

# Finding Data Contamination via TS-Guessing

Data Contamination WS24/25

Long Kim/Tim Kolber

## 1. Introduction

With the increase in training data used for model training, data contamination becomes a bigger issue every year. This is accompanied by closed model architecture, further making analysis and understanding behind LLMs more difficult. But even for smaller models, one can find contaminated corpora used in their training data, either intentionally to increase model performance, or unintentionally through massive corpus collection. In this paper we try to uncover some of this contamination in 3 different models via TS-Guessing (Deng et al., 2024<sup>1</sup>). In comparison to their approach, we include *multi-masking* to additionally test how context can affect the mask filling task and if models can fully answer with correct “wrong answers” or if it still only occasionally finds the correct ones. Our results show a similar trend to this paper, with Mistral being highly contaminated on *single* masks, but also that the ARC dataset as a whole has leaked into other models. Surprisingly the results for multi-masking show an unintuitive result: Models can guess multiple masks on a similar level, if not better level than single masks.

## 2. Method

Our method is inspired by Deng et al., 2024 who utilized masking in multiple choice datasets to detect data contamination. The important difference to other masking methods is the masking of an *incorrect choice* to not punish a model for simply predicting the correct answer of a question. Due to an almost infinite answer space for incorrect answers, predicting a wrong choice “correctly” should have very low probability, unless the model got contaminated with the exact choices and thus memorized them. With how datasets are set up though, the actual space of possibilities for incorrect answers is restricted to the context given. One of these restricted spaces can be “Yes/No” or “True/False” questions, which we filter out.

Deng et al. try to remove contextual clues by removing questions where the Rouge-L F1 score between any options exceeds 0.65. We on the other hand try to differentiate by implementing multiple option masking. With this we can test if by removing even more answers (therefore also more context and guides for constructing an answer) if the model is still capable of predicting correctly.

---

<sup>1</sup> <https://arxiv.org/pdf/2311.09783>

We utilize one-shot prompting to guide model responses. Our prompt overall is fairly simple, and while we attempt to restrict the response of a model to simply filling the mask with their predicted choice, it is not important for the actual accuracy calculation (as seen later in 3.3 *Metrics*).

## 3. Experiment

### 3.1 Domain

We evaluate 3 datasets: MMLU and OpenBookQA are used as reference to Deng et al., and to compare our results to their already available results. Both of these are used to test a models knowledge aspect. Furthermore we take a look at the “ARC-Challenge” and “ARC-Easy” datasets. These datasets evaluate the abstraction and reasoning capabilities of models. “ARC-Challenge” contains harder questions that where answered wrongly by both a retrieval-based algorithm and a word co-occurrence algorithm. There were other datasets we also considered, such as MedMCQA, though they lacked the correct choice in their test splits, thus making our template unusable.

### 3.2 Models

Our models are lightweight and open-source to ensure easier analysis. We utilize “LLama-3-8b-Instruct”, “Mistral-7b-Instruct” and “Phi-4”(14b). Both the LLama and Mistral models should act as reference to Deng et al. and helps us compare our method to theirs.

### 3.3 Metrics

We chose accuracy as the metric to measure how many wrong answers the model could correctly predict. Concretely, if the answer behind the mask is found anywhere in the model’s response, it counts as a correct prediction. We sum up all the correct predictions and divide by the total number of examples. As a result, we get a percentage of correct predictions the model should theoretically not be able to give.

Of course, in practice, the model will sometimes randomly guess the correct predictions, or the correct prediction can be logically determined. Therefore, it is important to compare the scores by different models on the same dataset.

For out multi-masking approach we utilize 2 different accuracy methods: One views each masked answer *separately*, and checks if this masked answer can be found anywhere in the model response. This full tally is then divided by the number of example multiplied by the how many masks were applied, to normalize the results. The second approach is the grouped accuracy which

only counts a response as correct if *all* masked answers can be found in the model response. We have set our mask amount to 2 for these experiments.

## 3.4 Results

### 3.4.1 Single masking

Single masking	LLama	Mistral	Phi-4
ARC-Easy	3.66%	7.58%	2.65%
ARC-Challenge	4.10%	6.31%	3.84%
OpenBookQA	0.80%	2.80%	0.80%

Fig.1: Exact matches via TS-Guessing

Our results show a similar picture as in the original paper: Mistral seems to be the most contaminated model by quite a wide margin, being able to correctly guess a significantly higher amount of incorrect choices. Exact Matches by Deng et al. for Mistral over OpenBookQA reach 10% vs. our 2.8%.

Interestingly, our results for our LLama3 model with 8b parameters over the OpenBookQA are significantly lower, only reaching 0.8% in comparison to exact matches in Deng et al. of 4% from LLama2 with 13b parameters. This could indicate

What is surprising is that Phi-4 with its 14B parameters actually reaches the lowest amount of exact matches, thus displaying potentially the lower amount of contamination. This indicates that stronger models do not necessarily perform “better”, both based on general performance, but also based on potential contamination.

Another comparison can be made between the contamination of the different partitions of ARC, where Mistral shows higher contamination for “ARC-Easy”, whereas the other 2 models show higher contamination for the harder “ARC-Challenge”.

### 3.4.2 Multi-Masking

Multi-masking (separate)	LLama	Mistral	Phi-4
ARC-Easy	2.52%	6.21%	10.44%
ARC-Challenge	2.30%	5.59%	7.25%
OpenBookQA	2.40%	3.90%	5.20%

Fig.2: Separate Accuracy via TS-Guessing

Our multi-masking results are significantly more surprising: Models often perform similarly, if not *significantly* better than on single masks. In terms of separate accuracy, LLama overall performs worse on ARC, but much better on OpenBookQA, with Mistral showing a similar trend. The massive outlier here is Phi-4, improving upon every single dataset and even eclipsing the results of Mistral.

Multi-masking (grouped)	LLama	Mistral	Phi-4
ARC-Easy	0.97%	1.85%	4.33%
ARC-Challenge	0.60%	2.30%	4.35%
OpenBookQA	0.40%	0.80%	0.80%

Fig.3: Grouped Accuracy via TS-Guessing

Adding our grouped accuracy we see a similar trend: LLama has the lowest scores, with Mistral seemingly more contaminated. Phi-4 again performs on a very high level, correctly guessing both masked answers on both ARC datasets. We assume that this behaviour of Phi-4 is caused by it initially, with just one mask, trying to give a general answer directly based on the prompt given, whereas with multiple masks it suddenly uses its background knowledge and training data to answer due to missing context. Looking at some of the entries, it's not just responses that are short or easily inferred from context (as seen by LLama and Mistral), but also longer responses that cannot have been guessed. Some of these grouped examples can be seen in Appendix B. One thing to note is that often times the positions of the correct wrong answers are switched.

## 4. Future Research

When it comes to masking multiple choices, our experiment merely checks if each masked answer is contained *anywhere* in the response, meaning that the actual position does not matter for our accuracy. It is somewhat important to differentiate wrong from correct positions, and this could for example be achieved by applying a penalty. Even just getting the choices themselves though already proposes potential data contamination, so in our opinion the penalty should be light, especially because responses won't always perfectly align with a suitable response format.

Another important distinction is how precise the output of the model is. With increasing amount of tokens contained in an option, occasionally the model will have a close, but not exact match. This is comparable to the issue of retrieval-based contamination detection where defining a proper threshold can be difficult, though in this case even a longer response can still be restricted by similar sentence structure where all options start with the same few tokens for example.

## 5. Conclusion

Based on our results we have found comparable results to the research done by Deng et al. when it comes to masking a singular answer from a multiple choice question. While our results are different, this should not be entirely surprisingly based on potentially very different code structures and random seeds. We have additionally tested a multi-masking approach to test how a model might respond to missing context, rather than removing questions with similar options outright. This showed us much higher than expected accuracies. Initially, while this seems unintuitive at first sight, we assume that missing context forces models to rely more on responding with leaked data. We also find that while Phi-4 performed the worst out of all models for single masking, it suddenly spikes with very high values on multi-masking, even getting several grouped predictions correct that contain many tokens. While some of these answers could be inferred from the still existing context, the other 2 models did not perform as successfully. An interesting quirk with Phi-4's correct predictions is that often times it would switch the position of the answers. Further research could apply penalties to such switched, resulting in more precise accuracy measures.

# Appendix

## A Prompt

*Please try to fill in a wrong option into [MASK]. Provide only that option. The answer should be different from the other options.*

*Example:*

**Question: Which of the following is a bird?**

**Options:**

A. Dove

B. [MASK]

C. Fox

D. Dolphin

**Answer: Dog**

**Question: Which common public relations tactic involves sending journalists on visits to appropriate locations?**

**Options:**

A. Media release

B. Media tour

C. Press room

D. [MASK]

**Answer:**

## B Grouped Examples of Phi-4

```
"['Velocity and acceleration both become negative.', 'Velocity becomes negative but acceleration remains positive.']", "userPlease try to fill in a wrong option into each [MASK]. Provide only those options. The answers should be different from the other options.
```

```
Example:
```

```
Question: Which of the following is a bird?
```

```
Options:
```

- A. Dove
- B. [MASK]
- C. Fox
- D. Dolphin

```
Answer: Dog
```

```
Question: When a person parachutes from a plane, she travels at about 67 meters per second. When she pulls her cord and her chute opens, she slows to about 3 meters per second. If the velocity and acceleration are both considered positive before the chute opens, how do her velocity and acceleration change when the chute opens?
```

```
Options:
```

- A. Velocity and acceleration both remain positive.
- B. [MASK]
- C. Velocity remains positive but acceleration becomes negative.
- D. [MASK]assistantB. Velocity becomes negative but acceleration remains positive.
- D. Velocity and acceleration both become negative."

Found in ARC-Challenge

```
"['during the winter near the North Pole', 'during the winter near the equator']", "userPlease try to fill in a wrong option into each [MASK]. Provide only those options. The answers should be different from the other options.
```

```
Example:
```

```
Question: Which of the following is a bird?
```

```
Options:
```

- A. Dove
- B. [MASK]
- C. Fox
- D. Dolphin

```
Answer: Dog
```

```
Question: When and where would the number of daylight hours each day be greatest?
```

```
Options:
```

- A. [MASK]
- B. during the summer near the equator
- C. [MASK]
- D. during the summer near the North PoleassistantQuestion: When and where would the number of daylight hours each day be greatest?

```
Options:
```

- A. during the winter near the equator
- B. during the summer near the equator
- C. during the winter near the North Pole
- D. during the summer near the North Pole

```
Answer for A: during the winter near the equator
```

```
Answer for C: during the winter near the South Pole"
```

Found in ARC-Challenge: Phi-4 gives the correct wrong answers directly in the options of its response, but then answers with 1 wrong answers additionally. This can be an example of the model simply already having the answers available, and then trying to actually answer “naturally”.



```
"['The mass of the statue is greater than the mass of the pieces.', 'The mass of the pieces is greater than the mass of the statue.']", "userPlease try to fill in a wrong option into each [MASK]. Provide only those options. The answers should be different from the other options.
```

```
Example:
```

```
Question: Which of the following is a bird?
```

```
Options:
```

- A. Dove
- B. [MASK]
- C. Fox
- D. Dolphin

```
Answer: Dog
```

```
Question: The Statue of Liberty was built in France and then taken apart into 350 pieces for shipping. When the pieces arrived in New York, the statue was put together. Which statement describes the relationship between the mass of the 350 pieces and the mass of the whole statue?
```

```
Options:
```

- A. [MASK]
  - B. The mass of the pieces is the same as the mass of the statue.
  - C. [MASK]
  - D. The mass of the statue is not related to the mass of the pieces.
- ```
assistantAnswer for A: The mass of the pieces is greater than the mass of the statue.
```

Found in ARC-Easy: These correct answers could be inferred from the structure of options given, but due to us testing it literally, it is still surprising to see an exact match.