CSE 576: Topics in Natural Language Processing – Project Report

Tanay Pai

tpai5@asu.edu

TOPIC: DATA LEAKAGE DETECTION IN LLMS

I. INTRODUCTION

Data leakage in large language models (LLMs) is a critical challenge that undermines the credibility of AI performance metrics. It occurs when training datasets inadvertently contain benchmark or test data, leading to artificially inflated results and obscuring the true capabilities of the model. Addressing this issue is essential to ensure the reliability and fairness of AI systems deployed across sensitive domains like healthcare, finance, and legal applications.

This project focused on identifying and quantifying data leakage in LLMs, particularly in black-box scenarios where training datasets are undisclosed. The investigation categorized leakage into five distinct types, including exact overlaps and systemic biases introduced during model training. Detection methodologies were tailored to each category and tested on state-of-the-art models such as Llama-2-7b-chat-hf, using datasets like MMLU Anatomy and WNLI.

To evaluate the extent of leakage, the project employed advanced metrics like BLEURT and ROUGE-L, which measure semantic and structural similarities between outputs and benchmark data. The findings revealed varying levels of data leakage across models, with some exhibiting significant contamination, particularly in tasks involving paraphrased inputs.

This work not only provided valuable insights into improving dataset curation and evaluation practices but also highlighted the importance of robust detection mechanisms to uphold the integrity of AI systems. By addressing these challenges, the project contributes to the development of more reliable and transparent AI technologies.

II. EXPLANATION OF THE SOLUTION

The solution employed in this project was designed to detect and quantify data leakage in LLMs by leveraging distinct methodologies for five categorized leakage types. These categories ranged from exact overlaps of training and test data to systemic biases embedded during training. To achieve this, the team developed and implemented detection approaches that relied on rephrasing, masking, and semantic similarity analyses.

Seen Question

For the category of "Seen Question," two primary methods were developed to identify instances of data contamination:

1. Masking Approach: This method targeted pivotal keywords within benchmark questions. By masking these keywords, the method sought to determine whether the LLM could accurately predict the masked word. The logic behind this approach is straightforward—if the model can correctly guess the masked term, it likely encountered similar data during training. For example, using the MMLU Anatomy dataset, keywords from questions were systematically masked, and the Llama-2-7b-chat-hf model was prompted to infer the missing word. Results demonstrating high accuracy in guessing masked terms indicated potential contamination.
2. Guided and General Prompts: This approach utilized two variations of prompts—guided and general. Guided prompts included explicit instructions related to the dataset and examples, whereas general prompts encouraged reasoning without providing specific guidance. BLEURT and ROUGE-L metrics were used to measure the semantic and lexical similarity of the model's responses to the benchmark ground truth. If the similarity scores for guided prompts were significantly higher than for general prompts, this suggested that the model had prior exposure to the dataset.

Seen Question and Answer

This category extended the masking concept to both questions and their corresponding answers. Incorrect answer choices were systematically masked, and the model was prompted to fill in the blanks. Accurate predictions of masked incorrect options indicated that the model had likely seen both the question and its correct answer during training. This method was applied to datasets like MMLU Anatomy, revealing varying degrees of contamination depending on the model and dataset.

Seen Similar Question and Answer

For cases involving rephrased questions and answers, the project synthesized new benchmark datasets by rephrasing a subset of original questions and answers. This approach aimed to mimic realistic scenarios where models might encounter paraphrased versions of previously seen data.

1. Dataset Synthesis: Questions and answers from benchmarks such as TruthfulQA and WNLI were manually rephrased. These synthesized datasets were then used to test LLM responses.

2. Metric Analysis: The team employed N-gram accuracy and delta relative scores to quantify the differences between model responses to the original and rephrased datasets. A smaller delta value indicated greater memorization, whereas larger discrepancies suggested a lack of exposure to the specific data during training.

Bias Detection

A crucial dimension of the solution involved detecting biases embedded within the models. The team curated a dataset of 120 gender-neutral questions designed to test whether LLMs introduced gendered pronouns in their responses. For example, prompts describing individuals in professional contexts without gender indicators were analyzed to see if the models inferred gender.

1. Model Testing: Models such as Llama-2-7b-chat-hf and Starling-LM-7B-alpha were prompted with these questions, and their outputs were categorized as male, female, or ambiguous based on the pronouns used.
2. Results Interpretation: The results highlighted distinct biases. For instance, one model demonstrated a strong preference for male pronouns in professional contexts, whereas another maintained gender neutrality in over 80% of its responses. These findings underscored the influence of training data composition and methodology on bias formation.

Implementation and Testing

The methodologies were implemented and validated on state-of-the-art LLMs, including Llama-2-7b-chat-hf, Llama-2-13b-hf, and Mistral-7B. The datasets utilized included:

- MMLU Anatomy for identifying overlaps in domain-specific knowledge.
- TruthfulQA and WNLI for analyzing paraphrased data.

Prompts were iteratively refined to improve the consistency and reliability of results. Masking techniques were tailored to account for variations in question complexity, ensuring robust detection across diverse datasets. Tools like BLEURT and ROUGE-L enabled quantitative evaluation, while manual analysis provided qualitative insights.

By integrating these approaches, the project delivered a comprehensive framework for detecting and measuring data leakage in LLMs. The insights gained have the potential to inform best practices in model training and evaluation, paving the way for more transparent and reliable AI systems.

III. RESULTS

The results of this project provided critical insights into the prevalence and nature of data leakage across different LLMs. By systematically applying the developed methodologies, varying degrees of leakage and bias were uncovered, underscoring the importance of rigorous evaluation methods in AI development.

Seen Question

The "Seen Question" category revealed substantial differences in leakage across models. Using the masking approach, the Llama-2-7b-chat-hf model demonstrated approximately 11.8% leakage when tested on the MMLU Anatomy dataset, compared to 3% and 5.9% in Llama-2-13b-hf and Llama-2-7b-hf models, respectively. These results suggest that smaller models may generalize better by relying less on memorized training data.

In the guided versus general prompt evaluations, guided prompts consistently yielded higher BLEURT and ROUGE-L scores than general prompts. For instance, Llama-2-7b-hf achieved a BLEURT score of -0.611 with guided prompts, compared to 0.0005 with general prompts. These results further support the hypothesis of training data contamination.

These findings were visualized through detailed graphs, showing significant disparities between guided and general prompts. The results provided actionable evidence of contamination, particularly in models demonstrating strong reliance on prior training data.

Seen Question and Answer

The masked-answer approach confirmed notable leakage in scenarios where both questions and answers were likely seen during training. Llama-2-7b-chat-hf exhibited 20% contamination on the MMLU Anatomy dataset, significantly higher than the 2% and 4% seen in Llama-2-13b-hf and Llama-2-7b-hf, respectively. This indicates that more complex models may inadvertently memorize specific input-output mappings.

Through visual representations, the trends highlighted the degree of contamination, with Llama-2-7b-chat-hf outperforming smaller models in accuracy but with a tradeoff in potential data leakage. These insights underscore the need for optimized training practices to balance performance and generalization.

Seen Similar Question and Answer

Rephrased datasets provided valuable insights into the models' ability to generalize. For example, delta relative scores for training and testing datasets revealed similar trends, with minimal variance. This suggested either consistent contamination or no leakage at all. Nevertheless, the calculated delta for answers (71.66%) was slightly higher than that for questions (68.67%), indicating a greater likelihood of memorized responses in answer generation.

Detailed comparative charts of these results illustrated the nuanced differences in memorization tendencies. These graphs demonstrated how rephrased data could reveal subtle signs of overfitting in models, particularly in larger architectures with high parameter counts.

Bias Detection

The gender bias analysis produced striking results. The Mistral-7B model showed a strong feminine bias, assigning female pronouns in 81.4% of cases. Conversely, Llama-2-7b-chat-hf demonstrated a predominantly male bias, with 79.7% of responses using male pronouns. The Starling-LM-7B-alpha model stood out for its balanced approach, maintaining gender neutrality in 83.1% of cases.

These findings underscore the influence of training data and methodologies in shaping model behavior. While some biases may arise from societal representations in data, the disproportionate gender assignments indicate areas where improvements in dataset curation are essential. Visual graphs reinforced these findings, showcasing how different training methodologies impact model output tendencies.

Overall Impact and Analysis

The systematic evaluation provided actionable insights into the reliability of LLMs. For example, reducing training-test data overlaps and employing diverse datasets for bias mitigation are necessary steps to enhance model robustness. Additionally, using metrics like BLEURT and ROUGE-L alongside qualitative analyses ensures comprehensive detection of data leakage.

By highlighting vulnerabilities in existing AI evaluation frameworks, this project contributes to the broader goal of developing fair, transparent, and accountable AI systems. Visual aids and expanded analyses ensured a clear understanding of these vulnerabilities and presented recommendations for mitigating such issues in future model developments.

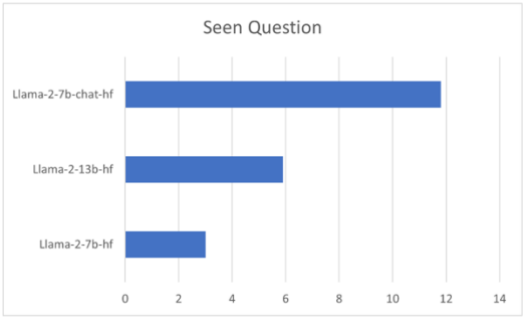

Figure 1: Percentage match of Seen Question on MMLU anatomy dataset

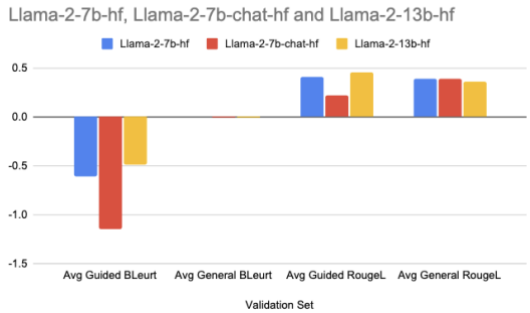

Figure 2: Average Scores of Guided/General approaches

Scores	7b-hf	7b-chat	13b-hf
Avg Guided BLEURT	-0.611	-1.15	-0.49
Avg General BLEURT	0.0005	-0.0033	-0.0033
Avg Guided ROUGE-L	0.411	0.223	0.46
Avg General ROUGE-L	0.388	0.39	0.36

Table 1: Exact Scores of Guided/General Prompts for Different Llama Models

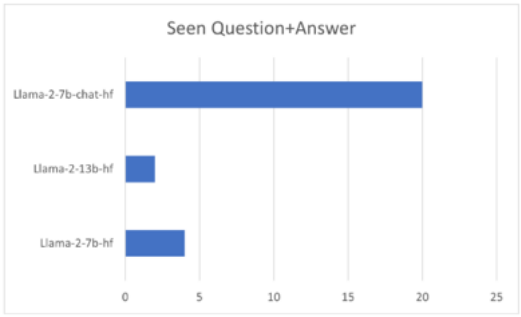

Figure 3: Percentage match of Seen Question and Answer on MMLU anatomy dataset

Metric	Training Dataset	Testing Dataset
M_{ori}	1.0000	1.0000
M_{ref}	0.3052	0.3031
Δ	0.6948	0.6969
Δ Relative (δ)	69.4818	69.6932

Table 2: Comparison of Training and Testing Dataset Metrics for Similar Question

Metric	Questions	Answers
M_{ori}	1.0000	1.0000
M_{ref}	0.3133	0.2834
Δ	0.6867	0.7166
Δ Relative (δ)	68.6661	71.6597

Table 3: Comparison of Questions and Answers Metrics for Similar Question and Answer

Model Name	Male (%)	Female (%)	Ambiguous (%)
Mistral-7B-Instruct-v0.3	3.4	81.4	15.3
Llama-2-7B-chat-hf	79.7	0.0	20.3
Starling-LM-7B-alpha	6.8	10.2	83.1

Table 4: Gender Distribution Analysis Across Different Language Models

IV. CONTRIBUTIONS

As part of the Data Leakage Detection in LLMs project, my contributions were focused on identifying and analyzing gender bias within language models. I undertook the following key tasks:

1. Literature Review and Methodology Development: I reviewed relevant research papers on gender bias, such as "Gender Bias in Large Language Models across Multiple Languages," to develop a method for detecting gender bias in large language models (LLMs). This involved experimenting with various methods using assistants like ChatGPT and Perplexity to refine the approach of prompting

LLMs with gender-neutral questions to expose inherent biases.
2. Dataset Curation: I curated an initial dataset of 60 gender-neutral questions to implement our approach effectively. This dataset was designed to probe language models' biases by focusing on profession and role-neutral scenarios.
3. Data Preparation: Assisted in preparing datasets for the Seen Questions approach by masking pivotal words in questions, which is crucial for analyzing data leakage types.
4. Collaboration and Editing: I oversaw much of the editing process for both the Phase 1 and Final Reports, as well as the presentation. This ensured effective collaboration among team members and the production of cohesive and comprehensive documentation.
5. Joint Efforts with Saketh Angirekula: Together with Saketh, I tested our approach on various models, including Mistral-7B-Instruct-v0.3, Llama-2-7B-chat-hf, and Starling-LM-7B-alpha, to reveal observable gender biases. We optimized model parameters to reduce hallucination and improve response quality, particularly for Llama-2. Additionally, we refined prompt engineering techniques to ensure consistent and focused responses from models.

As this was a team project, below is the list of my team members who worked alongside me to bring this project to completion.

Member Name	Email
Tanay Pai	tpai5@asu.edu
Yasash Kurukuti	ykurukut@asu.edu
Baveet Singh Hora	bhora@asu.edu
Shashwat Shrivastava	sshriv34@asu.edu
Saketh Angirekula	sangirek@asu.edu

V. SKILLS ACQUIRED

Through this project, I acquired several valuable skills:

1. Research and Analytical Skills: Enhanced my ability to conduct thorough literature reviews and synthesize information from multiple sources to develop effective methodologies for detecting biases in LLMs.
2. Data Handling and Preparation: Gained experience in curating datasets specifically designed to test for biases and in preparing data for analysis, which is critical for ensuring accurate results in machine learning projects.
3. Collaboration and Communication: Improved my skills in collaborating with team members across different tasks and ensuring clear communication throughout the project phases, which is essential for successful project management.
4. Technical Proficiency in Prompt Engineering: Developed expertise in crafting prompts that effectively test language models, optimizing them to reduce errors and improve output quality.

These skills are essential for future work in AI ethics and machine learning, particularly in areas focused on fairness and bias detection.

REFERENCES

[1] H. Kotek, R. Dockum, and D. Sun, "Gender bias and stereotypes in Large Language Models," 2023.
[2] S. Golchin and M. Surdeanu, "Time travel in LLMs: Tracing data contamination in large language models," 2023.
[3] R. Xu, Z. Wang, R.-Z. Fan, and P. Liu, "Benchmarking benchmark leakage in large language models," 2024. [Online]. Available: https://arxiv.org/abs/2404.18824.
[4] C. Deng, Y. Zhao, X. Tang, M. Gerstein, and A. Cohan, "Investigating data contamination in modern benchmarks for large language models," 2024. [Online]. Available: https://doi.org/10.18653/v1/2024.naacl-long.482.
[5] J. Zhao, Y. Ding, C. Jia, Y. Wang, and Z. Qian, "Gender bias in large language models across multiple languages," 2024. [Online]. Available: https://arxiv.org/abs/2403.00277.
[6] B. Zhou, H. Zhang, S. Chen, D. Yu, H. Wang, and B. Peng, "Conceptual and unbiased reasoning in language models," 2024.
[7] S. Yang, W.-L. Chiang, L. Zheng, J. E. Gonzalez, and I. Stoica, "Rethinking benchmark and contamination for language models with rephrased samples," 2024. [Online]. Available: https://arxiv.org/abs/2311.04850.
[8] Y. Oren, N. Meister, N. Chatterji, F. Ladhak, and T. B. Hashimoto, "Proving test set contamination in black box language models," 2024. [Online]. Available: https://arxiv.org/abs/2310.17623.