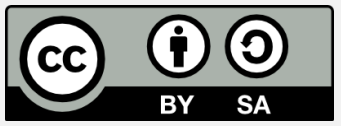
# How (Not) to Find Errors in LLM Outputs

**Ondřej Dušek**

University of Technology Nuremberg
3 June 2026

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



# LLMs need correctness evaluation

- PLMs & LLMs made (data-to-)text generation much better

| | System | Output | Rank | Score |
|---|---|---|---|---|
| | | name[Cotto], eatType[coffee shop], near[The Bakers] | | |
| (rule-based) → | TR2 | *Cotto is a coffee shop located near The Bakers.* | 1 | 100 |
| | SLUG-ALT | *Cotto is a coffee shop and is located near The Bakers* | 2 | 97 |
| (LSTM-based) → | TGEN | *Cotto is a coffee shop with a low price range. It is located near The Bakers.* | 3-4 | 85 |
| | GONG | *Cotto is a place near The Bakers.* | 3-4 | 85 |
| | SHEFF2 | *Cotto is a pub near The Bakers.* | 5 | 82 |

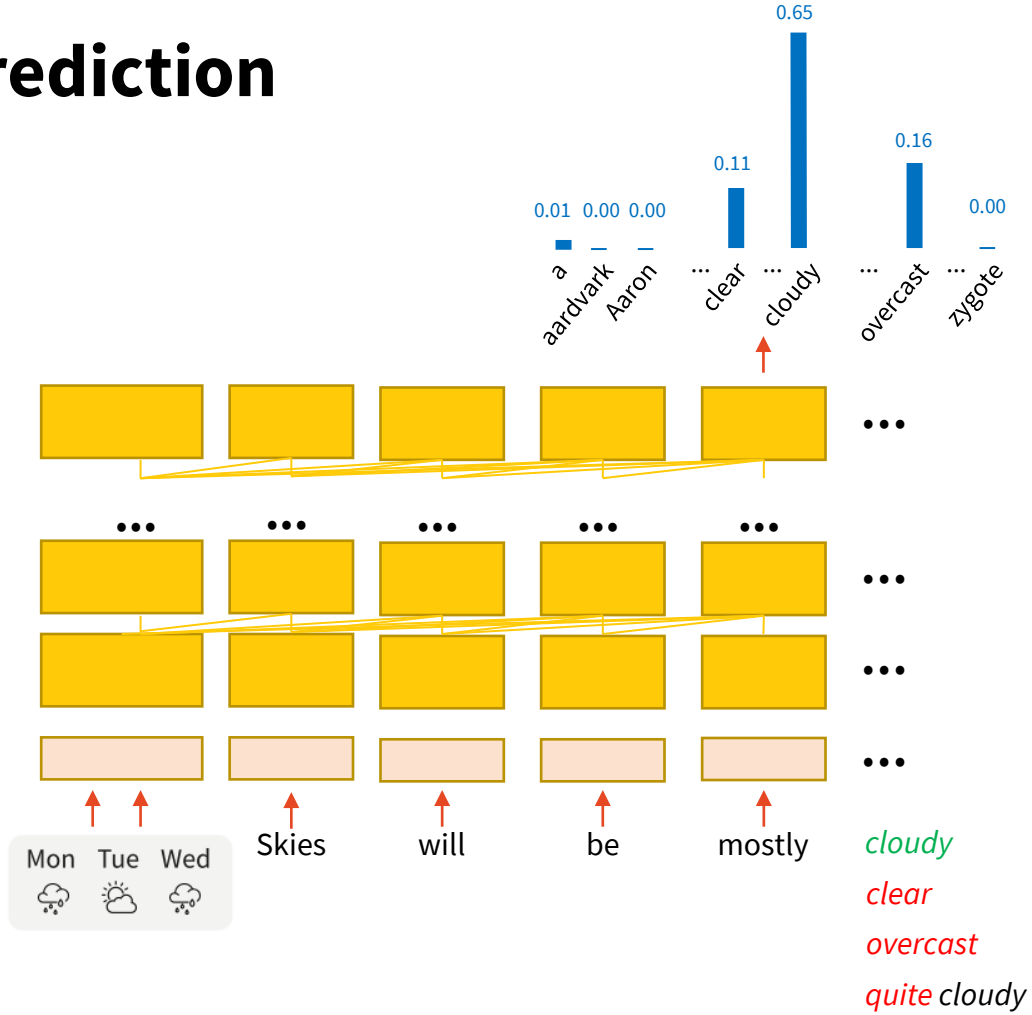
E2E Challenge (2018)

(Dušek et al., 2018)
https://aclanthology.org/W18-6539

- LLMs are now the default go-to method for everything
- They're still not perfect, though

# LLMs need correctness evaluation

- LM training / LLM pretraining: **next-word prediction**
  - just replicating training data
  - very low-level
  - no concept of sentence/text/aim/correctness
- LLM training:  still no correctness checks
  - **instruction tuning**: same as ↑, just better data
  - **RLHF**: global rewards, but only 👍 👎
    - rewards **plausible**, not necessarily correct
  - rewarded for always producing *an* answer
- → Errors are more subtle now




reference:
*Skies will be mostly cloudy, with occasional rain showers*

# How to evaluate: Metrics

(Schmidtová et al., 2024)
https://aclanthology.org/2024.inlg-main.44

- Default: **text** (+source/reference) → **score**
  - Surface overlap
  - Semantic similarity
  - Trained metrics
  - LLM scores

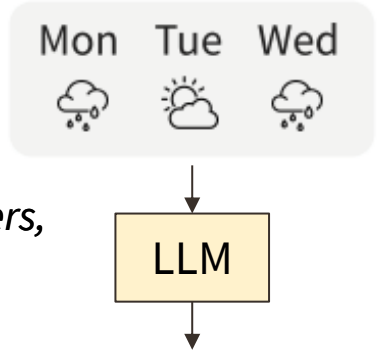
BLEU = **0.184**

BertScore = **0.916**

NUBIA Contradiction = **0.989**

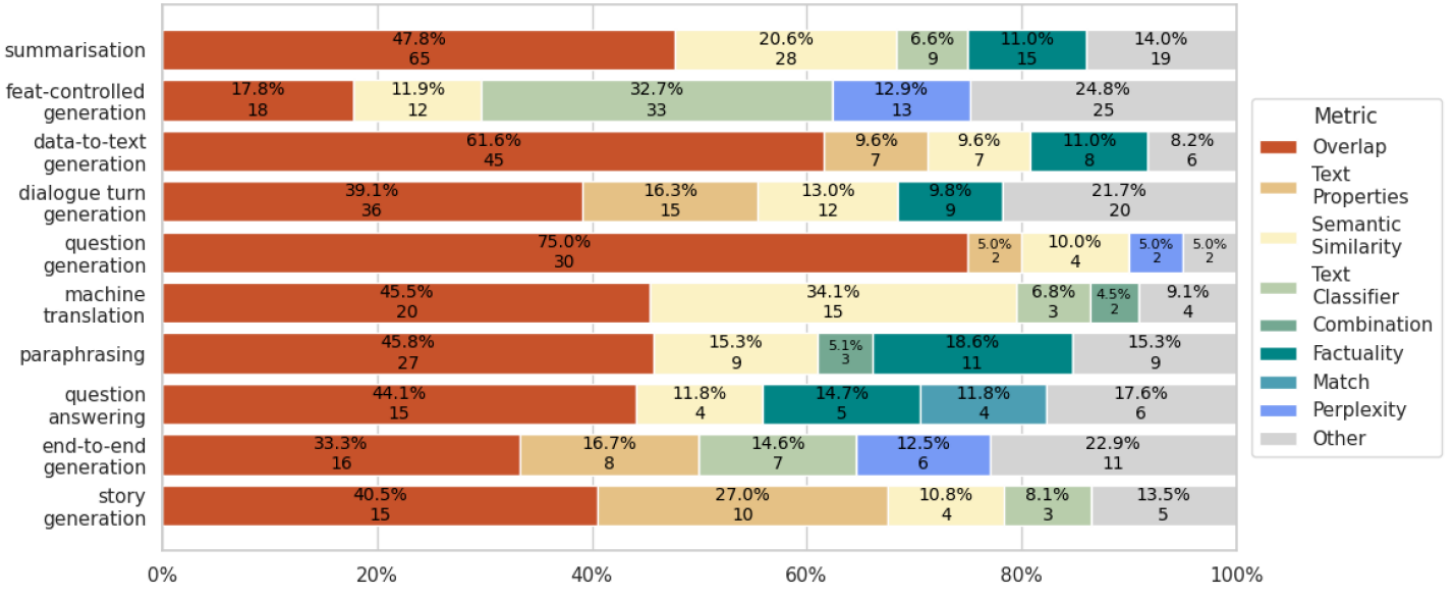
I'd rate the factual accuracy as a **3 out of 10**.

reference:
*Skies will be mostly cloudy with rain showers, but it will clear up on Tuesday.*




*Skies will be mostly clear, but winds will remain strong.*

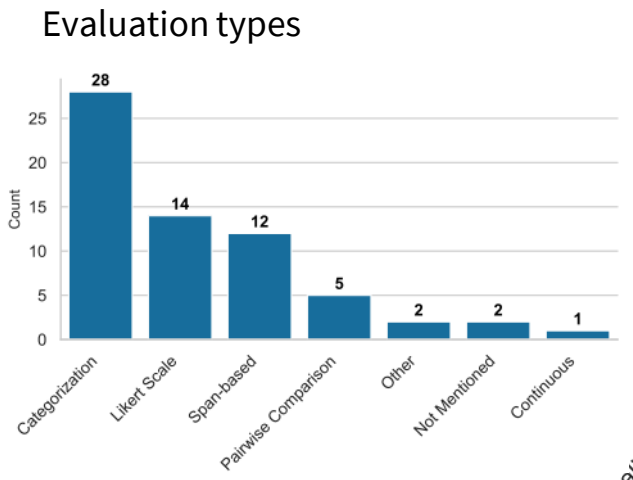
- Survey: overlap rules
  - 2023, now much more LLMs
- Often no stated reason
  - why was metric X picked?
- Good for comparison only
  - we don't know why the score is low/high

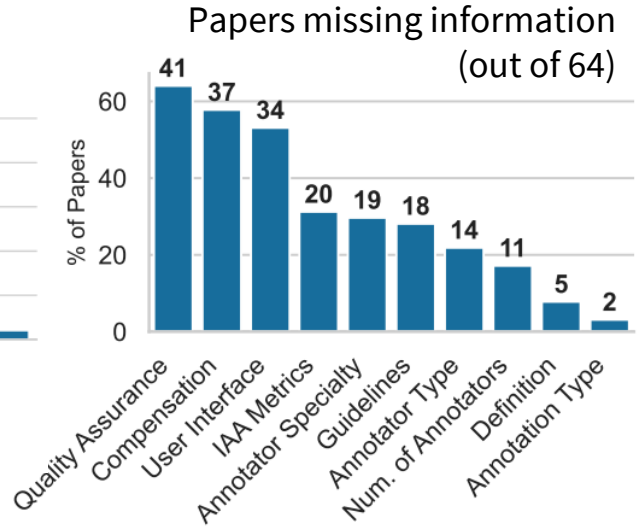


# Human Evaluation Issues

(Schmidtová et al., 2025)
https://aclanthology.org/2025.inlg-main.4/
(Schmidtová et al., 2026, to appear)

- Survey focusing on hallucination
- Mostly labels/Likert
  - some span annotation
- Often lacks detail
  - quality assurance
  - annotator compensation
  - interface
- Definitions of hallucination vary
  - + more or less precise
- Stagnation/decline in % papers
  - replaced by LLM eval? (44% LLM eval has no human check)
  - that shouldn't be the point



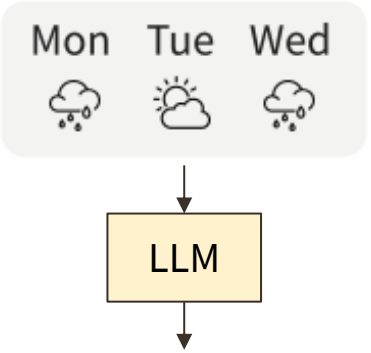

*factual accuracy, specifically looking for missing or incorrect information that could lead to errors in medical treatment after discharge*

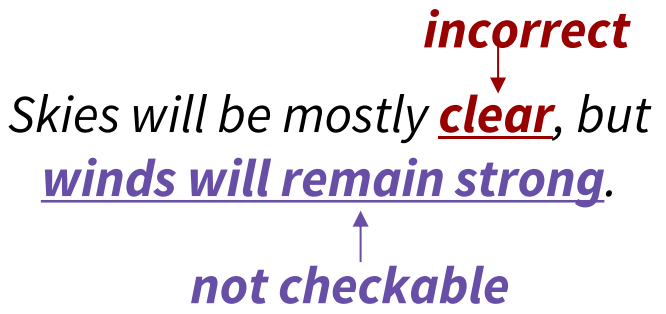
x

*Hallucinations - 0: no stuff that is not factual. - 1: even if there is one stuff that is not correct, gibberish also gets this*

# Span annotation

- **Find individual errors – mark error spans**
  - harder to annotate
  - more detailed feedback, actionable
  - allows similar LLM & human annotation
  - no references needed
- Still allows comparing models
  - use total number of errors (or some weighting)
  - ask for overall score based on the marked errors
- MT: MQM, ESA…
  - not so much in data-to-text & elsewhere



# How to annotate spans with LLMs?

(Semin et al., 2026)
https://arxiv.org/abs/2601.16946

- LLMs are decoder models – can only generate new text

**a) tagging** – regenerate with tags
  - imperfect copy: LLMs tend to "fix" stuff
  - robust but token-heavy

*There going to their house over their.*

<err>They're<err> going to their house over <err>their</err>.

**b) indexing** – refer by char/word numbers
  - LLMs suck at counting
  - adding indexes into text breaks flow

[0:5] = err  0:5 = "There"
[30:35] = err  30:35 = "r the"

*[1] There [2] going [3] to [4] their [5] house [6] over [7] their.*

**c) matching** – regenerate spans only
  - how to locate ambiguous spans?
  - otherwise similar to tagging & faster

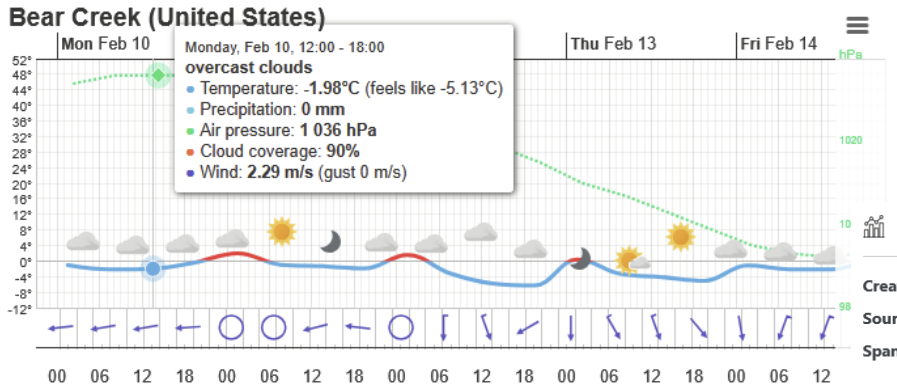
- err – their
- err – there

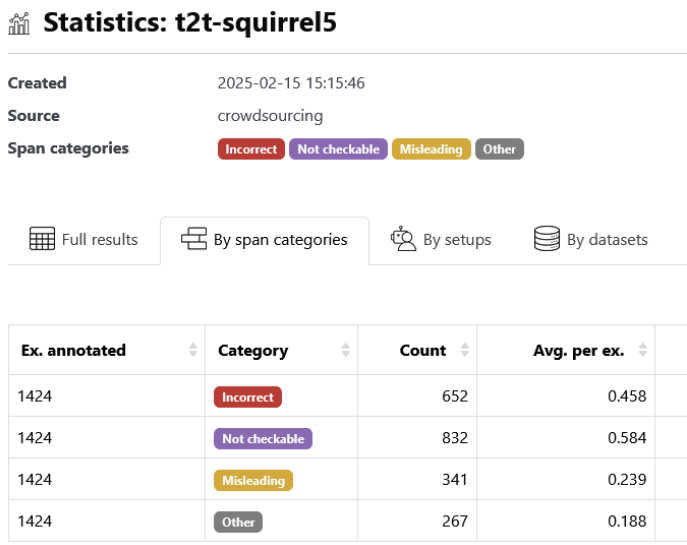
- We're generally using matching (JSON, including reasoning)

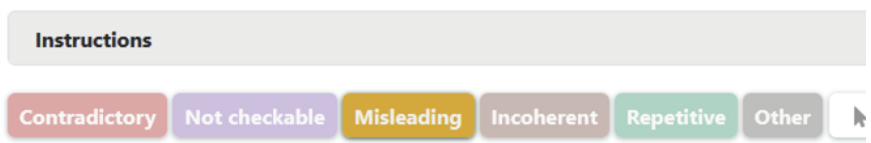
# Factgenie: simplifying evaluation

https://github.com/ufal/factgenie

- A tool for efficient span-based evaluation
  - data visualization
  - annotation
  - analysis
- Various input data formats
  - JSON, CSV, HTML
  - custom (e.g. weather)
- Humans & LLMs
  - internal campaigns
  - crowdsourcing (Prolific)
  - ollama
  - LLM APIs



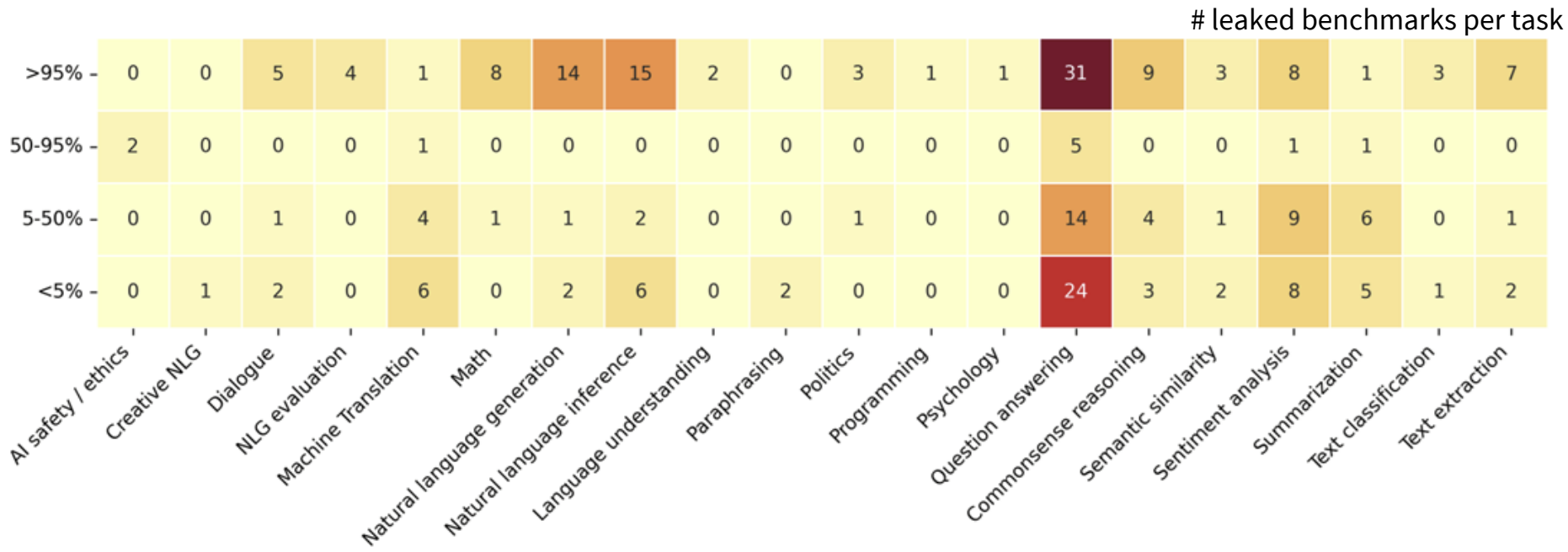

Drag your mouse over the text to highlight the span:

Over the next five days, Bear Creek can expect varying cloud cover wit periods of clear skies and overcast conditions. On Monday, the area wi experience mostly overcast clouds with temperatures ranging from -1.08 to -0.35°C. As the week progresses, there will be fluctuations in temperature, with lows reaching as cold as -10.53°C on Wednesday and highs reaching 2.02°C on Tuesday. Cloud cover will remain significant for much of Thursday and Friday, with some periods of broken clouds and clear skies. Throughout the period, wind speeds will generally be moderate, ranging from 0 to 4.63 m/s.

# What to evaluate on: Benchmarks?

(Balloccu et al., 2024)
https://aclanthology.org/2024.eacl-long.5

- Saturated: models constantly tested on them
  - MMLU, GSM8K – also, not that much generation
- Contaminated: data leaked to model training
  - not just during pretraining: **indirect leakage** to closed LLMs
  - >200 NLP benchmarks leaked to ChatGPT via web interface in the 1st year of use



# Benchmarks: Using new data

(Kasner & Dušek, 2024)
https://aclanthology.org/2024.acl-long.651

- data from different domains → open LLM → text
  - LLMs can read common data formats
  - simple prompting, 7B open models
- new data
  - from open APIs
- Evaluation via GPT-4 & crowdsourcing
  - referenceless, span annotation
- outputs are fluent, but ~80% has errors
  - >3 errors per output on average
  - depends on the domain
  - better with newer LLMs
  - worse with non-English

**Prompt**

```
Based on the given data:
```
{data}
```
Your task is to write a brief, fluent, and coherent single-paragraph {output_type} in natural language. The text should be balanced and neutral. Make sure that all the facts mentioned in the text can be derived from the input data, do *not* add any extra information.
```

**Output prefix**

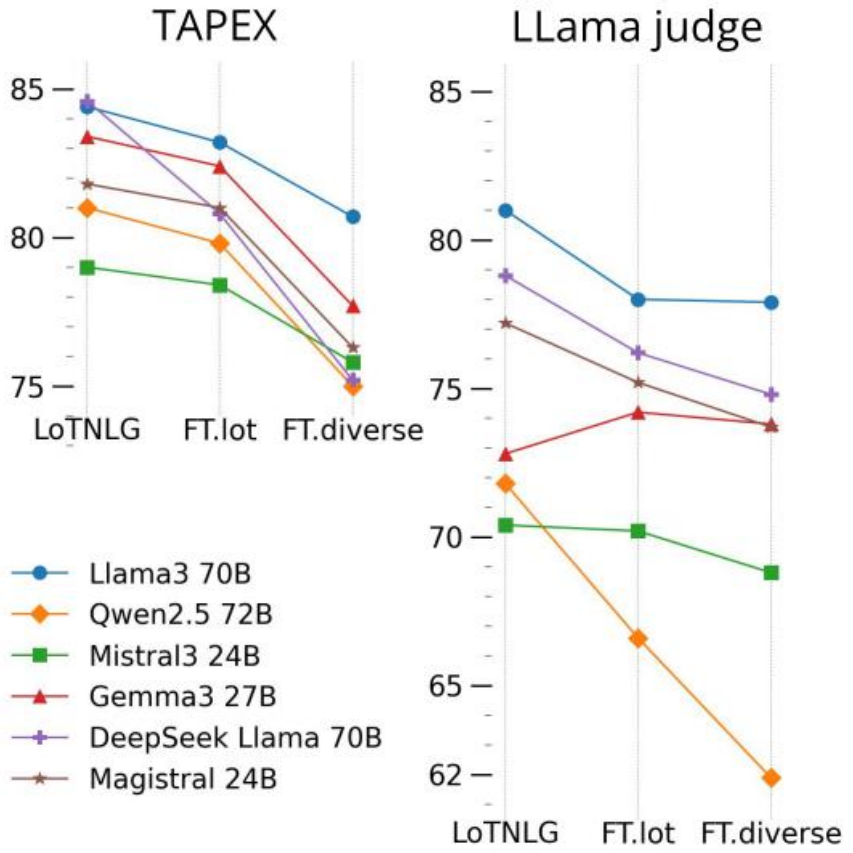
```
Sure! Here is the {output_type}:
"
```

| | Domain | Source of data | Format | Target output |
|---|---|---|---|---|
| | Weather | `openweathermap.org` | JSON | Weather forecast |
| | Technology | `gsmarena.com` | JSON | Product description |
| | Sport | `rapidapi.com` | JSON | Sports report |
| | Health | `ourworldindata.org` | CSV | Time series caption |
| | World facts | `wikidata.org` | MD | Graph description |

# Evaluating on new data vol. 2

(Onderková et al., 2025)
https://aclanthology.org/2025.inlg-main.7/

- Wikipedia tables: new pages & current events
  - multilingual
  - domain-balanced (sport/culture/politics/other)
- Evaluation via trained metrics, LLMs, humans
  - TAPEX / TAPAS: NLI-based consistency
  - human & LLM span annotation
- New data is an issue for the trained metrics
  - lower scores, not confirmed by humans
  - low correlation with humans 😳 (0.11 Pearson)
  - LLM better but not perfect: 0.53
- Fairly high rate of errors overall




Human evaluation

| | LoTNLG | FT.lot | FT.diverse |
|---|---|---|---|
| Gemma3 | 125 | 94 | 109 |
| Llama3 | 139 | 121 | 128 |
| Qwen2.5 | 150 | 127 | 133 |
| DeepSeek | 125 | 102 | 110 |

# Evaluating on new data vol. 3 – synthetic data

(Kartáč et al., 2026)
http://arxiv.org/abs/2603.20133

- Rephrase templates, make sure the answer is verifiable (easy extraction)
- Presentation matters
  - multi-turn input + assistant role make LLM perform worse

**(a) Reasoning in isolation**

Given the list of trains in JSON format below, select departure time of the latest train that arrives in Cambridge before sunset. Current date is 2025-08-27 and current time is 06:05. Sunset time is 19:57.
`Trains: [{"id": "TR5972", "departure": ...}]`

We are given: [...] The latest train that arrives in London Liverpool Street before sunset departs at 5:59pm.

Answer: **17:59** ✅

**(b) Reasoning within task-oriented dialogue**

You are a helpful assistant specialized in providing travel guidance for Cambridge. [... more instructions ...]
Current date is 2025-08-27 and current time is 06:05.
Sunset time is 19:57.

I need a train departing Cambridge and arriving at London Liverpool Street today. The earliest I can leave is 5:10pm.

`[{"function": {"name": "search_trains", ...}}]`

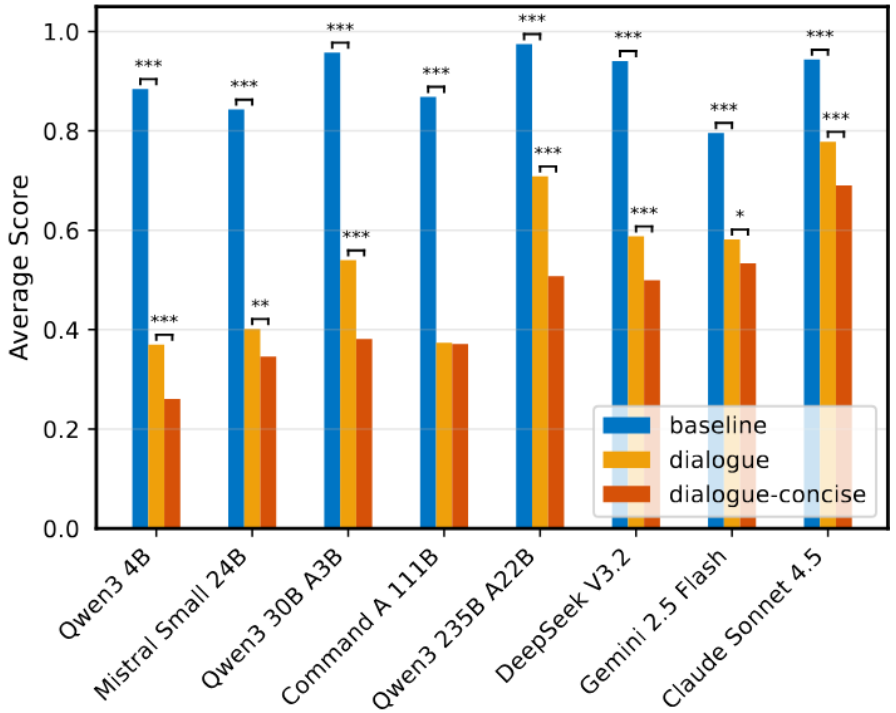
`[{"train_id": "TR5972", "departure": ...}]`

I found four trains departing today starting from 5:10pm.

What time is the last departure that will get me to London Liverpool Street before sunset?

The last train that will get you to London Liverpool Street before sunset (9:20pm) departs at 7:59pm and arrives at 9:27pm. Would you like to proceed with booking tickets for this train?

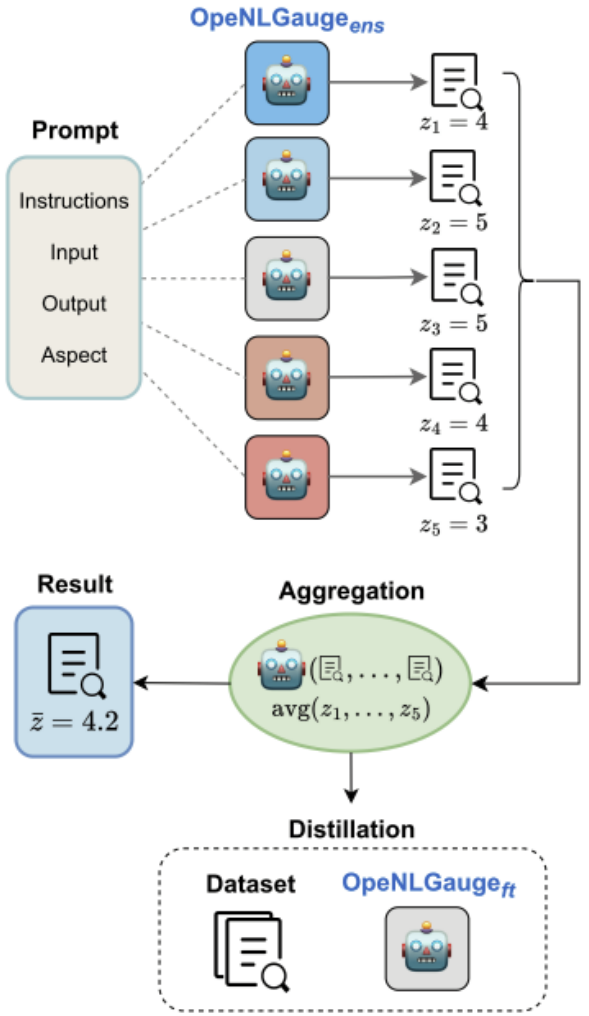
Answer: **19:59** ❌



# Span annotation with LLMs

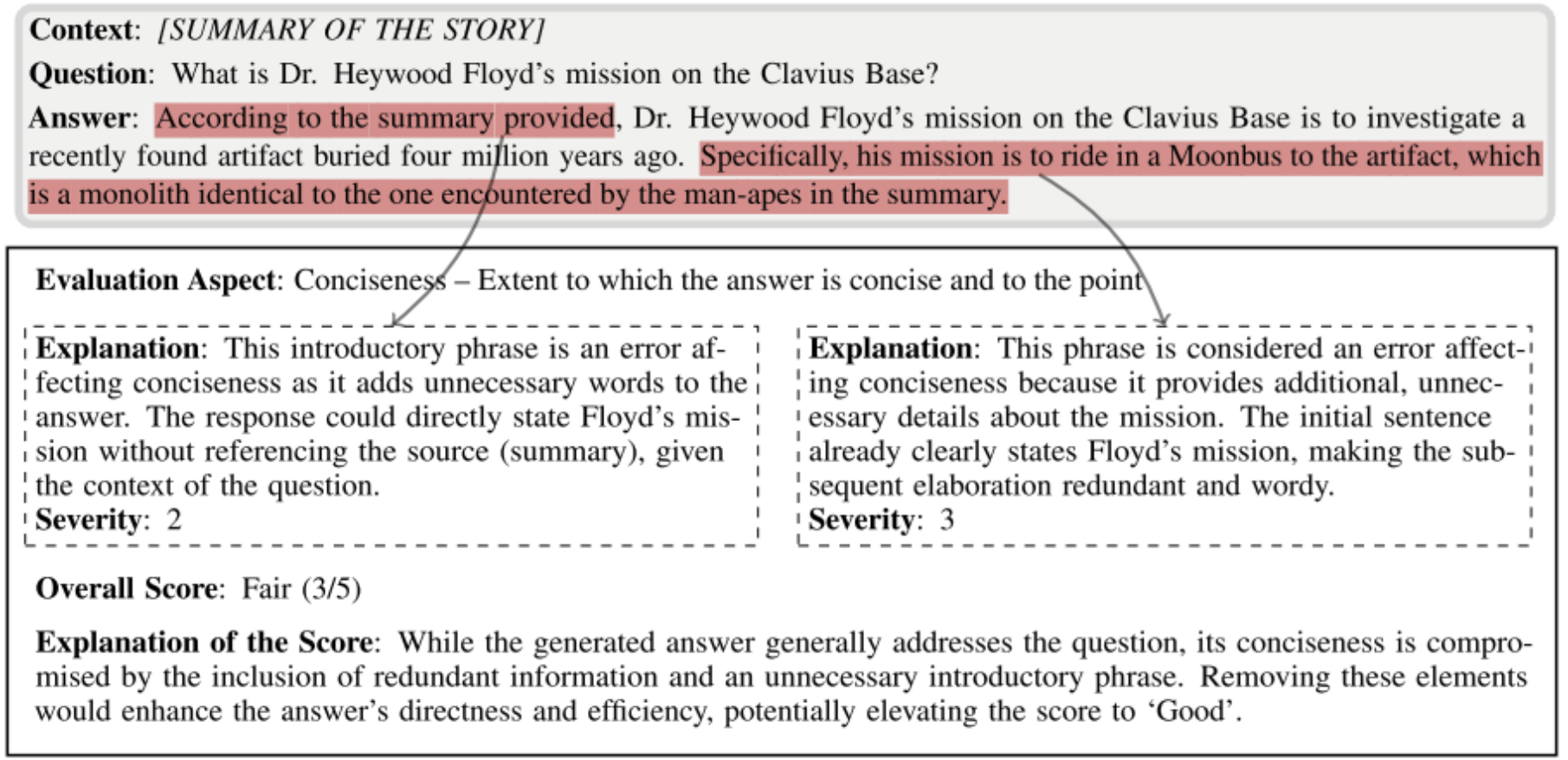
(Kartáč et al., 2025)
https://aclanthology.org/2025.inlg-main.19/

- Span annotation with LLMs taken further
  - open-source models
  - in an ensemble
  - + distilled version (via synthetic data)
  - with fine-grained explanations of each error span & overall
  - overall score grounded on error spans
- Aim: evaluate any aspect, any NLG task
  - 5 tasks, 15 datasets
    - summarization, data-to-text, dialogue, story generation, QA
  - 40+ aspects
    - correctness, informativeness, grammaticality, relevance…
  - 23 sources of generated texts to evaluate
    - LLMs, older pretrained LMs, RNNs, humans, extractive



# LLM span evaluation

- Detailed prompts, always aspect-specific
- Find spans & explain
- Provide overall score & explain



# Expanding LLM span evaluation

- Ensemble competitive with closed-LLM-based baselines
  - comparing on overall scores correlation
- Distilled model weaker but good for its size (8B)
- Performance is task-dependent (in baselines too)
  - may depend on human annotation quality
- Our approach has much better detailed outputs
  - only 1 baseline has comparable outputs
  - we mark >5x more errors, 32% → 83% correctly identified

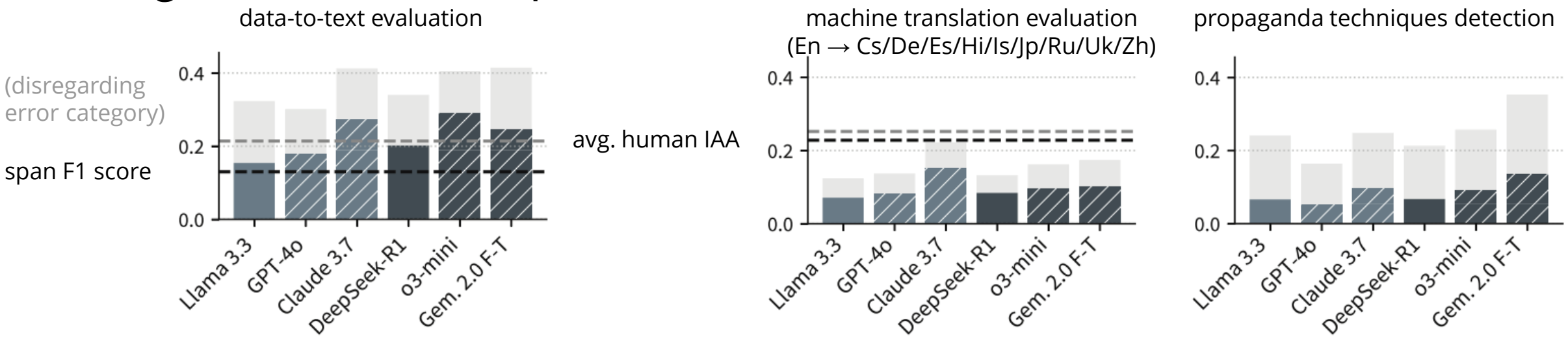
spearman correlation

| Metric | Fact | Summ | Dial | D2T |
|---|---|---|---|---|
| *ROUGE-L* | 0.156 | 0.165 | 0.244 | 0.142 |
| *BERTScore* | 0.256 | 0.231 | 0.273 | 0.172 |
| *UniEval* | 0.575 | 0.474 | 0.417 | 0.282 |
| G-Eval (GPT-4) | 0.611 | 0.523 | 0.588 | 0.269 |
| LLM eval. (GPT-4) | 0.637 | 0.511 | 0.746 | 0.320 |
| Auto-J | 0.226 | 0.198 | 0.425 | 0.141 |
| TigerScore | 0.504 | 0.384 | 0.346 | 0.200 |
| InstructScore | 0.072 | 0.258 | 0.241 | 0.247 |
| Themis | 0.684 | 0.553 | 0.725 | 0.333 |
| **OpeNLGauge_ens** | 0.689 | 0.534 | 0.653 | 0.299 |
| Command R+ | 0.608 | 0.395 | 0.374 | 0.198 |
| Gemma | 0.630 | 0.442 | 0.484 | 0.295 |
| Nemotron | 0.659 | 0.451 | 0.645 | 0.242 |
| Mistral | 0.637 | 0.502 | 0.596 | 0.254 |
| Qwen | 0.644 | 0.478 | 0.514 | 0.283 |
| Llama 3.1 8B | 0.236 | 0.186 | 0.309 | 0.108 |
| **OpeNLGauge_ft** | 0.651 | 0.502 | 0.578 | 0.315 |

# LLMs vs humans on error span annotation

(Kasner et al., 2026)
https://aclanthology.org/2026.mme-main.1/

- Comparing LLMs vs good humans (carefully selected crowd workers or experts)
- Average human IAA compared to model results

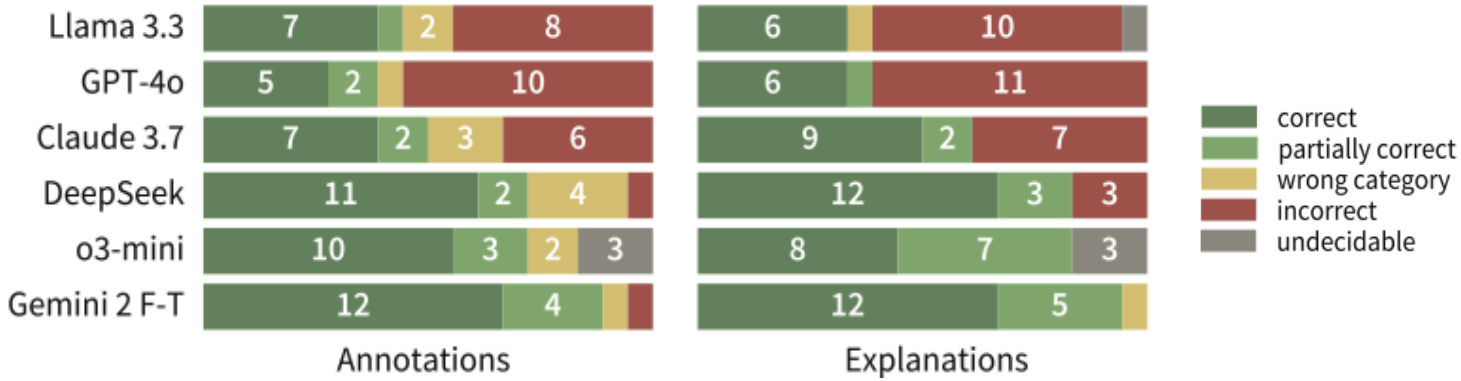



- Overall low agreements
  - room for subjectivity in span selection & error category assignment
- Performance is task-dependent
  - LLMs surpass average human agreement on data-to-text evaluation
  - reasoning models are mostly better

# LLMs vs humans on error span annotation

- Detailed analysis of error spans

data-to-text




- Models tend to nitpick
- Error spans ~ 50% correct, but so are human-annotated ones
  - the task is hard for humans too
- LLMs beat humans on cost & flexibility
  - good alternative to humans
  - task-dependent: performance needs to be validated first

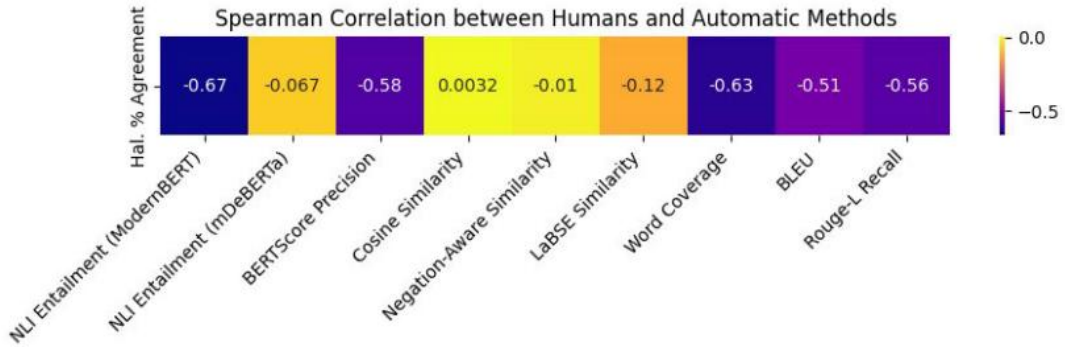
# LLM Metrics Caveats – Domains/Tasks

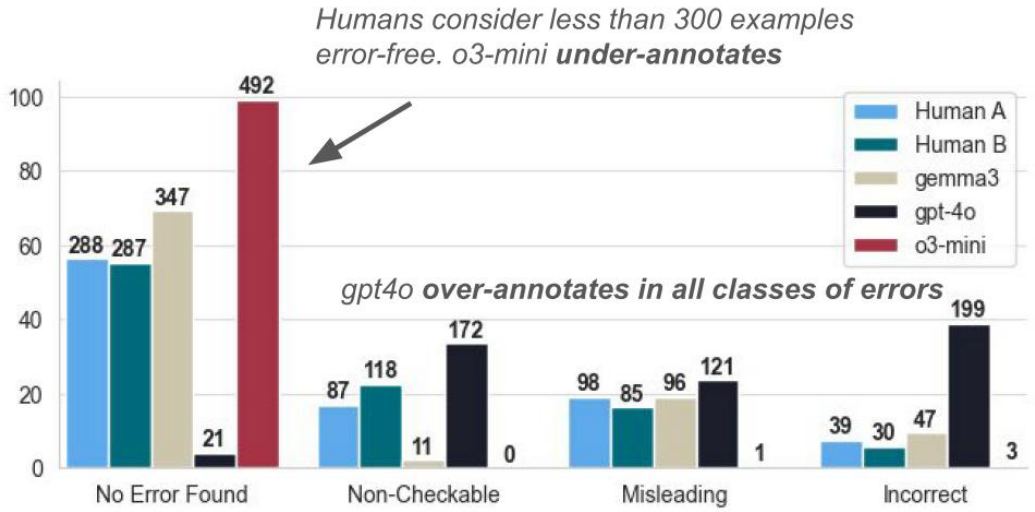
(Schmidtová et al., 2025)
https://aclanthology.org/2025.findings-emnlp.1363/

- Unconventional task: **highlight generation**
  - pick attractive points from long texts
  - hotel domain
- Similarity metrics correlate poorly with humans
- LLMs perform very inconsistently

Description: *Just a 5-minute walk from Mall of the Emirates, DoubleTree by Hilton Hotel and Residences Dubai offers modern accommodations. [...] The hotel is 7.0 km from Dubai Marina and 12.1 km from Dubai Mall. Dubai International Airport is 30 minutes away by car.*

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



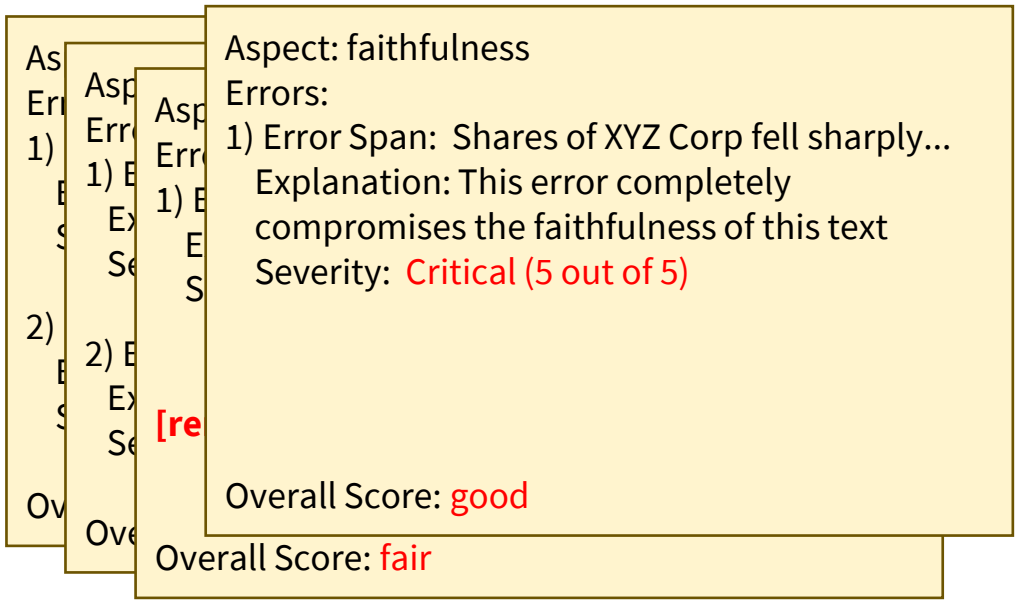
# LLM Metric Caveats – Consistency

(Terentowicz et al., 2025)
https://aclanthology.org/2025.inlg-main.37/

- Are model-assigned scores consistent with error spans?
- What happens to scores when we tamper with the analysis?
  - summarization (SummEval), story generation (HANNA), question answering (QAGS)

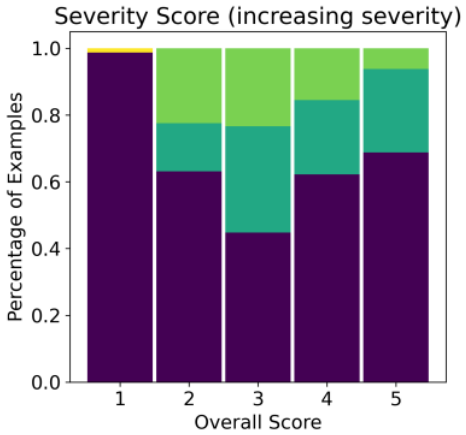
Input: *On the stock market today [...] XYZ Corp's stock is up 1.3%, surpassing competing ABC Inc., whose stock is down 2.5% […]*

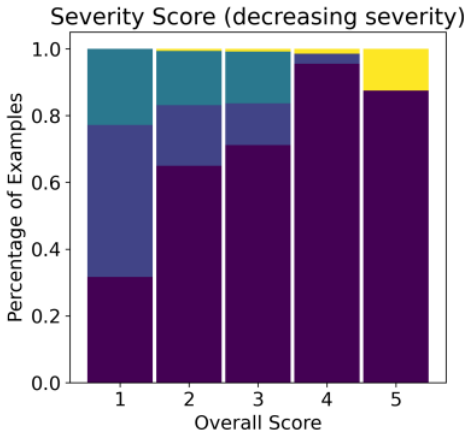
Summary: *Shares of XYZ Corp fell sharply today, trailing behind ABC Inc., whose profits soared by 40% this quarter.*

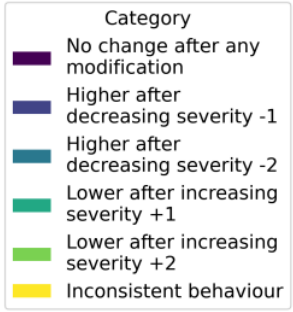


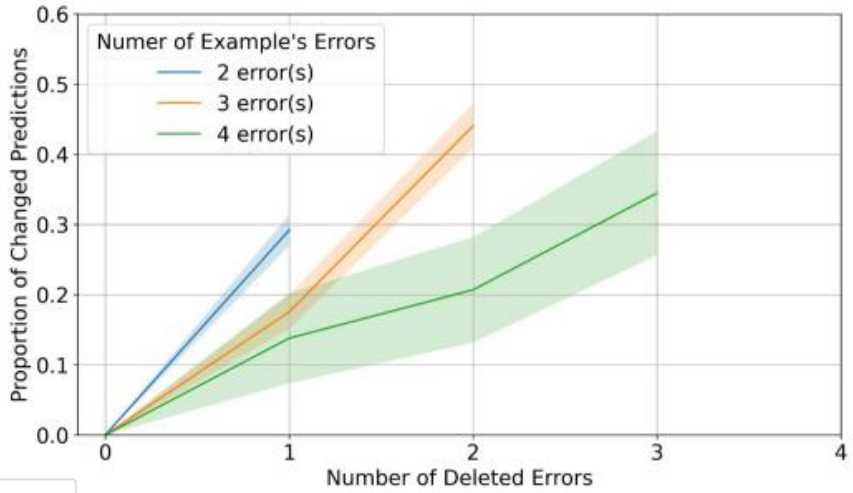
# LLM Metric Caveats – Consistency

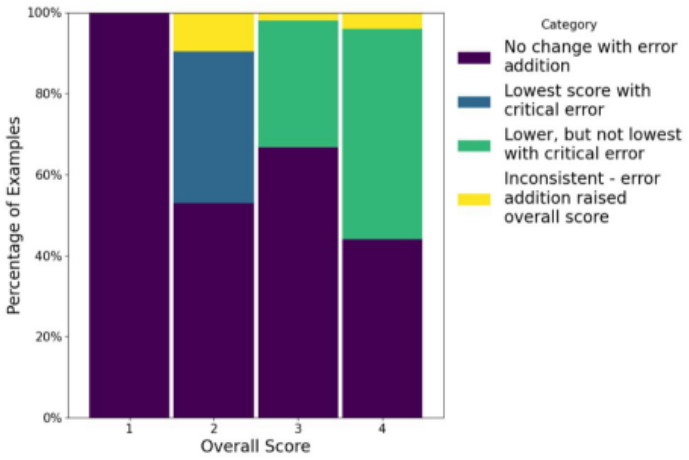
- Perturbations often have no effect
  - removing errors:
    18% scores change
  - decreasing severities:
    12% scores change
- Inconsistent
  - removing errors
    lowers score
    & vice-versa



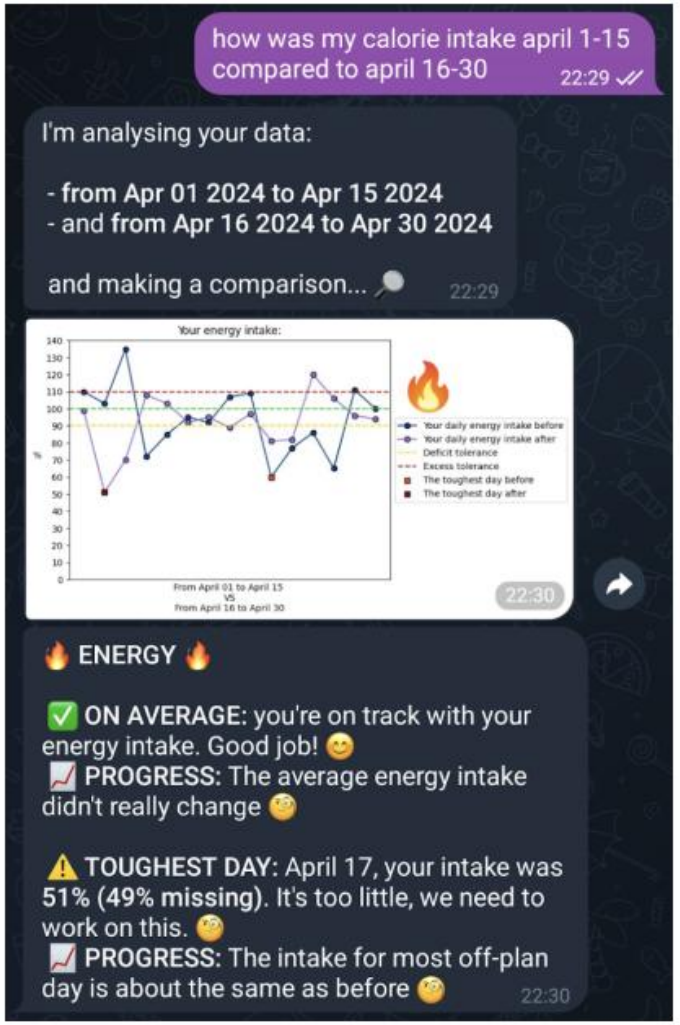



# General Caveats – Extrinsic Evaluation

(Li et al., 2025)
https://aclanthology.org/2025.inlg-main.44/

- Nutrition counselling chatbot
  - Rule-based in the baseline form
- LLM-based answer rephrasing
- LLM-based guided counselling (finetuned)
  - strict structure
  - data vetted by experts
- Intrinsic eval (metrics + manual) positive
- 7-week controlled trial (82 people across 3 settings)
  - No difference w.r.t. baseline



# Wrap up

- Use **new data**
  - without human references
  - real or synthetic (to check specific questions)
- Use **error span** annotation
  - harder & more subjective than scores, but more actionable
  - allows both comparisons & analysis
- Use **LLMs**, but be cautious
  - proprietary & reasoning models are better
  - ensemble of open models works well too
  - performance may depend on the task
  - consistency may not be ideal
- Verify performance with **humans**
  - ideally in an **extrinsic** evaluation

# Thanks

## Contacts

Ondřej Dušek
odusek@ufal.mff.cuni.cz
https://tuetschek.github.io
@tuetschek

**Link** to these slides:
https://bit.ly/od-utn26

## Thanks

**Simone Balloccu** (now TU Darmstadt)
Eduardo Calò (Utrecht Uni)
Albert Gatt (Utrecht Uni)
Dimitra Gkatzia (Edinburgh Napier Uni)
David M. Howcroft (Uni Aberdeen)
Rudali Huidrom (ADAPT)
**Zdeněk Kasner**
**Ivan Kartáč**
**Karen Jia-Hui Li**
Saad Mahamood (Shopware)
**Mateusz Lango**
**Kristýna Onderková**
Ondřej Plátek
Ehud Reiter (Uni Aberdeen)
**Danil Semin**
Fahime Same (Trivago)
**Patrícia Schmidtová**
Adarsa Sivaprasad (Uni Aberdeen)
**Alex Terentowicz** (TU Poznan)
Vilém Zouhar (ETH Zurich)

Funded by the European Union (ERC, NG-NLG, 101039303) and Charles University projects GAUK 252986 and SVV 260821.
Using LINDAT/CLARIAH-CZ Research Infrastructure resources (Czech Ministry of Education, Youth and Sports project No. LM2023062).