CS 2881 Mini Project: Follow My Instruction and Spill the Beans Reimplementation

Amir Amangeldi, Zaina Edelson, Prakrit Baruah

Abstract

By reimplementing the Spill the Beans paper (Qi et al.), we show that a simple instruction prompt can make instruction-tuned LMs verbatim copy retrieved text. While we could not recreate the exact findings from the paper, our results display significant data extraction from the models. Our variations on the original experiment include modifications to the system prompt, as well as efficiency tweaks to the chunking implementation.

Introduction

The Spill the Beans implementation included x core features:

1. System Prompt:
2. Chunking:
3. foo:
4. bar:

The Spill the Beans paper showcased strong extraction scores across model size, where extraction score increased as model size increased.

Size	Model	ROUGE-L	BLEU	F1	BERTScore
7b	Llama2-Chat-7b	**80.369**$_{\pm 1.679}$	**71.064**$_{\pm 2.033}$	83.415$_{\pm 1.375}$	**94.771**$_{\pm 0.301}$
	Mistral-Instruct-7b	79.121$_{\pm 0.653}$	68.426$_{\pm 0.857}$	**83.741**$_{\pm 0.446}$	94.114$_{\pm 0.134}$
≈13b	SOLAR-10.7b	46.109$_{\pm 3.55}$	38.595$_{\pm 3.677}$	51.224$_{\pm 3.302}$	88.148$_{\pm 0.706}$
	Llama2-Chat-13b	**83.597**$_{\pm 1.104}$	**75.535**$_{\pm 1.404}$	**85.806**$_{\pm 0.882}$	95.184$_{\pm 0.216}$
	Vicuna-13b	70.457$_{\pm 2.444}$	63.59$_{\pm 2.804}$	74.141$_{\pm 2.241}$	93.801$_{\pm 0.507}$
	Mixtral-Instruct-8x7b	80.862$_{\pm 1.226}$	70.697$_{\pm 1.501}$	85.725$_{\pm 0.979}$	**95.686**$_{\pm 0.232}$
	WizardLM-13b	74.923$_{\pm 2.399}$	66.468$_{\pm 2.468}$	77.355$_{\pm 2.279}$	92.759$_{\pm 0.517}$
≈70b	Llama2-Chat-70b	89.567$_{\pm 0.958}$	83.374$_{\pm 1.308}$	90.416$_{\pm 0.772}$	96.436$_{\pm 0.174}$
	Qwen1.5-Chat-72b	**99.154**$_{\pm 0.348}$	**98.412**$_{\pm 0.54}$	**99.138**$_{\pm 0.286}$	**99.757**$_{\pm 0.072}$
	Platypus2-Instruct-70b	83.383$_{\pm 2.235}$	80.693$_{\pm 2.39}$	83.884$_{\pm 2.125}$	96.15$_{\pm 0.463}$

Figure 1. Published results from Spill the Beans paper

References