Beyond Traditional Benchmarks:
Analyzing Behaviors of Open LLMs on Data-to-Text Generation

Zdeněk Kasner
kasner@ufal.mff.cuni.cz

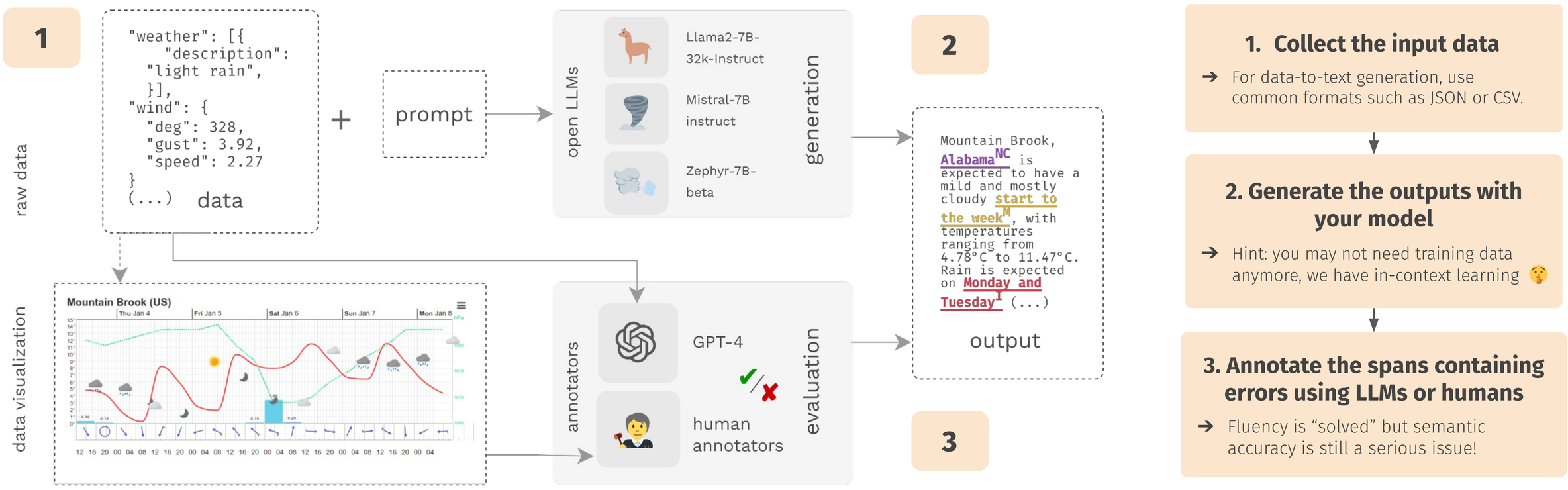
Ondřej Dušek
odusek@ufal.mff.cuni.cz

Charles University

ÚFAL

We need to **re-think evaluation of data-to-text generation** for the LLM era

How to evaluate LLMs that were trained on many existing benchmarks and produce fluent text, but still make semantic errors? Case study on data-to-text generation:



Our experiments

	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

Data:

- 1000 ex. (100 dev + 100 test per domain)
- collected from public APIs with our data collection tool **quintd**

Experiments:

- data-to-text: {Llama 2, Mistral, Zephyr}-7B
- evaluation: human annotators, GPT-4

4 out of 5 outputs of open LLMs contained **at least one semantic error**

Starter packs

for benchmarking data-to-text generation systems

2019*

*year only approximate, can be also seen in the present day

manually-crafted datasets · E2E · BLEU · **lexical similarity metrics** · WebNLG · crowdsourcing · **human references** · METEOR

output:

→ a handful of vague numbers

2024

chart (CSV) · crowdsourcing · **raw real-world data** · **referenceless evaluation** · LLM eval · weather (JSON) · NOT_CHECKABLE · **span categories** · INCORRECT · MISLEADING

→ token-level annotations, insights!

Main conference paper at ACL 2024, Bangkok, Thailand.

Supported by the European Union (ERC, NG-NLG, 101039303) and the Charles University project SVV 260 698.

Code & data explorer →
d2t-llm.github.io