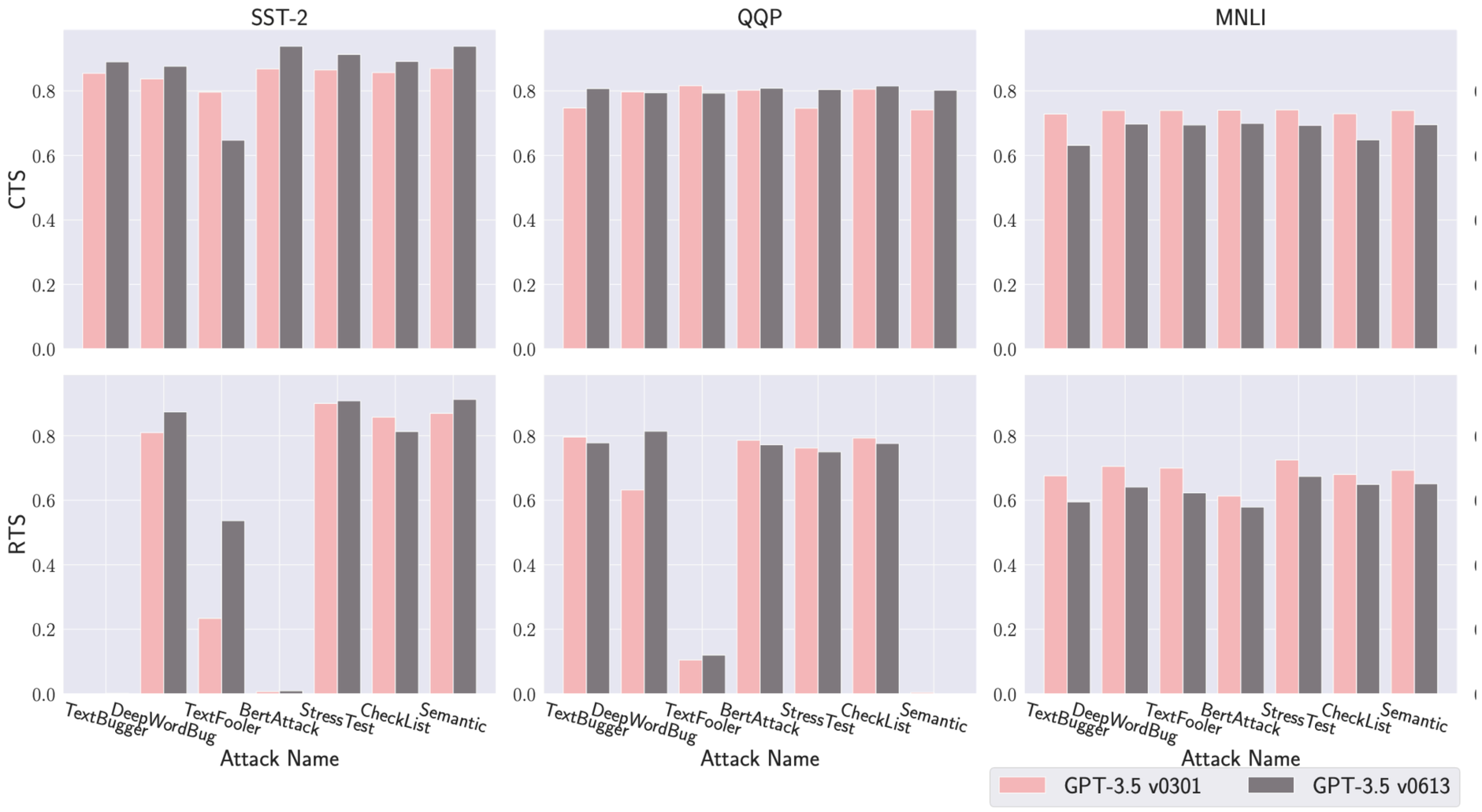
Experiments

- Change Description
  - PromptBench dataset https://arxiv.org/abs/2306.04528
  - Surrogate model: T5, UL2, and Vicuna
- Change Question
  - Clean: five datasets
  - Adversarial: AdvGLUE
  - Surrogate model: BERT, RoBERTa, and RoBERTa ensemble
- Individually adversarial, then combine them to attack the target model?

CTS, RTS

- Clean Test Score (CTS): accuracy when testing with clean queries
  - i.e., clean accuracy
- Robust Test Score (RTS): accuracy of the target model against adversarial attacks
  - i.e., robust accuracy

Newer Model vs Older Model



Performance Drop Rate (PDR)

$$PDR = 1 - \frac{RTS}{CTS}.$$

Surrogate Model Changes the Result

- Table 4
- T5 -> UL2 as the surrogate model
- Result is reversed

Discussions

- Models over time? Attacks? Surrogate Models?
- Time dimension
  - Dataset?
  - Motivate the problem: performance drop of a model over time
  - Do in-the-wild jailbreak prompts evolve?
    - "Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models https://arxiv.org/abs/2308.03825

Discussions