

Beyond Traditional Benchmarks:

Analyzing Behaviors of Open LLMs on Data-to-Text Generation

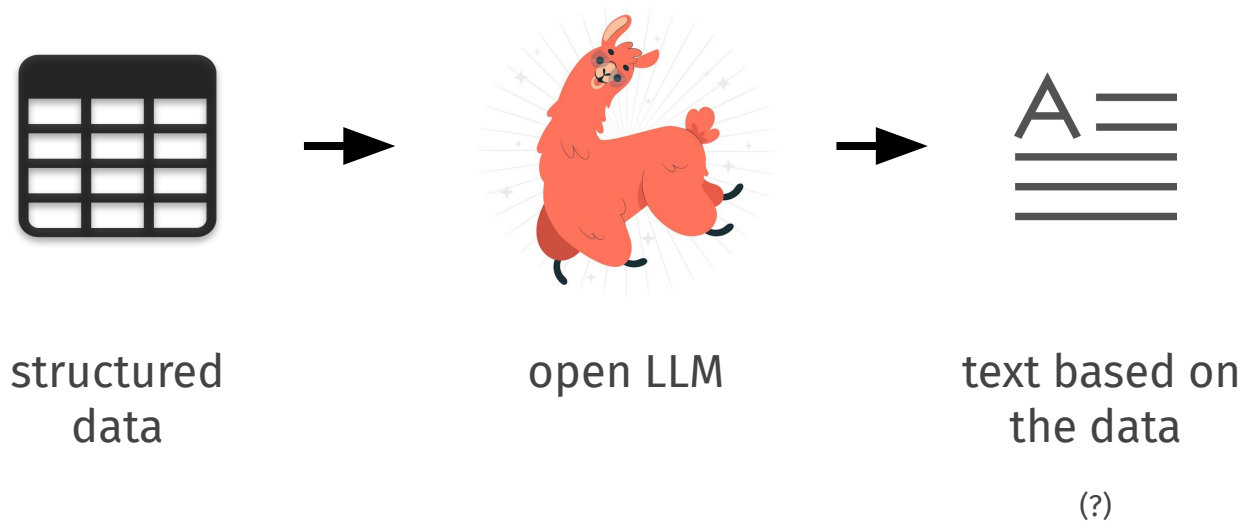
Zdeněk Kasner, Ondřej Dušek

Motivation

Data-to-text (D2T) generation



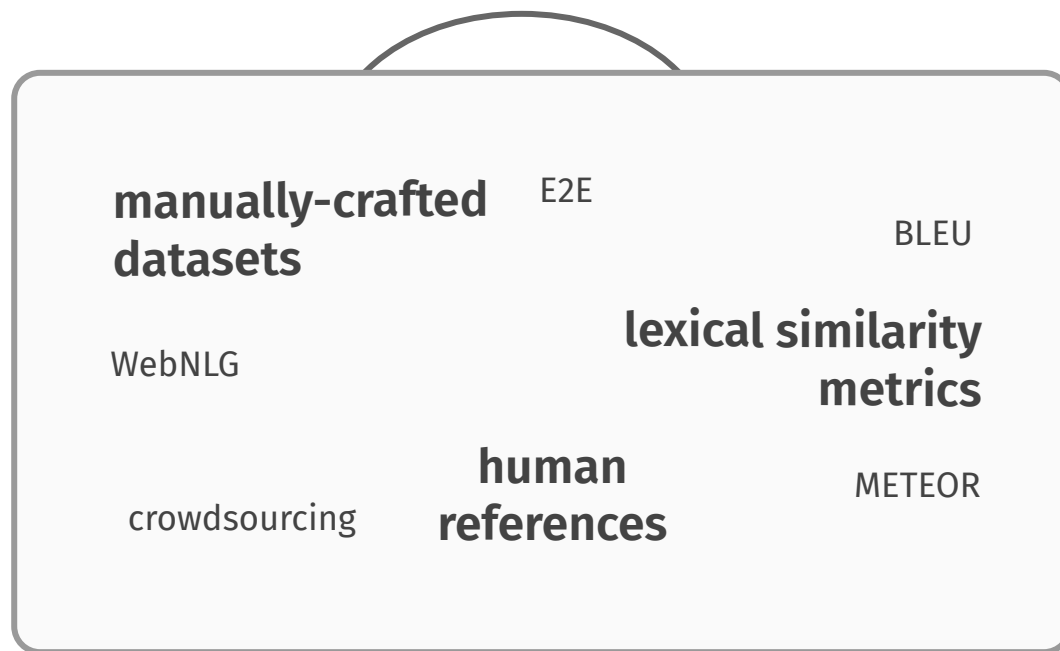
D2T generation with open LLMs



Research question:

How to faithfully observe the **behaviors of open LLMs** on data-to-text generation?

D2T benchmarking starter pack: 2019 edition



Problems

1

benchmarks are
small and saturated

*years of optimization
towards D2T test sets made
benchmarking unreliable*

2

benchmarks are
contaminated

*LLMs have seen many
datasets in their training
data*

3

traditional metrics
produce **a handful of
vague numbers**

*a single BERTScore is not
much better than a single
BLEU score*

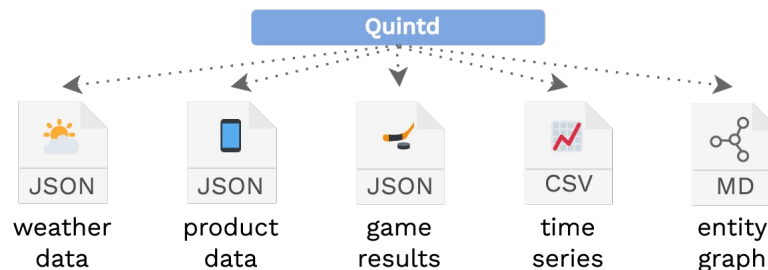
Our approach

Remedy #1: novel input data

Observation:

Structured data in common formats (JSON, CSV, ...) are plentiful.

Quintd: Our tool for downloading structured data for five tasks in distinct domains.

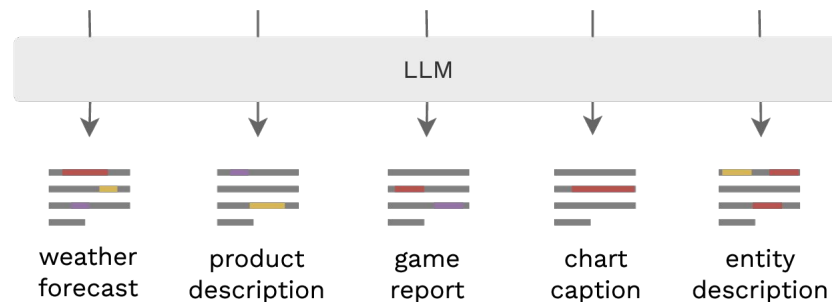


Remedy #2: in-context learning

Observation:

With LLMs, we can explain the task without references.

We use variations of the same prompt to produce the corresponding outputs.



Remedy #3: span-based annotations

Observation:

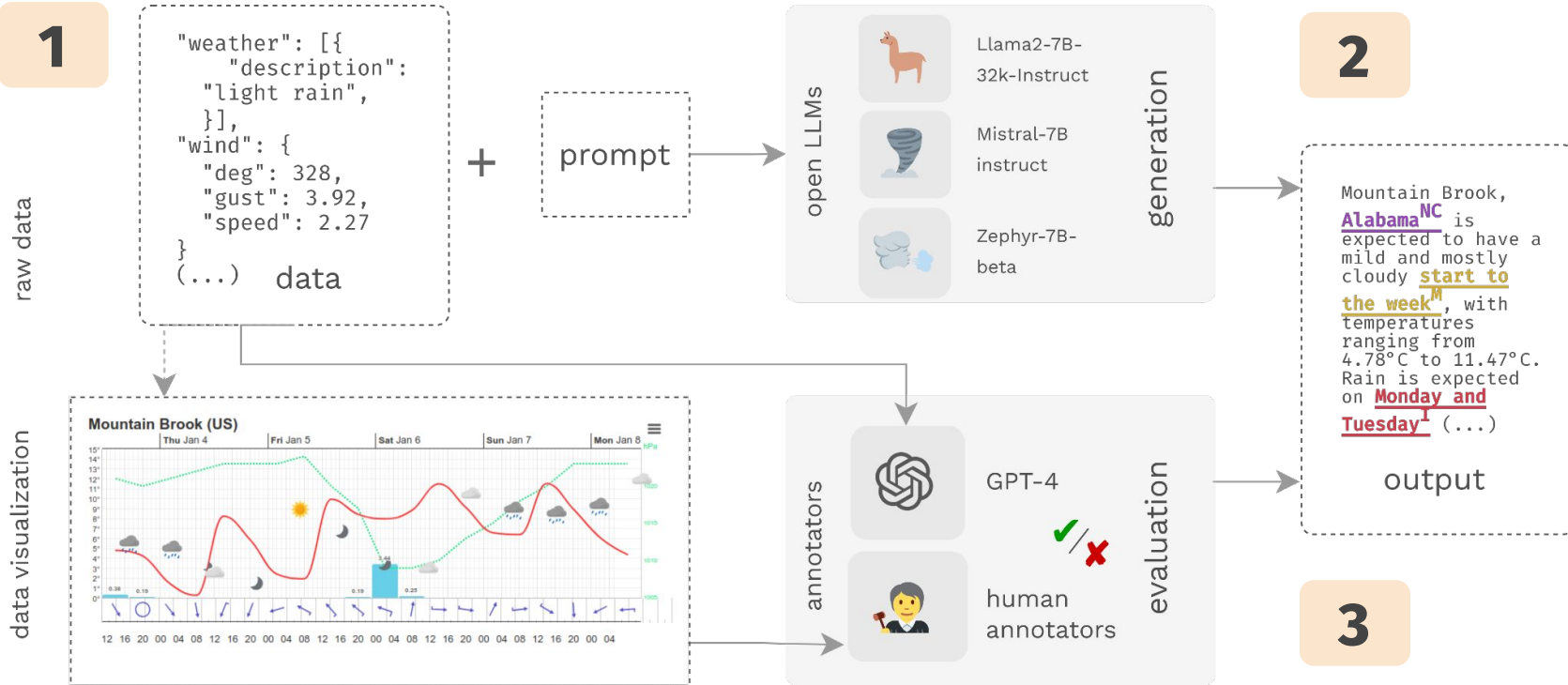
Outputs are fluent, but contain many semantic errors.

We annotate semantic errors based on the comparison with the input data.

llama2






The weather in Holt, England, will be mostly cloudy with scattered clouds for the next five days. The temperature will range from 4.2 to 7.4 degrees Celsius, with a high of 6.1 degrees Celsius on January 7th. The wind speed will be around 2.5 to 4.6 meters per second, with gusts of up to 9.32 meters per second on January 4th. There will be light rain on January 4th, with a total of 0.12 millimeters of precipitation. The humidity will be around 77% to 97%.

Putting it all together






Experiments & Results

Data & Tasks

	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



1000 examples in total (100 dev + 100 test per domain)

LLMs

	Llama2-7B-32k-Instruct
	Mistral-7B instruct
	Zephyr-7B-beta

(+GPT-3.5 for comparison)

Metrics

	GPT-4
	✓ / ✗
	human annotators

Quantitative

- on average, open LLMs produced:
 - at least one semantic error in **80% of outputs**
 - more than **3 errors per output**

Qualitative

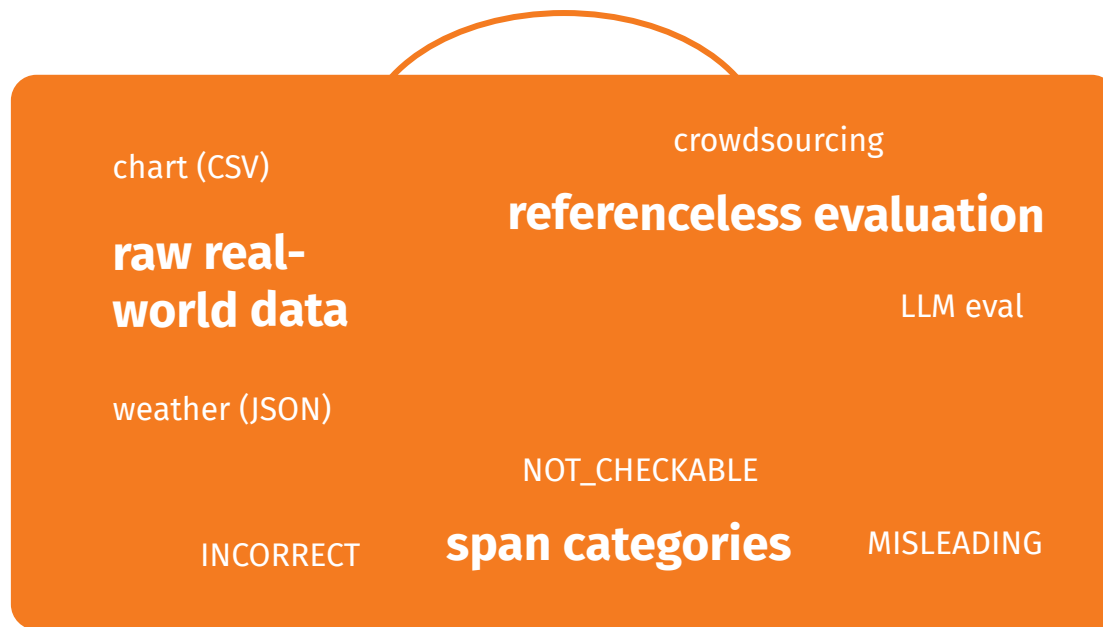
- long inputs cause practical issues
- how to get better outputs:
 - include units
 - remove unnecessary fields
 - prefix the model answer

Interactive data explorer

d2t-llm.github.io

(For the [data annotation toolkit](#), see our INLG 2024 demo paper:
factgenie: A Framework for Span-based Evaluation of Generated Texts)

D2T benchmarking starter pack: 2024 edition



Contact us



Zdeněk Kasner

`kasner@ufal.mff.cuni.cz`



Ondřej Dušek

`odusek@ufal.mff.cuni.cz`

<https://d2t-11m.github.io>



**Funded by
the European Union**



European Research Council

Established by the European Commission