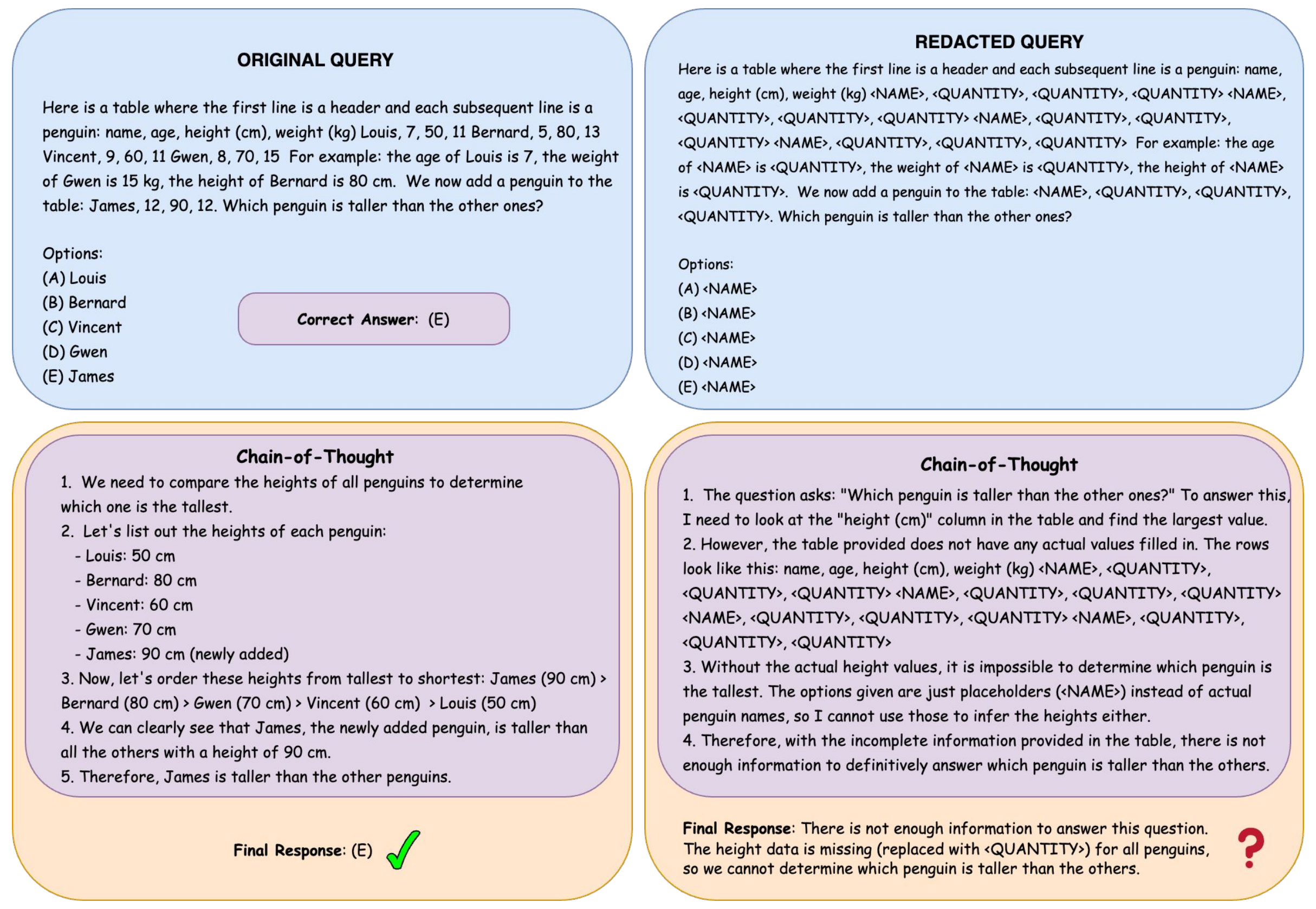
The Empirical Impact of Data Sanitization on Language Models

Anwesan Pal, **Radhika Bhargava**, Kyle Hinsz,
Jacques Esterhuizen, Sudipta Bhattacharya

Problem statement

Data privacy is a critical concern in the development and use of language models (LMs) due to presence of personally identifiable information (PII).

One approach to ensure data privacy is **Data Sanitization** which involves complete and irreversible removal of PII from data. Despite the wide-adoption of data sanitization methods, its impact on the performance of language models has not been studied in-depth.



ORIGINAL QUERY

Here is a table where the first line is a header and each subsequent line is a penguin: name, age, height (cm), weight (kg) Louis, 7, 50, 11 Bernard, 5, 80, 13 Vincent, 9, 60, 11 Gwen, 8, 70, 15 For example: the age of Louis is 7, the weight of Gwen is 15 kg, the height of Bernard is 80 cm. We now add a penguin to the table: James, 12, 90, 12. Which penguin is taller than the other ones?

Options:
(A) Louis
(B) Bernard
(C) Vincent
(D) Gwen
(E) James

Correct Answer: (E)

REDACTED QUERY

Here is a table where the first line is a header and each subsequent line is a penguin: name, age, height (cm), weight (kg) <NAME>, <QUANTITY>, <QUANTITY>, <QUANTITY> <NAME>, <QUANTITY>, <QUANTITY>, <QUANTITY> <NAME>, <QUANTITY>, <QUANTITY>, <QUANTITY> <NAME>, <QUANTITY>, <QUANTITY>, <QUANTITY> For example: the age of <NAME> is <QUANTITY>, the weight of <NAME> is <QUANTITY>, the height of <NAME> is <QUANTITY>. We now add a penguin to the table: <NAME>, <QUANTITY>, <QUANTITY>, <QUANTITY>. Which penguin is taller than the other ones?

Options:
(A) <NAME>
(B) <NAME>
(C) <NAME>
(D) <NAME>
(E) <NAME>

Chain-of-Thought

1. We need to compare the heights of all penguins to determine which one is the tallest.
2. Let's list out the heights of each penguin:
   - Louis: 50 cm
   - Bernard: 80 cm
   - Vincent: 60 cm
   - Gwen: 70 cm
   - James: 90 cm (newly added)
3. Now, let's order these heights from tallest to shortest: James (90 cm) > Bernard (80 cm) > Gwen (70 cm) > Vincent (60 cm) > Louis (50 cm)
4. We can clearly see that James, the newly added penguin, is taller than all the others with a height of 90 cm.
5. Therefore, James is taller than the other penguins.

Final Response: (E) ✓

Chain-of-Thought

1. The question asks: "Which penguin is taller than the other ones?" To answer this, I need to look at the "height (cm)" column in the table and find the largest value.
2. However, the table provided does not have any actual values filled in. The rows look like this: name, age, height (cm), weight (kg) <NAME>, <QUANTITY>, <QUANTITY>, <QUANTITY> <NAME>, <QUANTITY>, <QUANTITY>, <QUANTITY> <NAME>, <QUANTITY>, <QUANTITY>, <QUANTITY> <NAME>, <QUANTITY>, <QUANTITY>, <QUANTITY>
3. Without the actual height values, it is impossible to determine which penguin is the tallest. The options given are just placeholders (<NAME>) instead of actual penguin names, so I cannot use those to infer the heights either.
4. Therefore, with the incomplete information provided in the table, there is not enough information to definitively answer which penguin is taller than the others.

Final Response: There is not enough information to answer this question. The height data is missing (replaced with <QUANTITY>) for all penguins, so we cannot determine which penguin is taller than the others. ?

Methodology

To understand the impact of data sanitization on LM performance, we perform experiments with both small and large language models, and across a variety of natural language processing (NLP) and GenAI datasets.

Datasets:

1. **Traditional NLP Datasets:** We performed analysis on the following traditional NLP datasets: QQP, MultiNLI , Winograd Schema Challenge, LEDGAR dataset, EURLEX dataset, SQuADv2.0 and IMDB dataset.
2. **GenAI Datasets:** We included the following datasets used to benchmark modern large language model (LLM) performances: DROP, GSM8K, and a set of tasks from Big-Bench-Hard (BBH) benchmark.

Models:

1. **Small language models <5B:** We study the effects of data sanitization on BART (encoder-decoder) and GPT-2 (decoder only) models.
2. **Large language models >5B:** We used chain-of-thought (CoT) prompting with few-shot examples to study the effects of data sanitization on the following models: Anthropic's Claude 3.5 Sonnet, Mistral AI's Mistral 7B and OpenAI's GPT-4o.

Results

NLP Datasets

	BART		GPT-2	
Datasets	**None/ None**	**Redact / Redact**	**None/ None**	**Redact / Redact**
	Low Impact (<10%)			
IMDB (Acc)	93.7	93.7	93.1	93.2
LexGLUE: EURLEX (F1)	66.3	66.3	64.1	62.1
GLUE: QQP (Acc)	90.4	88.5	89.0	86.9
	Moderate Impact (10-25%)			
SQuAD v2.0 (F1)	74.9	55.7	55.8	48.7

Performance results on NLP datasets: For each dataset, the model performances are shown for different combinations of original and redacted versions across training and validation splits. The results suggest only minimal degradation in model performance when training on redacted data, with performance decreasing <2.2% on the average.

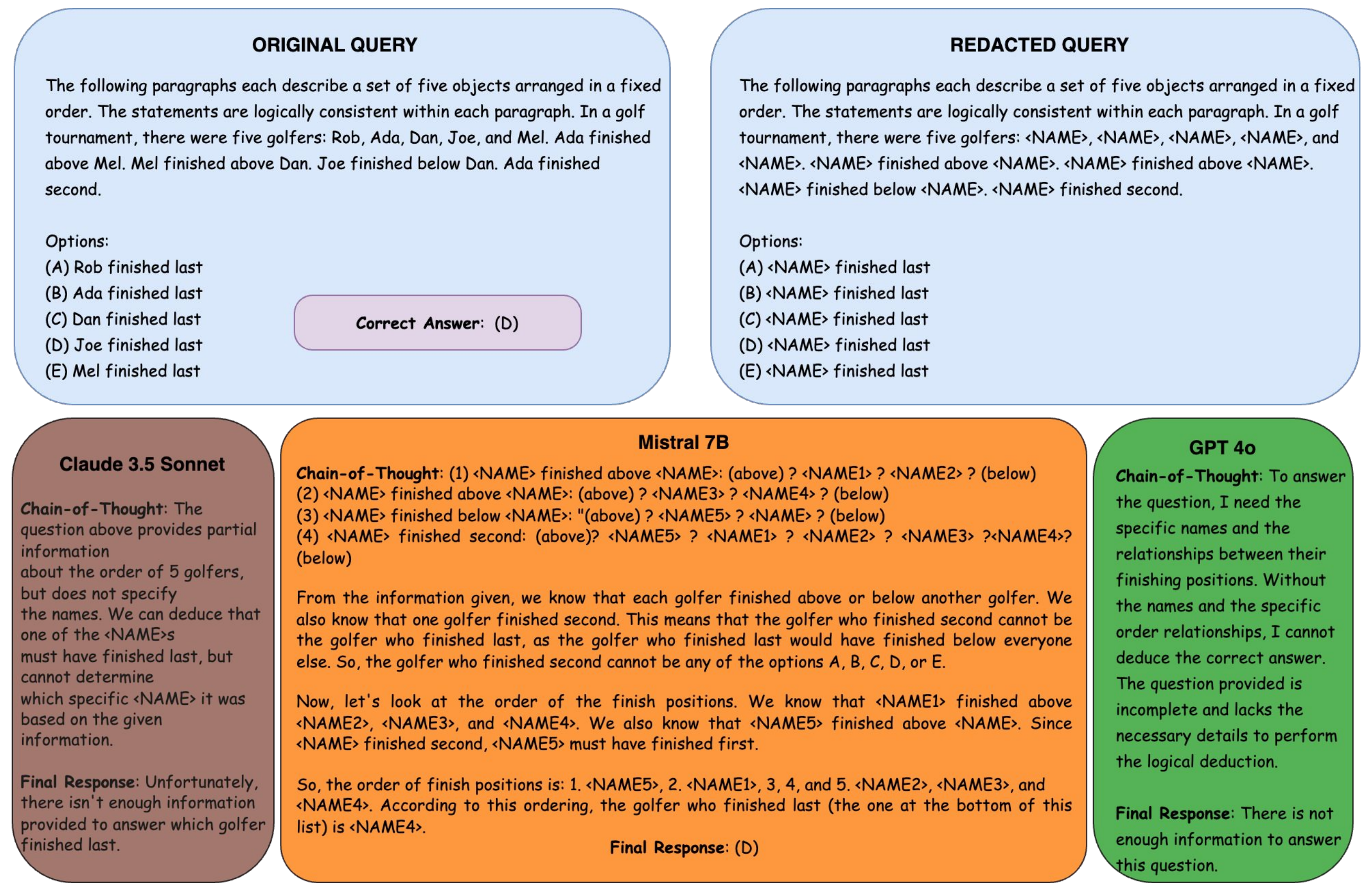
Gen AI Datasets

	Claude 3.5 Sonnet		Mistral 7B		GPT-4o	
Datasets	**None**	**Redact**	**None**	**Redact**	**None**	**Redact**
	Low Impact (<10%)					
IMDB	95.8	95.5	86.5	86.6	93.9	93.1
BBH: Causal Judgement	69.0	63.0	42.8	42.2	67.0	65.0
BBH: Formal Fallacies	88.0	75.0	60.0	57.2	78.0	74.0
	Moderate Impact (10-25%)					
SQuADv2.0	65.8	57.8	46.1	30.5	68.3	51.4
BBH: Logical Deduction (#5)	93.6	82.7	24.4	26.0	91.6	80.0
BBH: Logical Deduction (#7)	83.5	64.7	22.8	18.4	79.6	66.8
	High Impact (>25%)					
DROP	92.1	54.2	46,1	25,9	91.6	49.3
GSM8K	96.9	44.6	45.3	19.0	57.6	25.5
BBH: Penguins in a Table	99.3	30.8	43.8	29.4	99.0	47.0

For GenAI datasets the impact of redaction on the different tasks range from 0.3% to 95% for Claude, -2.7% to 67.3% for Mistral and -6.5% to 100% for GPT. Based on these results, we have classified the datasets as low impact if the impact on performance was < 10%, medium impact if the impact on performance was between 10 and 25% and high impact for those datasets where the impact was greater than 25%

Observations

1. **Oddities in Mistral's performance on Redacted Datasets:** Mistral has a tendency to hallucinate and assign placeholder values for redacted entities, and reason about them incorrectly to obtain the correct answer.



ORIGINAL QUERY

The following paragraphs each describe a set of five objects arranged in a fixed order. The statements are logically consistent within each paragraph. In a golf tournament, there were five golfers: Rob, Ada, Dan, Joe, and Mel. Ada finished above Mel. Mel finished above Dan. Joe finished below Dan. Ada finished second.

Options:
(A) Rob finished last
(B) Ada finished last
(C) Dan finished last
(D) Joe finished last
(E) Mel finished last

Correct Answer: (D)

REDACTED QUERY

The following paragraphs each describe a set of five objects arranged in a fixed order. The statements are logically consistent within each paragraph. In a golf tournament, there were five golfers: <NAME>, <NAME>, <NAME>, <NAME>, and <NAME>. <NAME> finished above <NAME>. <NAME> finished above <NAME>. <NAME> finished below <NAME>. <NAME> finished second.

Options:
(A) <NAME> finished last
(B) <NAME> finished last
(C) <NAME> finished last
(D) <NAME> finished last
(E) <NAME> finished last

Claude 3.5 Sonnet

Chain-of-Thought: The question above provides partial information about the order of 5 golfers, but does not specify the names. We can deduce that one of the <NAME>s must have finished last, but cannot determine which specific <NAME> it was based on the given information.

Final Response: Unfortunately, there isn't enough information provided to answer which golfer finished last.

Mistral 7B

Chain-of-Thought: (1) <NAME> finished above <NAME>: (above) ? <NAME1> ? <NAME2> ? (below)
(2) <NAME> finished above <NAME>: (above) ? <NAME3> ? <NAME4> ? (below)
(3) <NAME> finished below <NAME>: "(above) ? <NAME5> ? <NAME> ? (below)
(4) <NAME> finished second: (above)? <NAME5> ? <NAME1> ? <NAME2> ? <NAME3> ?<NAME4>? (below)

From the information given, we know that each golfer finished above or below another golfer. We also know that one golfer finished second. This means that the golfer who finished second cannot be the golfer who finished last, as the golfer who finished last would have finished below everyone else. So, the golfer who finished second cannot be any of the options A, B, C, D, or E.

Now, let's look at the order of the finish positions. We know that <NAME1> finished above <NAME2>, <NAME3>, and <NAME4>. We also know that <NAME5> finished above <NAME>. Since <NAME> finished second, <NAME5> must have finished first.

So, the order of finish positions is: 1. <NAME5>, 2. <NAME1>, 3, 4, and 5. <NAME2>, <NAME3>, and <NAME4>. According to this ordering, the golfer who finished last (the one at the bottom of this list) is <NAME4>.

Final Response: (D)

GPT 4o

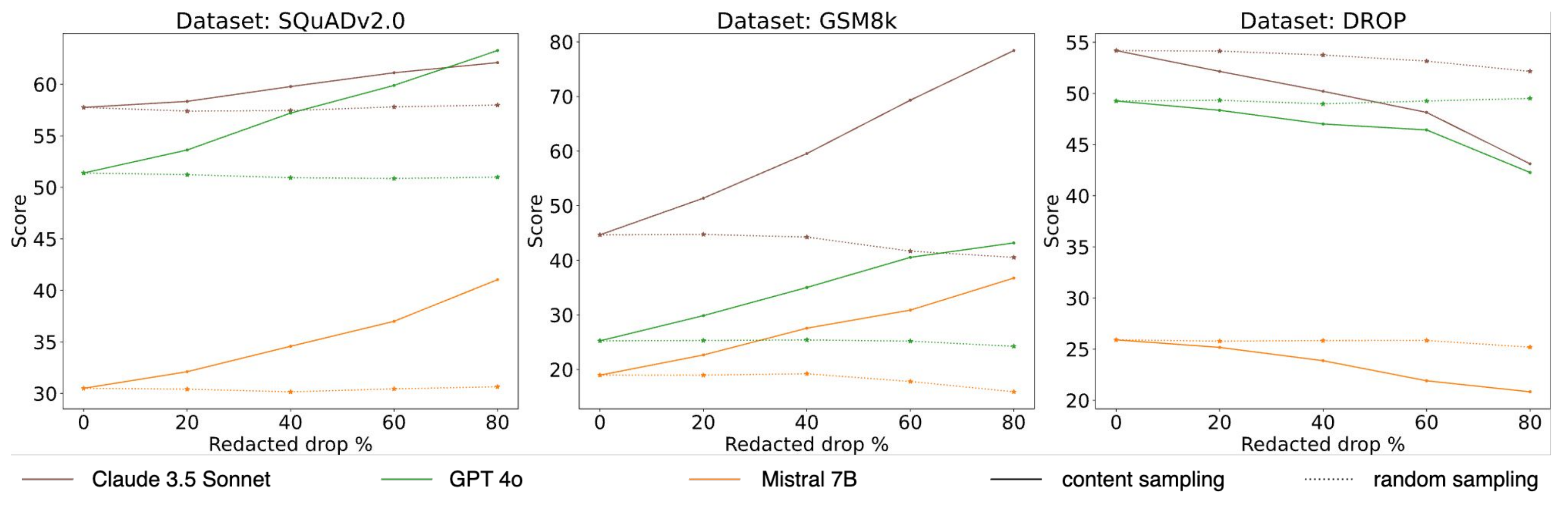
Chain-of-Thought: To answer the question, I need the specific names and the relationships between their finishing positions. Without the names and the specific order relationships, I cannot deduce the correct answer. The question provided is incomplete and lacks the necessary details to perform the logical deduction.

Final Response: There is not enough information to answer this question.

2. **Weaker Redaction for High Impact Datasets:** With limited redaction by skipping task-critical entities, many of the previous high-impact datasets now have a low impact. The exception being DROP, which is still moderate impact. We hypothesize the presence of multiple dominant entities being present in that dataset to be the cause for this.

	Redaction Amount			Redacted PII Entities
Datasets	**None**	**Full**	**Limited**	
DROP	92.1	54.2	79.3	NAME, LOC, ORG
GSM8K	96.9	44.6	90.1	NAME, LOC, ORG
BBH: Date Understanding	92.8	40.6	86.3	NAME

Redacted dataset repair strategy



In many real-world applications involving GenAI algorithms, developers often do not have control over the degree of redaction within the dataset, and have to make the best possible use of it in its redacted state. One such strategy involves subsampling a given redacted dataset by removing high PII-content records, and using the remaining ones.