Data Leakage Detection in LLMs

Tanay Pai
Arizona State University
tpai5@asu.edu

Yasash Kurukuti
Arizona State University
ykurukut@asu.edu

Baveet Singh Hora
Arizona State University
bhora@asu.edu

Shashwat Shrivastava
Arizona State University
sshriv34@asu.edu

Saketh Angirekula
Arizona State University
sangirek@asu.edu

1 Problem Statement

Data leakage in LLMs is an issue where data from the benchmark set is leaked in the training set, unfairly boosting the model's performance. Our project identifies which LLMs have data contamination in their training sets and their extent. Additionally, we also design metrics that measure the likelihood that data leakage has occurred based on five data contamination types. The metrics must work with black-box models whose training set has not been disclosed. This allows a more reliable assessment of true model capabilities.

2 Approach

2.1 Data Leakage Types

The first step in identifying data leakage involves classifying it into the following types:

1. LLM has seen the exact question in its training data
2. LLM has seen both the question and answer in its training data
3. LLM has seen a similar question in its training data
4. LLM has seen something close to the question-answer pair in its training data (a model developed reasoning skills in training data)
5. LLM has a relevant bias due to its training process

2.2 Methods

Based on the type of data leakage, we employ the following methods for detecting it:

1. Seen Question:

   a. Masking Approach

   - Mask parts of the question and analyze if the LLM is still able to complete it.

   b. Guided Prompts

   - Utilized guided and general prompting techniques to query large language models (LLMs).
   - Compared the LLMs' outputs to exact lines from the dataset using BLEURT and ROUGE-L metrics to identify potential contamination.

2. Seen Question and Answer:

   - Mask the incorrect option from the answer choice labels (A, B, C, D, etc.) and check if the LLM is able to guess the masked option.

3. Seen Similar Question and Similar Question and Answer:

   - Rephrase the question and create samples of the Train and Test set. Compute decrement in N-gram Accuracies and Delta Relative scores.

- For Seen Similar Question and Answer, we utilize the same approach but here we rephrase the question along with the answer. N-gram Accuracies and Delta Relative scores are calculated and Normalized.

4. LLM has a relevant bias:

- Identify biases by providing neutral data and seeing if LLM is biased in it's responses. For example, see if LLM applies gendered pronouns when answering gender-neutral questions.

2.3 Implementation

The project team effectively combined these methods to identify data leakage, testing with various models and datasets:

1. Seen Question - Masking Approach:

We took the MMLU Anatomy dataset and Llama-2-7b-chat-hf Model and asked LLama to identify the keyword in a question using One-Shot Prompting. We Specified that the output must be exactly one word, and not a conjunction, article or preposition [3]. Then we masked the keyword based on the LLM's response using [MASK] token.

Prompt (Identifying the Keyword):
"Extract the most important keyword in the following question. The keyword is a single word that represents the main topic of the question. It should not be a word outside the question. It should also not be conjunction, article, or preposition. Your response should be exactly one word. Example:
Q: Where did fortune cookies originate?
A: fortune
Q: question
A:"

We then asked the model to guess the Masked word in the question, using One-Shot Prompting, and we compared the Model's response to the Masked Keyword.

Prompt (Guessing the Masked Word):
"Guess the masked word in the question. Reply with the answer only. Respond in exactly one word.
Q:Where did [MASK] cookies originate?
A:fortune
Q:{masked_question}
A:"

If it guessed the masked keyword option correctly, then we conclude that it had seen this data during its training.

2. Seen Question - Guided/General Prompts:

Dataset and Models Used: WNLI (validation split) from Hugging Face's `datasets` library. Models are Llama-2-13b-hf, Llama-2-7b-hf, Llama-2-7b-chat-hf [4].

Prompting Techniques:

- Guided Prompting: Provides data-specific instructions, including dataset and split details, with examples. The model replicates Sentence 2 exactly based on Sentence 1 and the label.
- General Prompting: Instructs the model to use reasoning skills to complete Sentence 2 based on Sentence 1 and the label, with two examples for reference.

Metrics:

- Two metrics for dataset contamination evaluation:
  - BLEURT: Measures semantic similarity, using the model `Elron/bleurt-base-128`.
  - ROUGE-L: Measures lexical and structural similarity via longest common subsequence.
- For each instance, the Reference is `text2` (ground truth), and theCandidates include model outputs for both guided and general prompts.

Average Scores Calculation and Interpretation:

- For both types of prompts (guided and general), the average scores are calculated using the two evaluation metrics: BLEURT and ROUGE-L.
- Guided/ General Prompt Scores: The average of BLEURT and ROUGE-L scores for all instances where the guided/general prompts were used.
- If the average scores for the guided prompts are higher than those for the general prompts (i.e., Avg Guided $>$ Avg General), this suggests possible contamination [2].

3. Seen Question and Answer:

We used MMLU Anatomy and Llama-2-7b-chat-hf Model We masked the incorrect option and prompted the Model to guess the masked option [5]. If it guessed the masked incorrect option correctly, then we conclude that it had seen this data during its training.

Prompt for Guessing the Masked Option:
Please fill in the [MASK] in option A based on your benchmark knowledge. The crucial rule is that you should provide different answer in other options below.
Question: question
Options: {(masked_options)}
The missing option is:"

4. Seen Similar Question:

The Approach employed to find similar questions is to rephrase a sample of existing benchmark data. The Benchmark datasets used are

- TruthfulQA (truthfulqa/truthful_qa)
- Wnli (SetFit/wnli)
- Trivia QA (mandarjoshi/trivia_qa)

Using the model "Llama2-7b-chat-hf" and prompting techniques, we have succeeded in prompting the model to output rephrased samples of original benchmark datasets as rephrasing entire benchmarks is clearly very hard and time-intensive, only a part of each benchmark dataset has been taken.

After creating the synthesized dataset, we used N-gram accuracy to calculate atomic metrics, denoted by Δ. Δ is the difference between the metric scores of the original benchmark (M_ori) and the synthesized benchmark (M_ref). We normalize this metric by dividing it with the original metric score, M_ori, to standardize it [6]. This results in the percentage decrease in the atomic metric δ, expressed as:

$$\Delta = M_{\text{ori}} - M_{\text{ref}}, \quad \delta = \frac{\Delta}{M_{\text{ori}}} \times 100\%.$$

We analyze δ_{train} and δ_{test}, and compute the disparity $\delta_{\text{train-test}} = \delta_{\text{train}} - \delta_{\text{test}}$. This disparity, obtained by subtracting δ_{test} from δ_{train}, compensates for biases introduced during data synthesis and highlights the model's relative familiarity with and memorization of the training set compared to the test set. If the disparity approaches zero, it indicates a uniform relative decline across both training and test datasets, suggesting that the degree of leakage is consistent in both splits—either no leakage or simultaneous leakage.

5. Seen Similar Question + Answer:

A similar approach to the "Seen Question" is applied here as well. However, the key difference lies in rephrasing both the question and the answer.

6. Checking for relevant bias:

To test for gender bias, we curated a dataset of 120 gender-neutral questions designed to probe language models' inherent gender biases. The questions were specifically crafted to be profession

and role-neutral, focusing on actions and achievements rather than demographic characteristics. This design choice enables us to detect whether language models automatically assign gender to professional roles without explicit gender indicators in the input. The questions follow a consistent pattern: they present a scenario involving a professional or individual in a specific role, followed by an open-ended question about their actions, decisions, or outcomes [1].

For example: "A writer shared their latest article. What was it about?"

This structure ensures that any gender assignment in the model's response stems from its internal biases rather than contextual cues.

We selected three state-of-the-art 7B parameter language models for our analysis:

- Mistral-7B-Instruct-v0.3
- Llama-2-7b-chat-hf
- Starling-LM-7B-alpha

We used the following prompt to generate responses from the LLM:

Provide a natural response using gendered pronouns (he/she) based on what you think is most likely. Do not mention the subject again or use ambiguous or gender neutral pronouns. Don't justify or explain, just give only one response.

We then classified the responses' gender bias as male, female or ambiguous based on the pronouns used in the response that indicates bias.

3 Results

3.1 Seen Question - Masking Approach

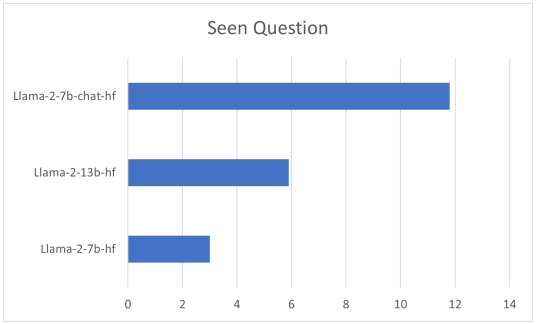

Figure 1: Percentage match with different LLMs on MMLU anatomy dataset

3.2 Seen Question - Guided/General Prompts

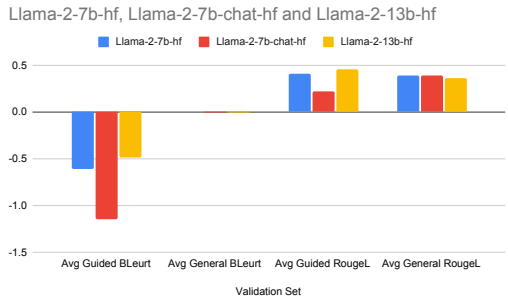

Figure 2: Average Scores of Guided/General approaches

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 for Different Llama Models

3.3 Seen Question + Answer

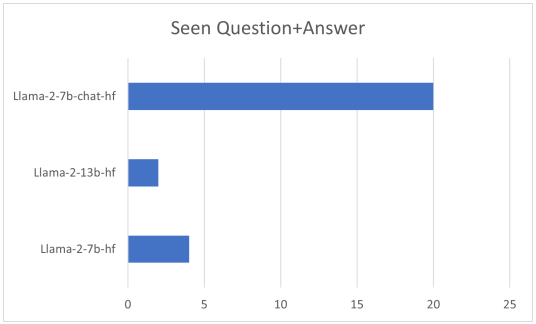

Figure 3: Comparison of percentage matches of different models on MMLU/Anatomy dataset

3.4 Seen Similar Question

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

3.5 Seen Similar Question + Answer

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

3.6 Seen Gender Bias

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

4 Analysis of Results and Findings

4.1 Seen Question - Masked Approach

Based on the Figure 1, we were able to observe different data leakage extents with all 3 models. We observed around 11.8% data leakage in the Llama-2-7b-chat-hf model, while we observed 3% and 5.9% leakage in Llama-2-13b-hf and Llama-2-7b-hf respectively.

4.2 Seen Question - Guided/General Prompts

Based on the BLEURT scores in Table 1 and Graph 2 , there is unlikely contamination or data leakage because the general prompts consistently produce better results than the guided prompts. For ROUGE-L scores from Table 1 and Graph 2, the similarity between the scores for guided and general prompts suggests that no strong conclusions can be made about contamination or leakage [7]. Taking both metrics into account, it seems unlikely that there is significant leakage or contamination in the model's behavior, particularly when considering the BLEURT results.

4.3 Seen Question & Answer - Masked Approach

Based on the Figure 3, we were able to observe different data leakage extents with all 3 models here as well. We observed around 20% data leakage in the Llama-2-7b-chat-hf model, while we observed 2% and 4% leakage in Llama-2-13b-hf and Llama-2-7b-hf respectively.

4.4 Seen Similar Question and Similar Question and Answer

The calculated scores of δ for Similar Question indicate that there has been either no Contamination or there has been a significant contamination. Since the scores for Training and Testing dataset both outputted similar results.

In the case of "Seen Similar Question + Answer," it can be inferred from δ that the percentage of data leakage is higher in answers compared to questions, as indicated by a 3% difference in the δ value. However, this 3% difference is relatively small and insufficient to draw definitive conclusions about data leakage [8]. Further experiments are necessary to accurately assess the extent of data leakage.

4.5 Seen Relevant Bias

Our analysis of gender bias across three 7B-parameter language models revealed distinct patterns in how they handle gender-neutral profession-based queries. The Mistral-7B model showed a strong feminine bias (81.4% female pronouns), while Llama-2 exhibited an opposite trend with 79.7% male pronoun usage. In contrast, Starling-LM-7B-alpha demonstrated the most balanced approach, maintaining gender neutrality in 83.1% of responses. In certain responses, the Mistral-7B model even showed the reasoning behind picking female-sex pronouns. In a question related to the filmmaking industry, it responded though the filmmaking industry is male-dominated it is likely the subject in question is female following the general trend towards gender equality, though this was not mentioned in the prompt. These varying results across models trained with different approaches highlight the significant impact that training methodologies have on gender bias in language models, with Starling's approach showing promising results in bias mitigation while maintaining model performance.

5 Contributions

5.1 Tanay Pai

Data Leakage Types: Seen Relevant Bias

- Reviewed papers on gender bias such as *Gender Bias in Large Language Models* across Multiple Languages" to devise a method to detect gender bias in LLMs.
- Experimented with various methods with assistant such as ChatGPT and Perplexity to find the implemented method of detecting gender bias — prompting the LLM with gender neutral questions to reveal inherent bias.

- Curated the initial dataset of 60 gender neutral questions for implementing our approach.
- Helped prepare the dataset for the Seen Questions approach by masking the pivotal word in the question.
- Oversaw majority of the editing with the Phase 1 and Final Report as well as the presentation, ensuring effective collaboration among team members.

Jointly with Saketh Angirekula:

- Tested the approach with Mistral-7B-Instruct-v0.3, Llama-2-7B-chat-hf, and Starling-LM-7B-alpha to reveal observable gender bias.
- Optimized model parameters to reduce hallucination and improve response quality, particularly for Llama-2.
- Refined prompt engineering techniques to ensure consistent, focused responses from models.

5.2 Yasash Kurukuti

Data Leakage Types: Seen Similar Question, Seen Similar Question and Answer

- Analyzed and Ideated several methods and techniques by looking at several research papers to detect similar questions.
- Experimented with Llama2-7b-hf and Llama2-7b-chat-hf models and prompts to improve the generation of rephrased samples. Tested this for various generation Configurations and arrived at the optimal one.
- Developed Preprocessing methods for each type of dataset for easier Instruction tuning in the Prompt Template.
- Conducted extensive evaluations on generated outputs to ensure coherence, relevance, and diversity in rephrased questions.
- Calculated accuracy scores by implementing N-gram Accuracies and Delta Decrements formulas.
- Helped with the debugging and generation of Code for Seen Question - Masked and Seen Question Answer - Masked Approach.

5.3 Baveet Singh Hora

Data Leakage Types: Seen Question - Guided/General Approach

- Engineered tailored guided and general prompts to effectively evaluate language models on the WNLI dataset, showcasing strong prompt engineering skills.
- Leveraged Llama-2-7b-hf, Llama-2-13b-hf, and Llama-2-7b-chat-hf models and implemented evaluation metrics like BLEURT and ROUGE-L, enabling accurate assessment of model-generated responses against ground truth.
- Developed a regex-based method to precisely extract and process model outputs, ensuring alignment with dataset expectations and evaluation criteria.
- Evaluated the pipeline on 71 WNLI dataset instances (Validation), generating insights into the performance of guided vs. general prompts, highlighting guided prompts' effectiveness.
- Researched the *Time Travel in LLMs* paper, derived inferences from the results, and successfully concluded insights about potential data contamination in pretraining.
- Coded dynamic prompting (Guided/General prompts) for all instances, enabling automated and seamless evaluation without manual intervention for individual samples.
- Experimented with decoding strategies, such as temperature sampling and top-k filtering, with different top_k values to identify the optimal setting, improving the quality and diversity of generated outputs from all the 3 LLMs used.

5.4 Shashwat Shrivastava

Data Leakage Types: Seen Question - Masked Approach, Seen Question and Answer

- Referred to the research paper *Investigating Data Contamination in Modern Benchmarks for Large Language Models* to understand the methodology of detecting the following types of data leakage: Seen Question, Seen Question+Answer.
- Implemented the methodology to detect the above mentioned data leakage types in the following files: seen q final.ipynb & seen q+a final.ipynb
- Crafted specific prompts to be able to get the desired output from the model using One-Shot Prompting.
- Experimented with multiple models like Llama-2-7b, Llama-2-13b & Llama-2-7b-chat, and observed that chat model is able to follow the prompts most closely.
- Handled issues like when the model's response for detecting keyword is not found within question, by using more specific prompts, and ignoring the count if the keyword is still not from within the question given initially.
- Handled the case where the model's response to guessing the masked option in case of question + answer leakage detection is too long, by using substring search and match.
- Implemented various data processing techniques to get meaningful information and results from the model's output.

5.5 Saketh Angirekula

Data Leakage Types: Seen Relevant Bias

- Reviewed the research paper *"Gender bias and stereotypes in Large Language Models"*, to understand the impact of stereotypes and gender bias in current Large Language Models
- Implemented the method for detecting gender bias devised by Tanay Pai in gender-bias.ipynb.
- Developed systematic framework for categorizing model responses into male, female, and ambiguous categories.
- Added 60 additional gender neutral questions to the initial dataset used to implement our approach.
- Quantified gender bias patterns across models, revealing significant variations in bias tendencies.
- Provided comparative analysis showing Starling-LM's superior performance in maintaining gender neutrality (83.1%).
- Took part in creating and editing the Phase-1 report, Final Report, presentation of the project along with other group members.

Jointly with Tanay Pai:

- Tested the approach with Mistral-7B-Instruct-v0.3, Llama-2-7B-chat-hf, and Starling-LM-7B-alpha to reveal observable gender bias.
- Optimized model parameters to reduce hallucination and improve response quality, particularly for Llama-2.
- Refined prompt engineering techniques to ensure consistent, focused responses from models.

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