SCAI 18. 8. 2025

Evaluation Challenges in the LLM Era

Patrícia Schmidtová

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics


We will use an LLM
for a brand new task
we will evaluate it somehow
we will evaluate it somehow

#1: Data Contamination

Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs

Simone Balloccu **Patrícia Schmidtová** **Mateusz Lango** **Ondřej Dušek**
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czech Republic
{balloccu,schmidtova,lango,odusek}@ufal.mff.cuni.cz

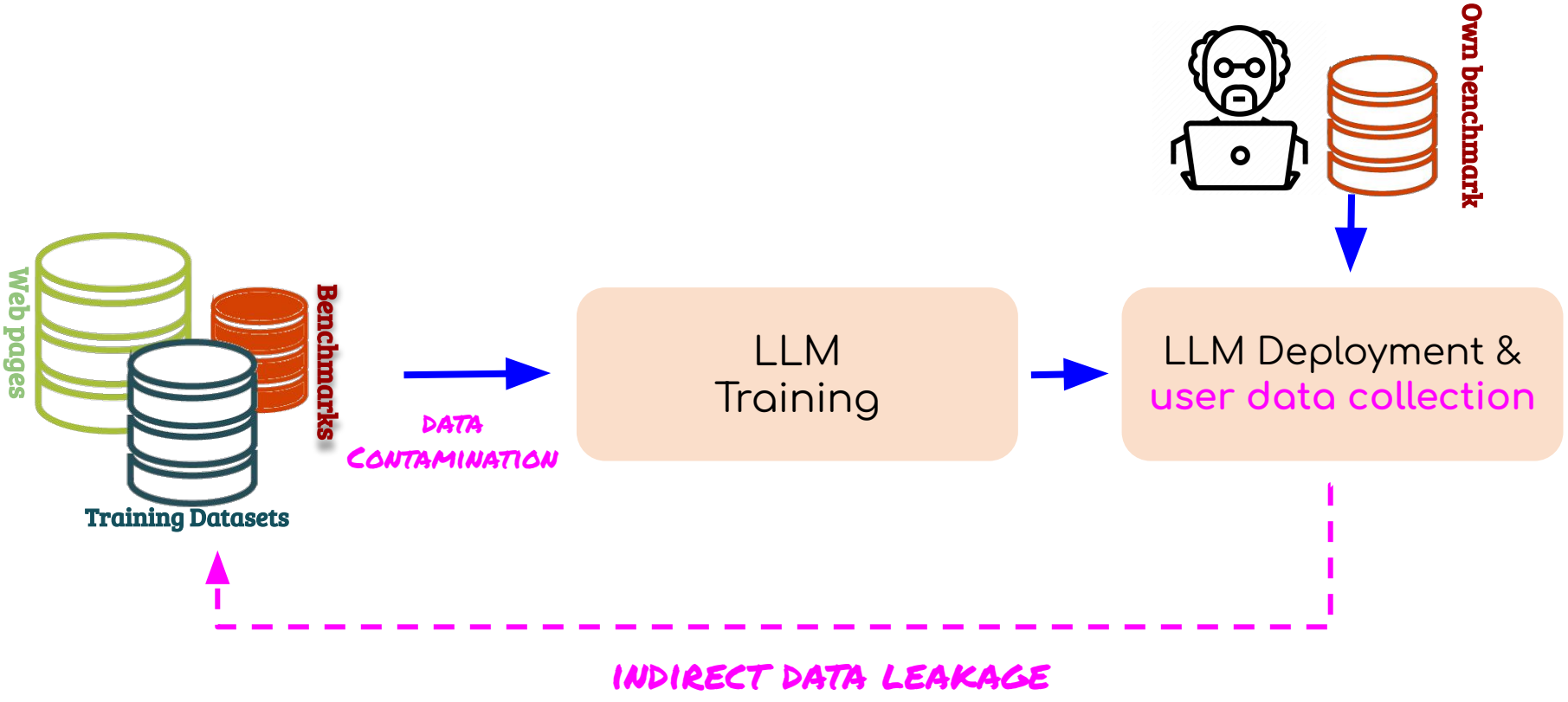
Overview

- The lack of details on training data for closed-source LLMs raised concerns on the issue of data contamination.
- Existing research overlooks when this happens indirectly - for example when models are updated from user data containing benchmarks.
- We review 255 papers causing an indirect data leak by evaluating GPT-3.5 and GPT-4 through the ChatGPT interface.
- We find that these models have been exposed to millions of samples from hundreds of NLP benchmarks.

Closed-Source LLMs & Data Contamination

- **Closed-Source**: LLMs only accessible via APIs or UIs
- For such models, researchers don't have access to:
  - Model weights
  - **Training data**
  - Other infrastructural details
- **Data contamination**: pre-training data may contain training, validation and **test sets** of NLP benchmarks

Indirect Data Leakage



Why is Indirect Data Leakage important?

1. It’s more difficult to trace due to possible subtle alterations

2. It comes with instructions included

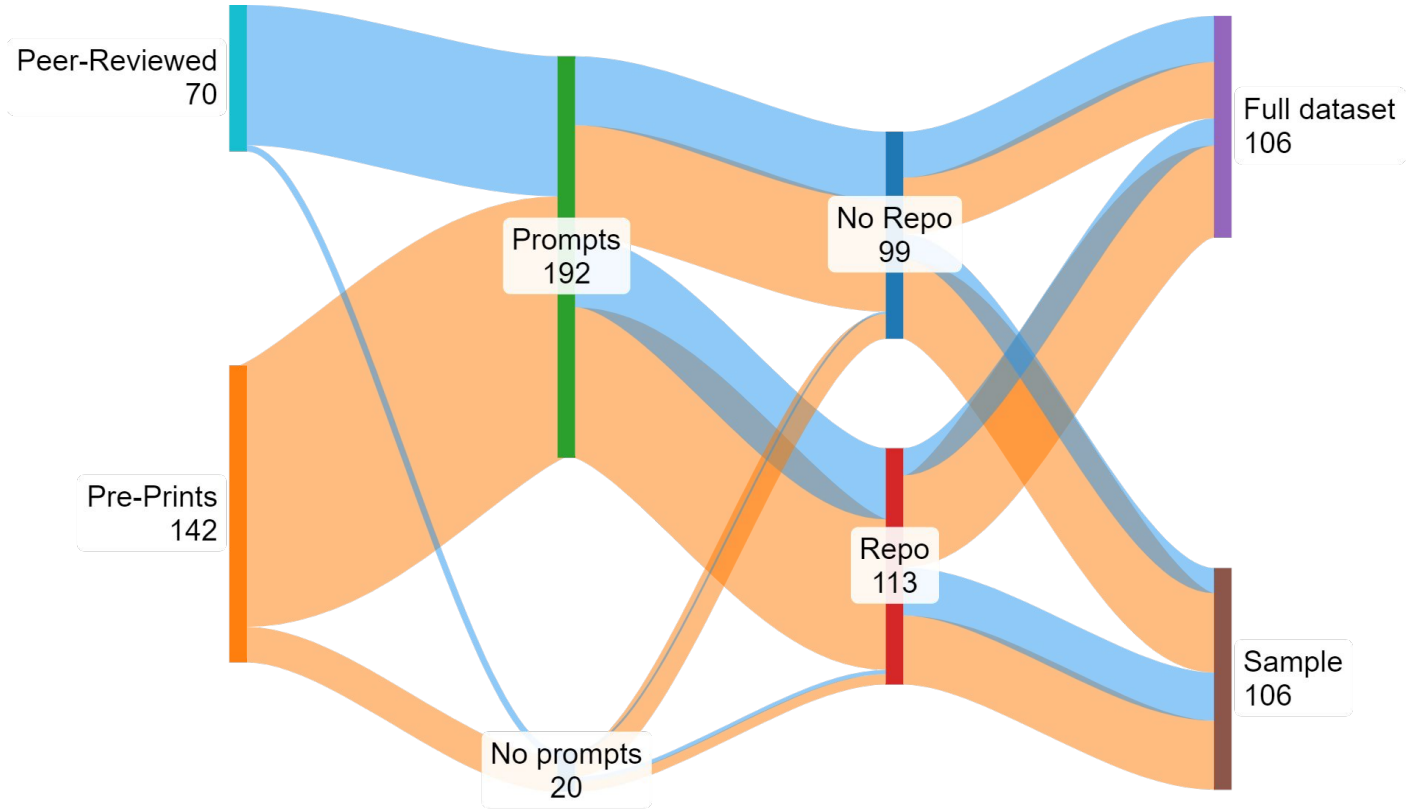
Results

We examined **255** papers, **212** of them interacted with closed-source models.

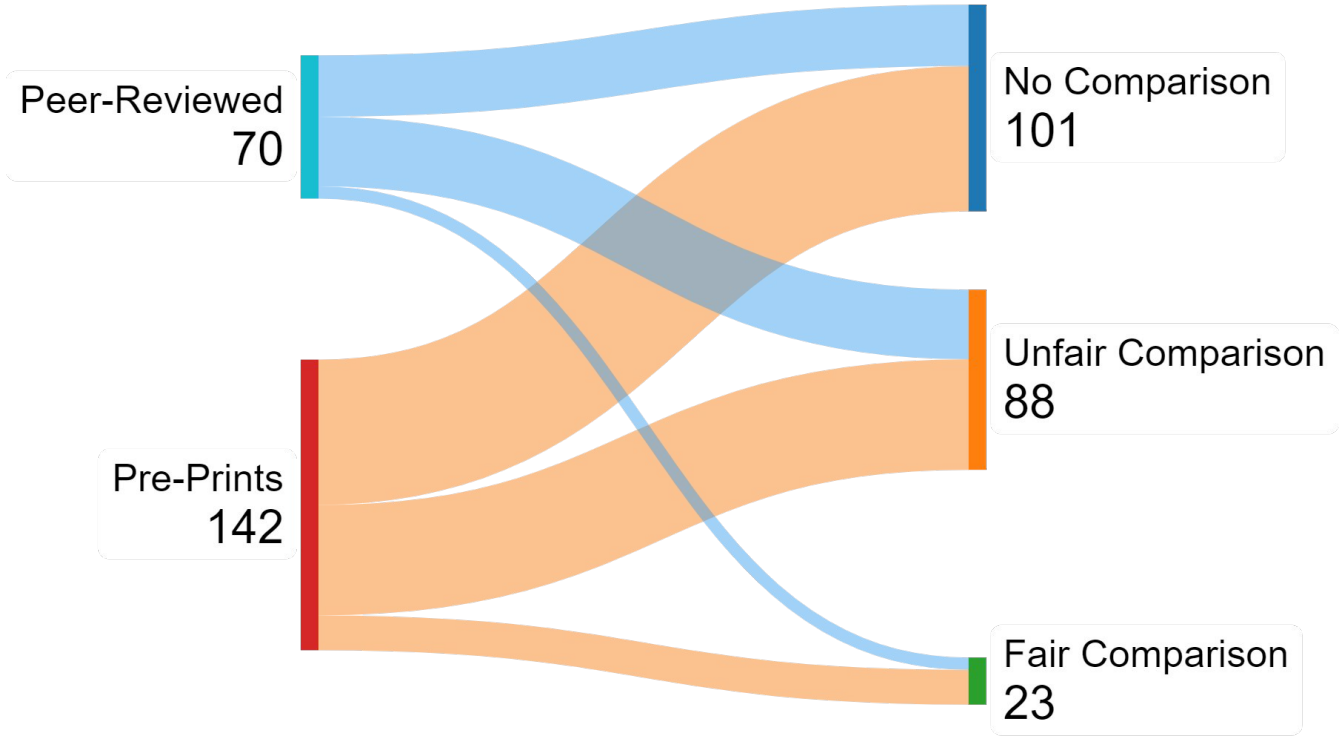
Out of these **212** papers, **90** (~**42%**) indirectly leaked data.

90 papers leaked ~**4.7M** samples form **263** NLP benchmarks.

Results – Reproducibility



Results – Fairness



Unfair comparison: comparing the performance on different samples of a dataset.

Suggested practices

- Access the model in a way that does not leak data
- Interpret performance with caution
- When possible, avoid using closed-source models
- Adopt a fair and objective comparison
- Make the evaluation reproducible
- Report indirect data leakage

#2: What Are We Even Measuring?

Automatic Metrics in Natural Language Generation: A Survey of Current Evaluation Practices

Patrícia Schmidtová[1] ✉, **Saad Mahamood**[2], **Simone Balloccu**[1], **Ondřej Dušek**[1], **Albert Gatt**[3], **Dimitra Gkatzia**[4], **David M. Howcroft**[4], **Ondřej Plátek**[1], and **Adarsa Sivaprasad**[5]

Introduction

- Automatic metrics are quick proxies, but...
  - Some have a poor correlation with human judgment
  - Many cannot capture factuality or faithfulness issues in text
  - Different implementations make results hard to interpret and reproduce
  - They can be over-reported without adding any informational value

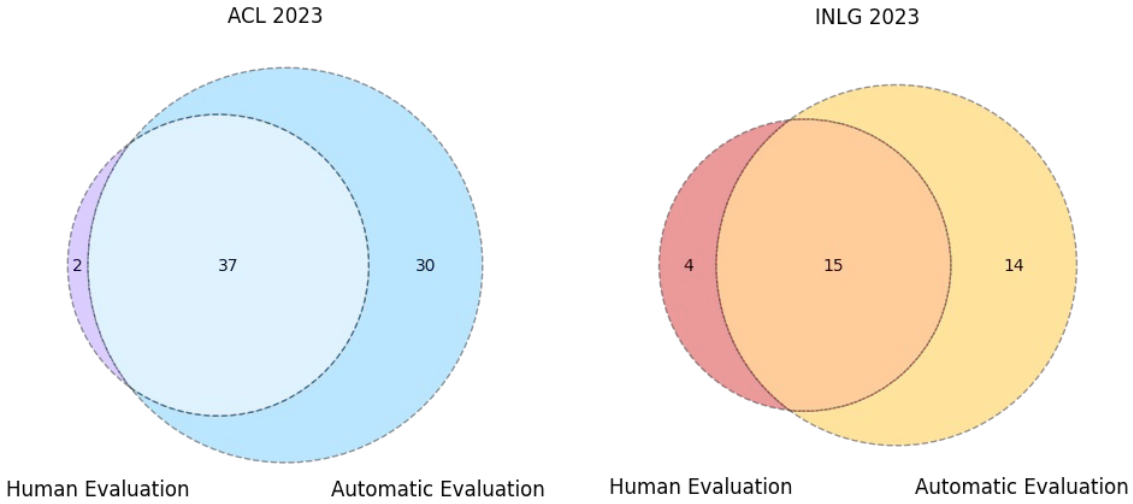
Method

We collected papers from INLG 2023 and ACL 2023 Generation track and annotated the following information:

- **Name** of the evaluation method
- Was the method **newly introduced**?
- Which **task**(s) was this metric used to evaluate?
- Did the authors comment on any **correlation between automatic** and **human evaluation**?
- Did the authors provide **implementation details** for the metric?
- Was the metric **only** reported in the **Appendix**?
- Did the authors explain the **rationale** for the metric?

Overview of Results

- **110** papers total (**36** from **INLG** and **74** from **ACL**)
- **102** papers **included** any **evaluation**
- **57%** use **human** evaluation
- **94%** use **automatic** evaluation
- **51%** use **both**
- **634 counts** of automatic metrics (**283 unique**)

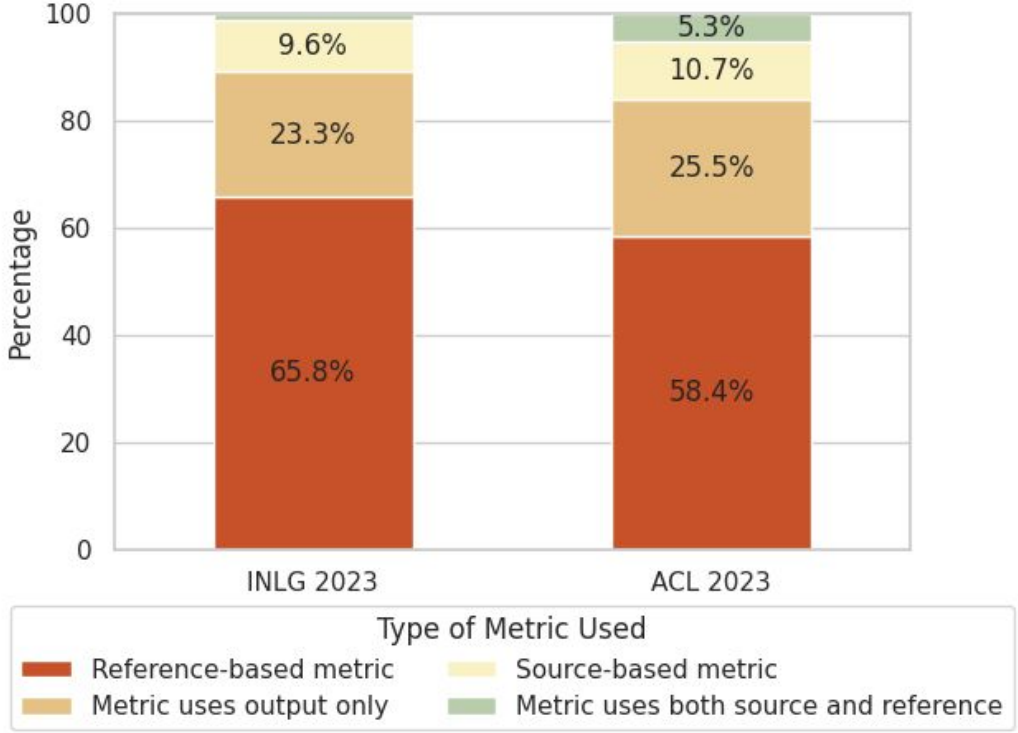


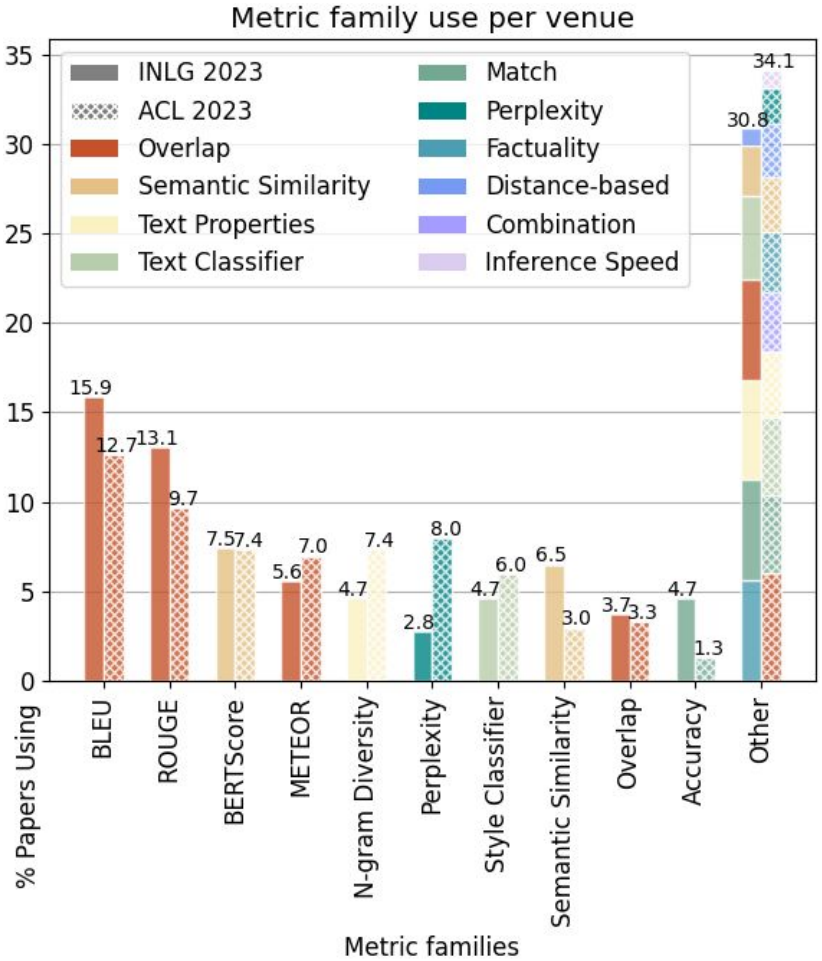
Metric Families & Categories

Metric Task Name	INLG	ACL	Total
Overlap	71	201	272
Semantic Similarity	20	59	79
Match	15	61	76
Text Properties	12	63	75
Text Classifier	17	57	74
Factuality	49	21	70
Perplexity	3	37	40
Distance-based	1	15	16
Combination	0	14	14
Inference Speed	0	4	4

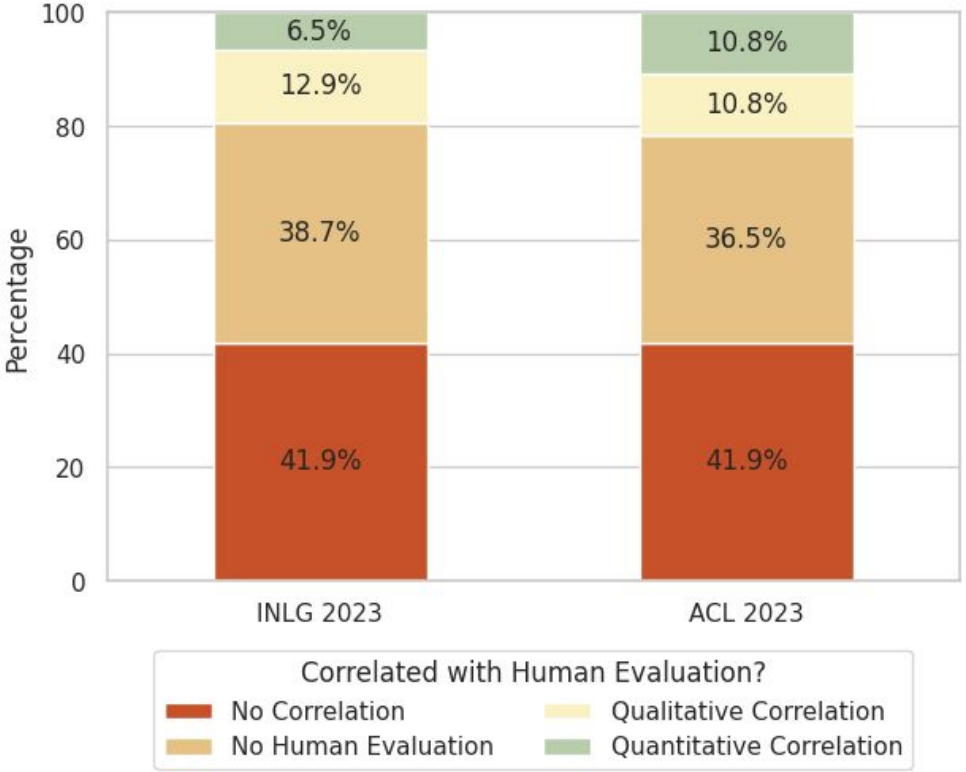
Metric Family Name	INLG	ACL	Total
BLEU	26	69	95
ROUGE	27	65	92
N-gram diversity	6	49	55
Style Classifier	5	37	42
BERTScore	8	32	40
Perplexity	3	29	32
METEOR	6	21	27
Semantic Similarity	9	12	21
Overlap	6	21	27
Factuality	5	13	18
Accuracy	8	8	16
Quality Estimation	7	7	14
Combination	0	14	14
BARTScore	2	10	12
. . .			
Recall	2	44	6
Edit Distance	1	5	6
Flesch Readability	1	3	4
Inference Speed	0	4	4
Precision	1	2	3
loss/error	0	3	3
chrF++	1	1	2

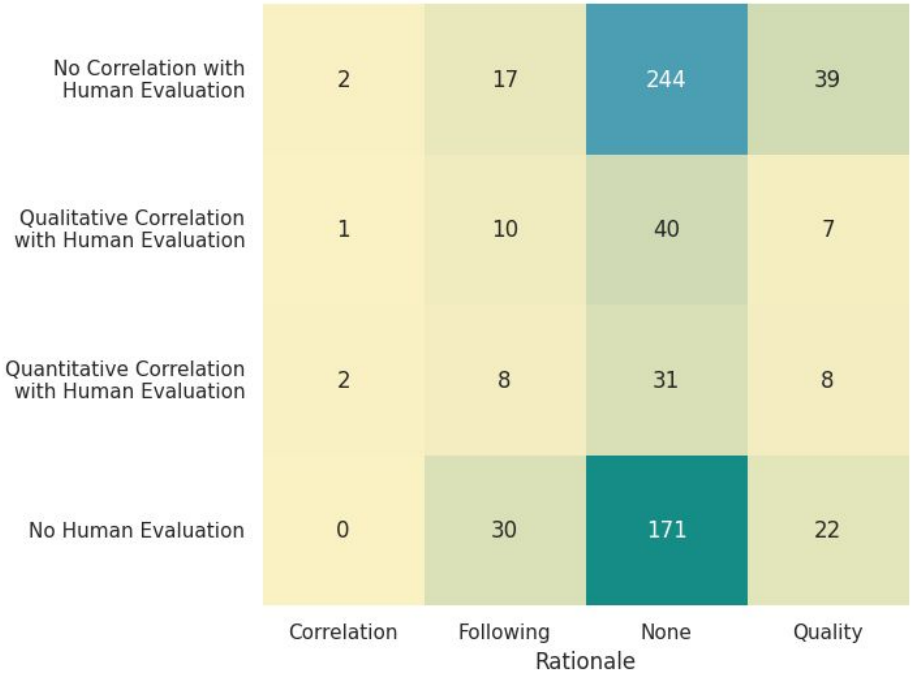
What kinds of metrics were used?



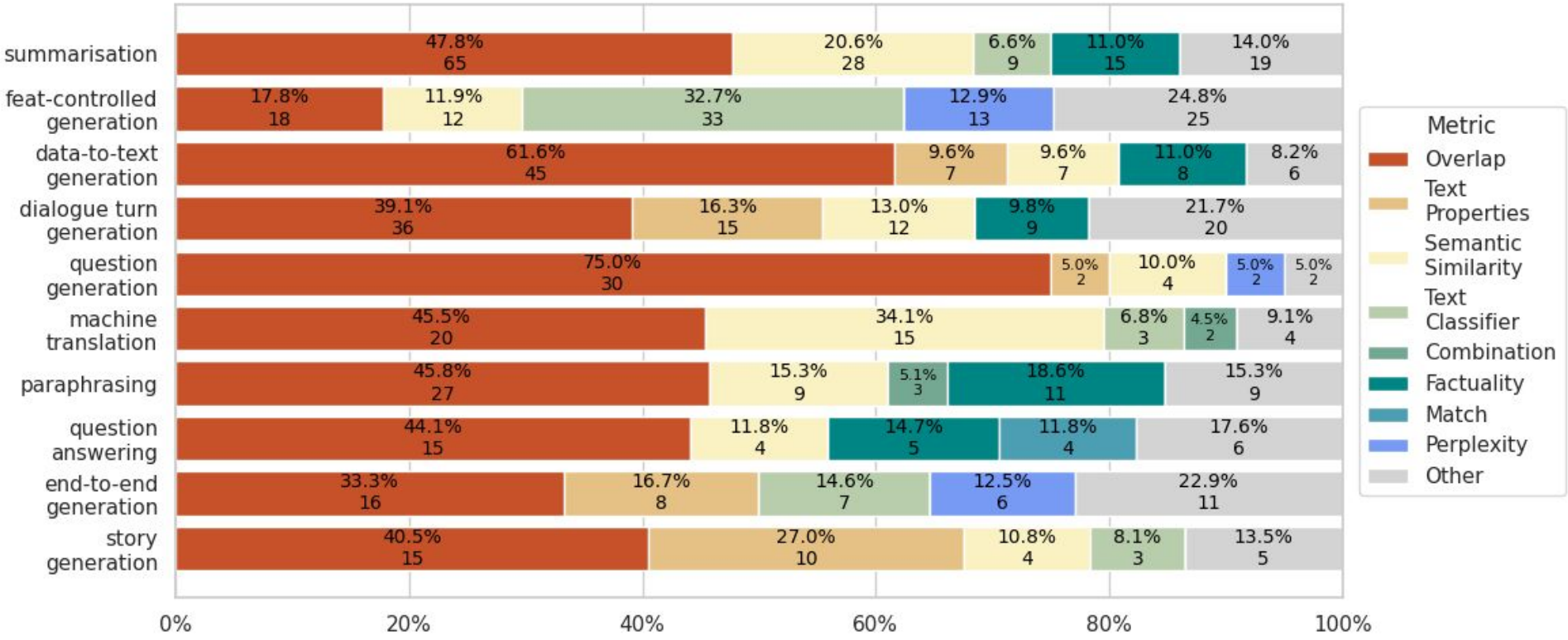


Correlation with Human Evaluation





Results per Task



Recommendations - Evaluation Quality

- Rationalize your selection of metrics
- Comment on metric combinations
- Do not copy-paste widely used metrics
- Respect the intended use of metrics
- Discuss (dis)agreements between human and automatic evaluation

Recommendations - Evaluation Reproducibility

- Share evaluation details
- Share data samples
- Release code

#3: Human Evaluation is Silver, Can We Make it Gold?

Experts or Stakeholders

Pros

- Understanding of the topic
- High quality feedback
- Don’t require instruction

Cons

- Cost of time and effort
- Low quantity of data
- Potentially biased
- Scalability

Non-Experts

Crowdworkers (Prolific or Amazon Mechanical Turk)
In-house annotators

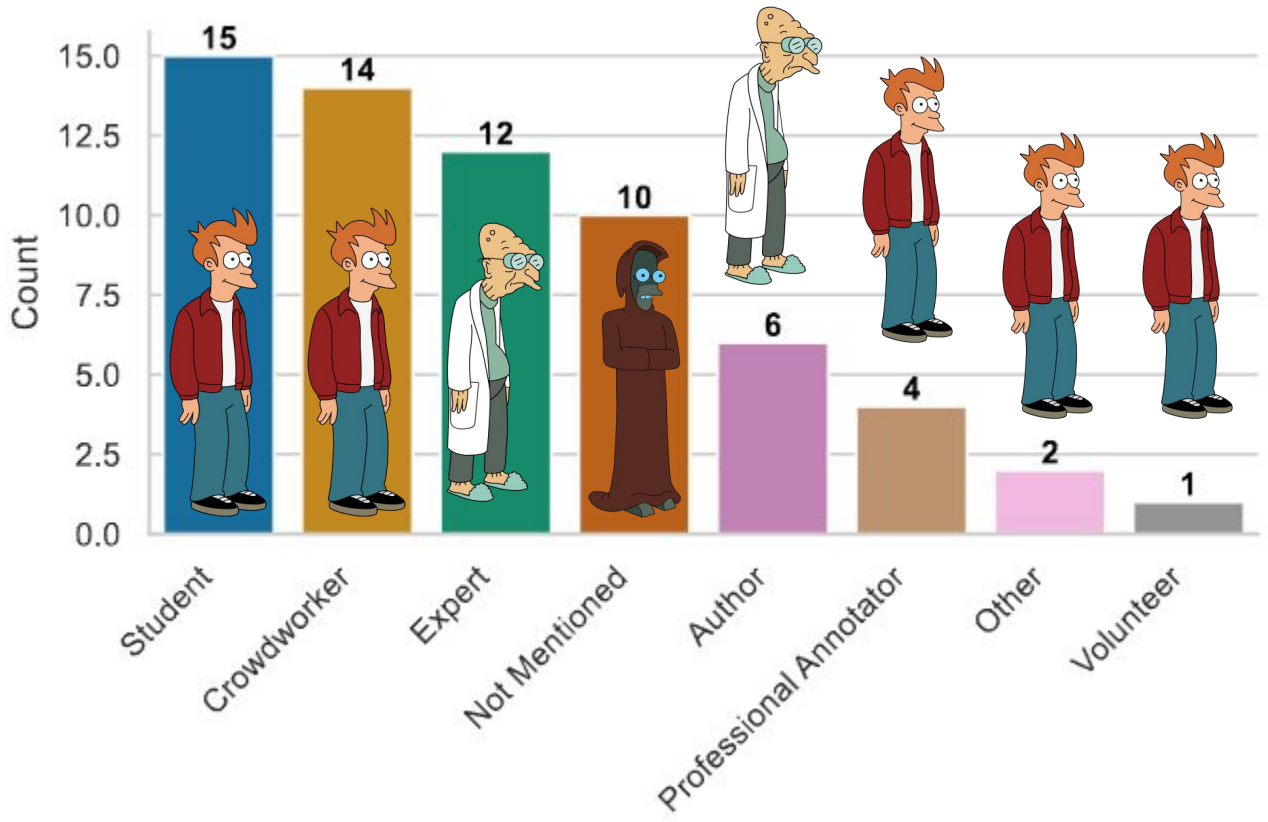
Pros

- Gold standard of evaluation
- Flexible
- High quantity of data
- Significantly cheaper than experts

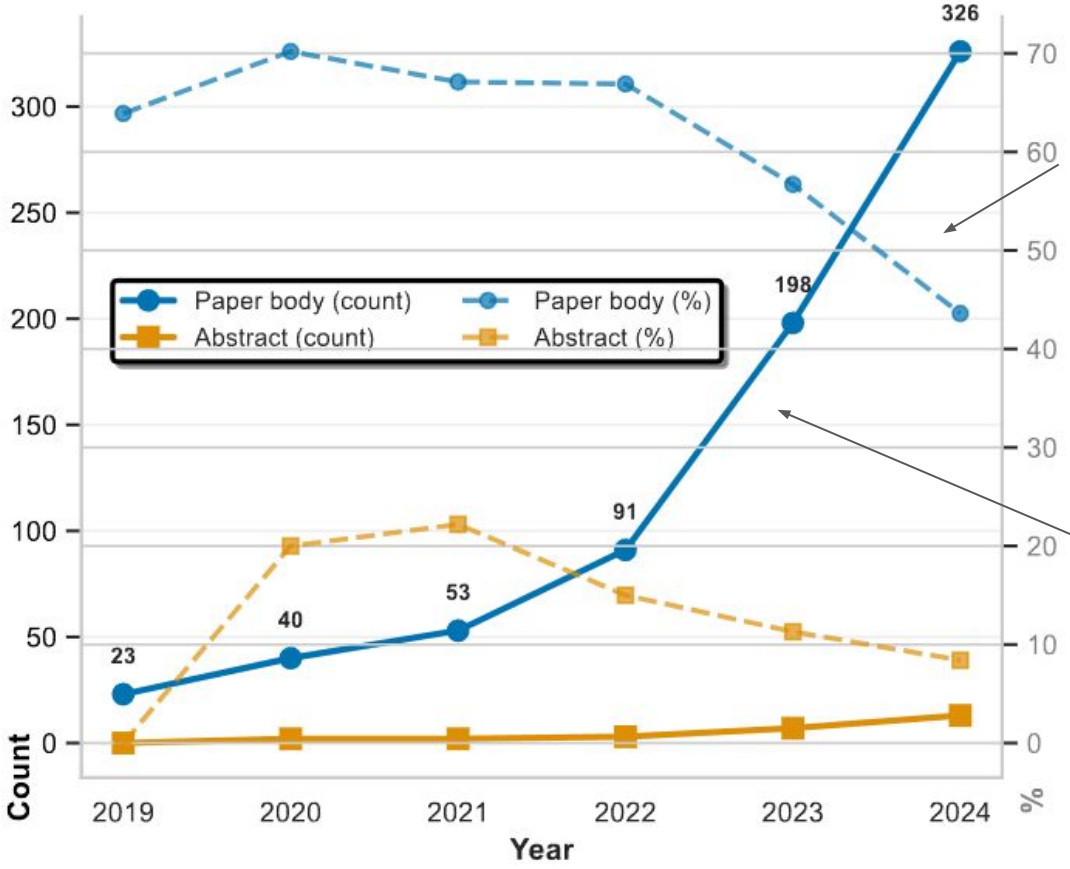
Cons

- Slow
- Require thorough instruction
- Questionable quality of data
- Still expensive

Who are the people evaluating hallucinations?



The # of human evals is growing, but their popularity is decreasing.



The percentage of papers that perform human evaluation.

The number of papers that perform human evaluation.

Ensuring Quality of Crowdsourced Human Annotation

- Pilot, pilot, and pilot!
- Use a small dataset annotated by experts as an attention check
- Use filters to pick out annotators with a high approval rate
- Carefully consider what kind of feedback you want to collect
- Mind the cognitive load of the task
- Some types of annotations are more subjective than others – you will need more data to accurately capture trends
- Share the details about the evaluation – and the data too!

#4: How (not) to do LLM as a Judge?

Large Language Models as Span Annotators

Zdeněk Kasner[1], **Vilém Zouhar**[2], **Patrícia Schmidtová**[1],
Ivan Kartáč[1], **Kristýna Onderková**[1], **Ondřej Plátek**[1],
Dimitra Gkatzia[3], **Saad Mahamood**[4], **Ondřej Dušek**[1], **Simone Balloccu**[5]

Real-World Summarization: When Evaluation Reaches Its Limits

Patrícia Schmidtová
Charles University*
{schmidtova,odusek}@ufal.mff.cuni.cz

Ondřej Dušek
Charles University

Saad Mahamood
trivago
saad.mahamood@trivago.com

Span Annotation

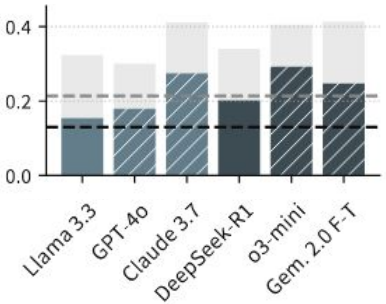
Just a 5-minute walk from Mall of the Emirates, DoubleTree offers modern accommodations. The hotel is 7.0 km from Dubai Marina and 12.1 km from Dubai Mall.

✓ *Enjoy easy access to the Mall of the Emirates.*

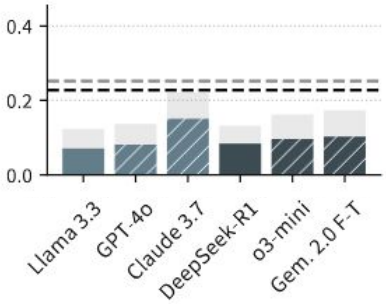
× *Enjoy breathtaking views across* ***the Hudson River to New Jersey and Liberty Island*** *from select suites.*

LLMs as judges work for tasks where the data is online.

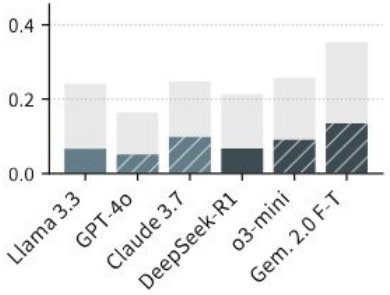
Data-to-Text



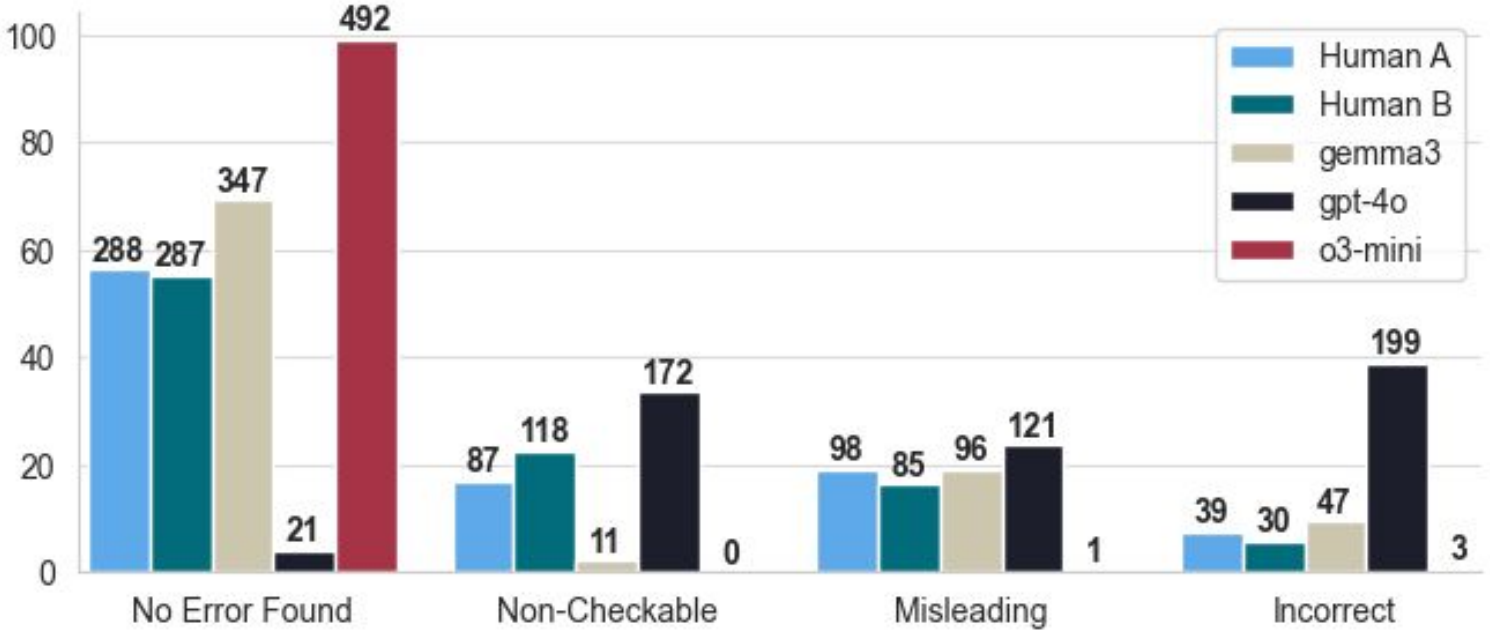
MT



Propaganda detection



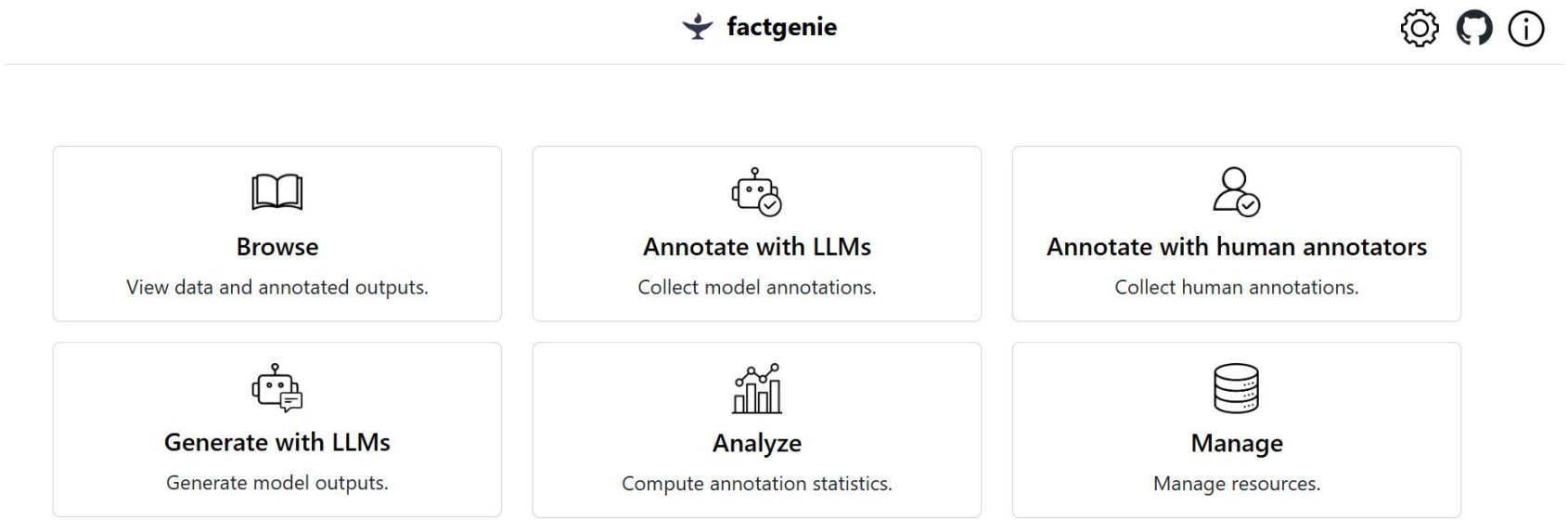
But they don’t have to work for new tasks…



If you want to use them, validate them first!

FactGenie: A Tool for Span Annotation

github.com/ufal/factgenie



Thank you!

Correspondence to: schmidtova@ufal.mff.cuni.cz

Or my LinkedIn:

This research was co-funded by the European Union (ERC, NG-NLG, 101039303) and by Charles University projects GAUK 252986 and SVV 260 698.