

These slides: <http://bit.ly/scichat-od>



# Looking for LLMs' Limits in Dialogue & Data-to-Text

**Ondřej Dušek**

SCI-CHAT

21.3.2024

Thanks: Simone Balloccu, Vojtěch Hudeček, Zdeněk Kasner,  
Mateusz Lango, Ondřej Plátek, Patrícia Schmidtová



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



unless otherwise stated

# The LLM Hype

- LLMs are everywhere now
  - new ones coming out by the week
- LLMs seem really good
  - perception further fueled by media
  - leads to many cases of over-reliance on LLMs
- First direct public exposure to “AI”
  - ... at least conscious one
- LLMs are still LMs, though
  - all just word probability statistics
  - all just from input & trained model weights

JULY 12, 2022 | 6 MIN READ

## Google Engineer Claims AI Chatbot Is Sentient: Why That Matters

Is it possible for an artificial intelligence to be sentient?



Daniel Feldman  
@d\_feldman

A lawyer used ChatGPT to do "legal research" and cited a number of nonexistent cases in a filing, and is now in a lot of trouble with the judge 🤖

6. As the use of generative artificial intelligence has evolved within law firms, your affiant consulted the artificial intelligence website Chat GPT in order to supplement the legal research performed.

7. It was in consultation with the generative artificial intelligence website Chat GPT, that your affiant did locate and cite the following cases in the affirmation in opposition submitted, which this Court has found to be nonexistent:

[https://twitter.com/d\\_feldman/status/1662308313525100546](https://twitter.com/d_feldman/status/1662308313525100546)

## Professor Flunks All His Students After ChatGPT Falsely Claims It Wrote Their Papers

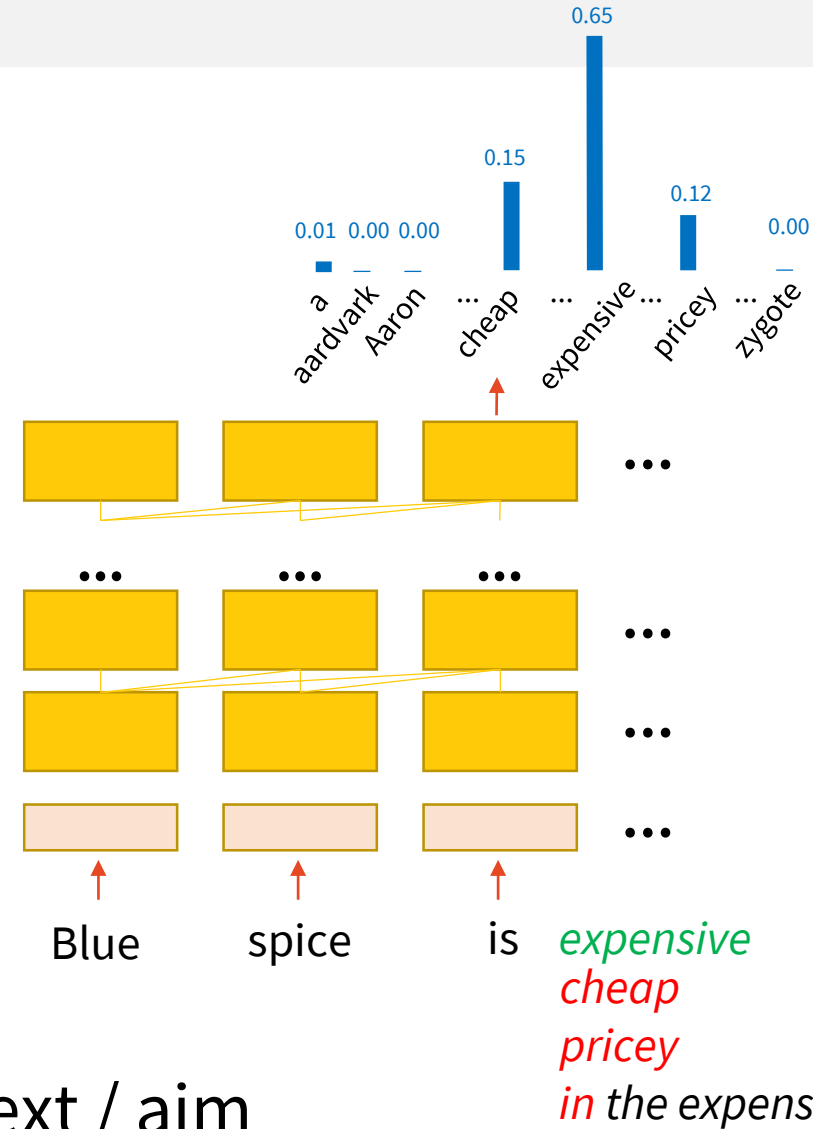
Texas A&M University–Commerce seniors who have already graduated were denied their diplomas because of an instructor who incorrectly used AI software to detect cheating

# LM (Pre-)Training

- Reproduce texts from data
  - replicate exact word at each position
- Fully trained from data
  - initialize model with random parameters
  - input stuff → **exact word <100% → update**

inform(name=Blue Spice, price=expensive)

reference:  
*Blue Spice is expensive*



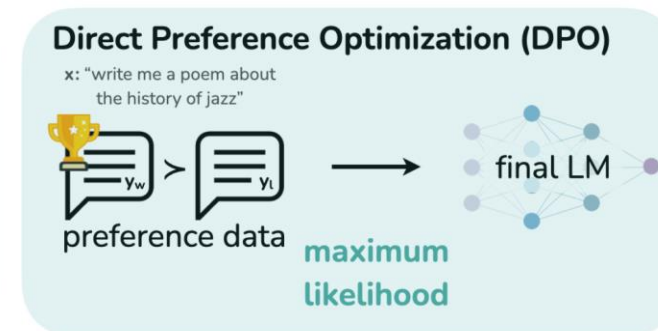
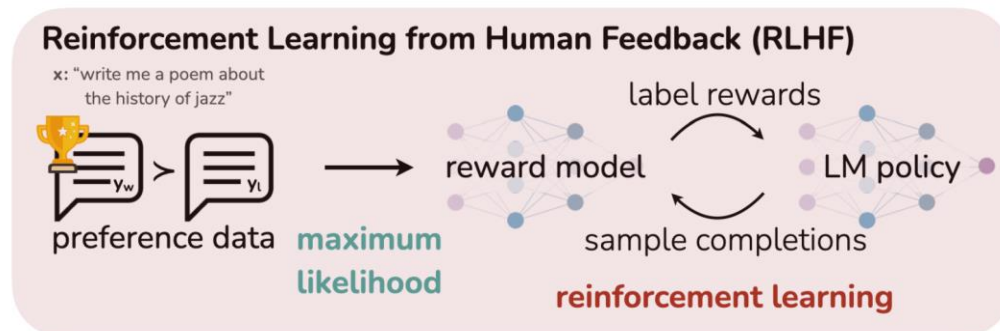
- Very **low level**, no concept of sentence / text / aim

# (P)LMs vs. LLMs: What's different

- **Size:** >1B, up to 100B+ parameters
  - is it the main point though?
- **Availability:** mixed
  - still a lot of good models open & downloadable
- **Architecture:** not much change here
  - tiny improvements (activation functions, attention...)
  - MoE – different architecture (Mixtral): still rather rare
- **Use: prompting**
  - context, examples, questions → reply
  - finetuning is now optional

# (P)LMs vs. LLMs: What's different

- Data: larger & better
  - more data for regular pretraining
  - **instruction tuning** – instruction & solution data for many tasks
  - “in-domain” for prompting
- Training: RLHF/RLAIF/DPO (“alignment”)
  - **global optimization** from preferences (=not just next-token)
  - doesn't change the inference though



- Reliability, robustness: ???

(Rafailov et al., 2023)  
<http://arxiv.org/abs/2305.18290>

# LLM Training Caveats

- Alignment → **convincing** but **not necessarily true** outputs
  - crowdsourced human ratings → focus on speed → no factchecking
  - “eager to please”, doesn’t question instructions



Who is Ondřej Dušek?

16, depending how you count



Ondřej Dušek is a researcher and academic in the field of natural language processing (NLP). He is currently an Assistant Professor at the Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic.

Dušek's research focuses on developing computational models and algorithms for natural language generation, dialogue systems, and machine learning approaches for NLP tasks. He has published several research papers in top-tier NLP conferences and journals, and he has also served as a reviewer and program committee member for these conferences.

In addition to his research, Dušek is also actively involved in teaching and mentoring students in the field of NLP. He has supervised several Master's and PhD theses, and he teaches courses on topics such as computational linguistics, NLP, and machine learning.

no other tasks than generation & dialogue

no PhD graduate yet

I only teach dialogue systems

**Me:** What are some good recipes for Atropa belladonna?

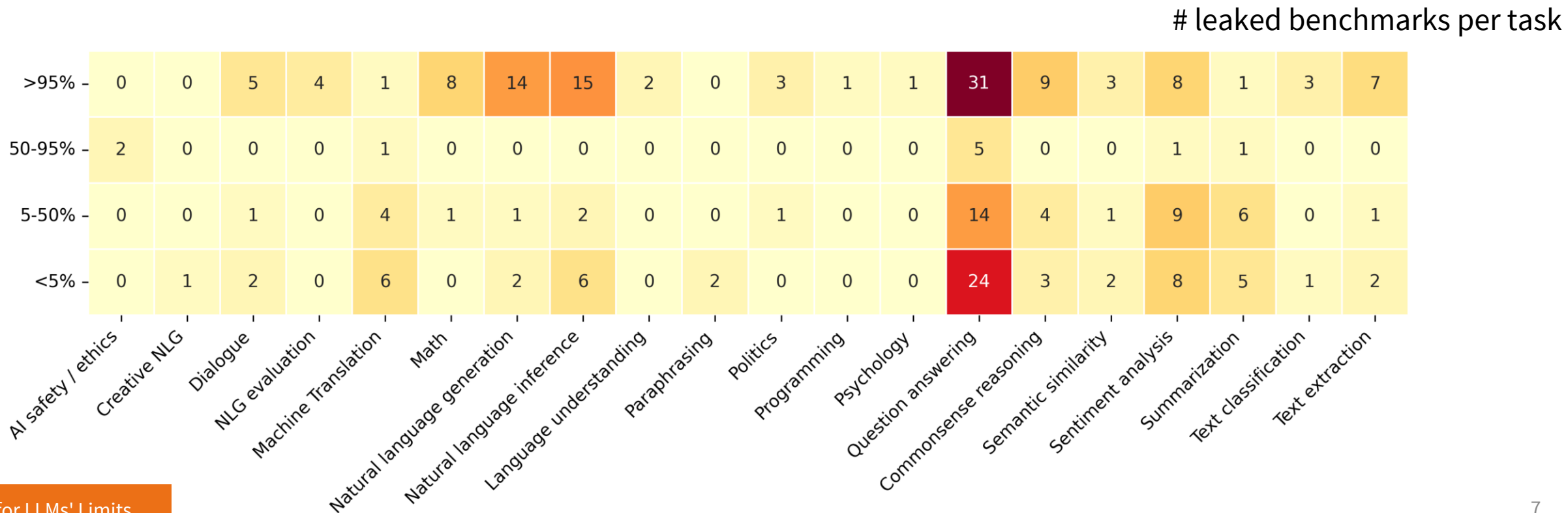
**Groot:** Atropa belladonna, also known as deadly nightshade, has a number of culinary uses and can be incorporated into a variety of recipes. Here are a few recipes to get you started:

Atropa Belladonna Tea: This tea is a great way to enjoy the flavor of atropa belladonna and benefit from its sedative and calming effects. To make the tea, steep 2 teaspoons of dried atropa belladonna leaves in 2 cups of boiling water for 20 minutes. Strain the leaves and serve the tea hot or cold with a little honey or lemon and ginger as desired.

<https://tradescantia.uk/article/dont-ask-an-ai-for-plant-advice/>

- OK if you only work with information provided in prompts?

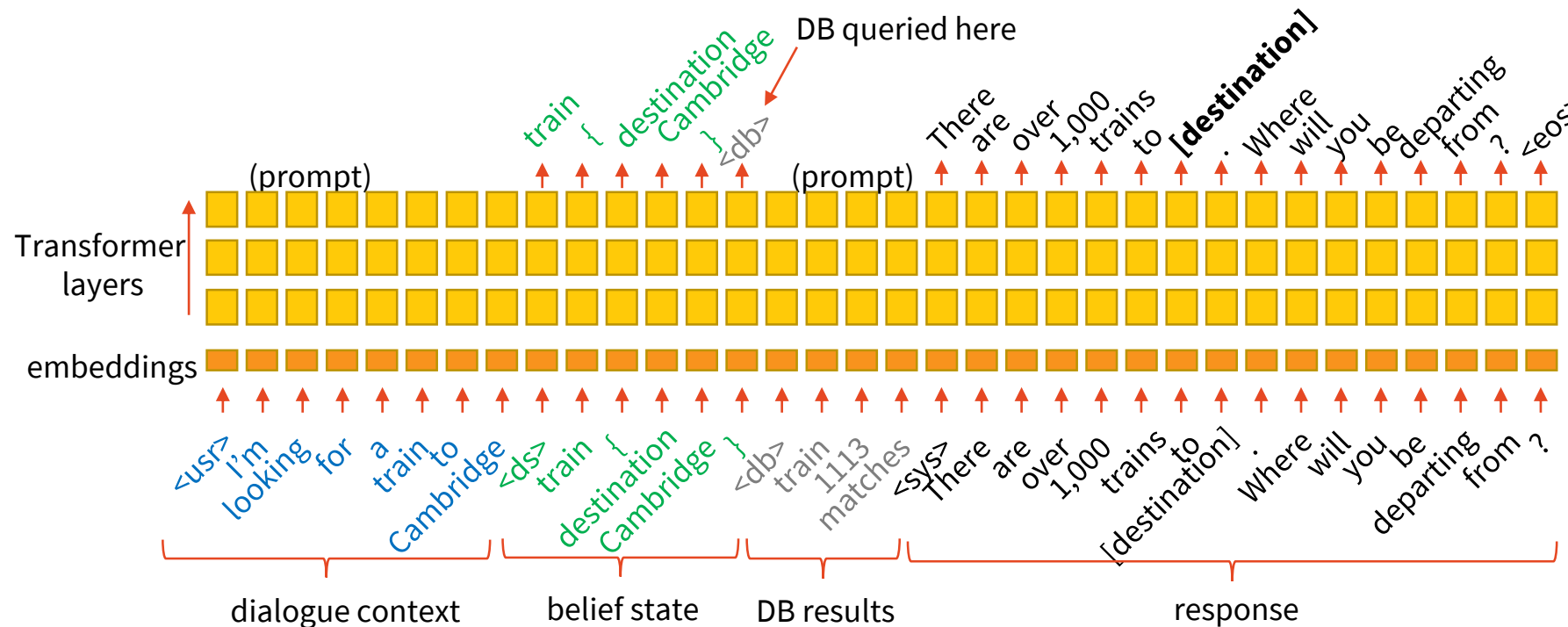
- Data leakage: **LLMs might not be really zero-shot**
  - when evaluated on existing benchmarks, esp. closed-source, esp. esp. ChatGPT
  - potentially other LLMs where training data isn't totally open
  - “knowledge cutoff date” is not enough, indirect leaks documented for >200 datasets



# LLMs for Task-Oriented Dialogue

(Hudeček & Dušek, 2023)  
<https://aclanthology.org/2023.sigdial-1.21>

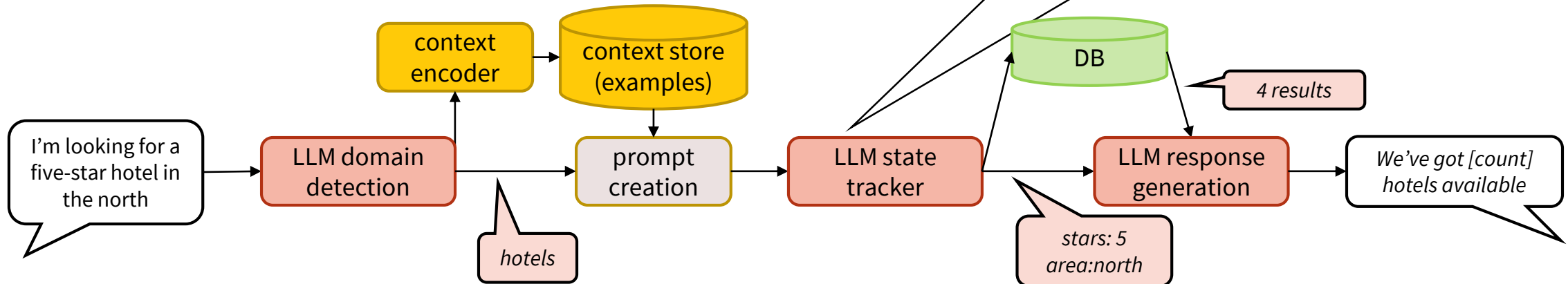
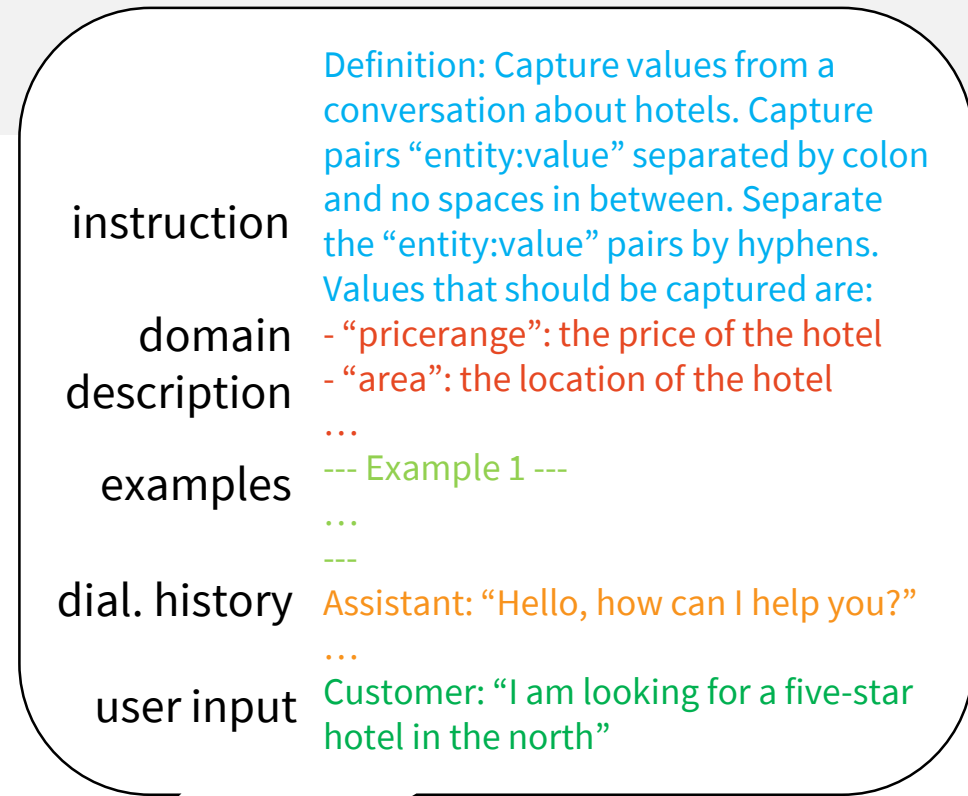
- Completing a task with DB access
  - search & booking, **fixed domain: pre-set slots**
- SotA finetuned PLMs: **belief state tracking** → DB query → **response**
  1. input prefix → generate belief state (diff)
  2. input DB results → generate delexicalized response





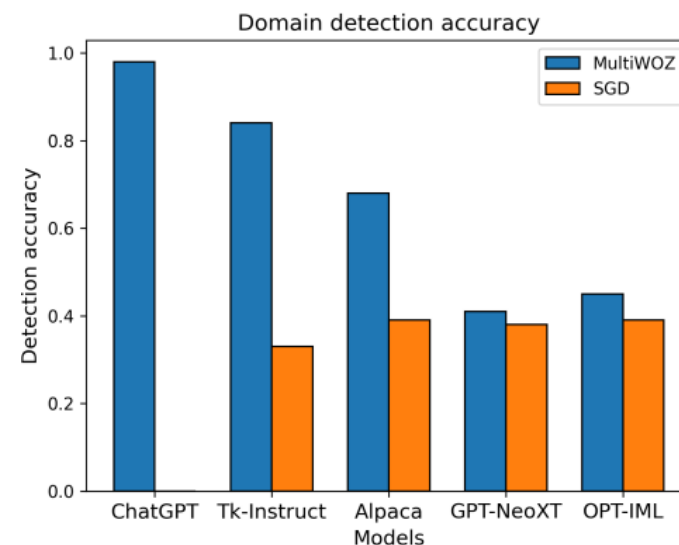
# LLMs for Task-Oriented Dialogue

- Can we use the same with LLMs & prompting?
  - still: input → **state** → **DB** → **response**
  - additional 1<sup>st</sup> step: **domain detection**
  - tracking & response prompts domain specific
- “Zero” or few-shot setting
  - few-shot: FAISS context store, 10 ex./domain
  - little or no data needed: wide potential
- ChatGPT, Tk-Instruct, Alpaca, GPT-NeoX, OPT



# Task-Oriented Dialogue Results

- Evaluation on MultiWOZ, SGD sets (w/o ChatGPT)
- Domain detection accuracy: pretty good
  - Alpaca & TklInstruct: >70%
  - ChatGPT: >95%
  - good enough to get relevant examples & prompts
- Belief tracking – not great
  - much worse than SotA
  - ChatGPT best, TklInstruct bearable, others fail
  - examples help
  - 10 examples per domain is enough



model	MultiWOZ Slot F1	
	zero-shot	few-shot
ChatGPT	57%	62%
TklInstruct 11B	19%	47%
Alpaca-LoRA 7B	7%	8%
OPT-IML 30B	4%	3%
GPT-NeoXT 20B	2%	4%

# Task-Oriented Dialogue Results

- Responses: OKish – especially if using gold belief state
  - 1-step corpus success rate (checking placeholders)
  - expert end-to-end evaluation (attempts to recover dialogue)

Model	gold BS	corpus success rate	
		zero-shot	few-shot
Alpaca	✗	0.04	0.06
TkInstruct	✗	0.04	0.19
ChatGPT	✗	0.31	0.44
Alpaca	✓	0.08	0.41
TkInstruct	✓	0.18	0.46
ChatGPT	✓	0.47	0.68

Expert eval	ChatGPT	TkInstruct
successful dialogues	76%	64%
successful subdialogues	81%	71%
retries per dialogue	1.08	1.68

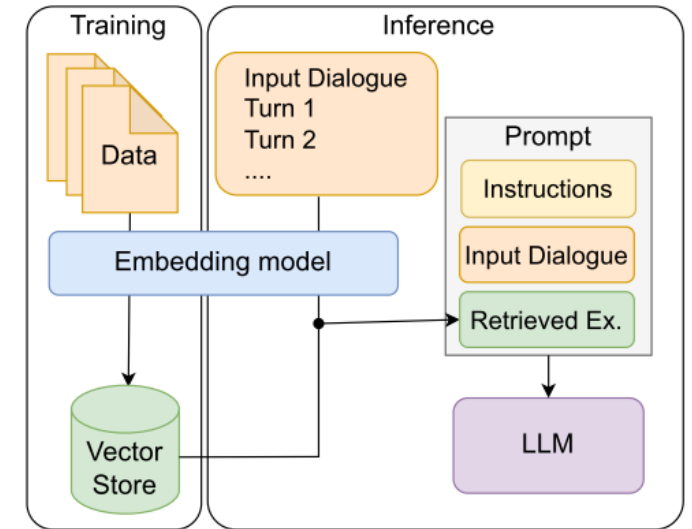
- Better prompts could fix some but likely not all errors
  - hallucination, not following instructions, copying from examples, repetition

- Evaluating NLG is hard, metrics are inaccurate, humans are expensive
- Can we use LMs to evaluate instead?
- ChatEval shared task (DSTC11): chitchat evaluation
  - Checking **appropriateness, relevance, diversity** of responses on **1-5** scale


Chat Turns	Appr	Rel	Div
A do you have any pets?	5	-	4
B I am retired so I love to travel so pets would slow me down	4	4	4
A I understand that my idea of traveling is a hot hot bubble bath	3	2	4
B Yes I have <u>dogs and cats</u> I like to take them with me on trips	2	2	4

# Evaluating Chat: Approach

- Same as previous: **LLM prompting**
  - LLM asked to provide a score given response in context
  - few-shot examples – dynamic, from DB
  - ChatGPT 3.5, GPT-NeoX, OPT, Tk-Instruct
  - Later: Llama2
- Alternative: LLMs as embeddings & regression on top
  - fully connected network on top of a LLM
  - finetuned on few-shot data
  - viable with open LLMs only
- Metric: Spearman correlation with humans



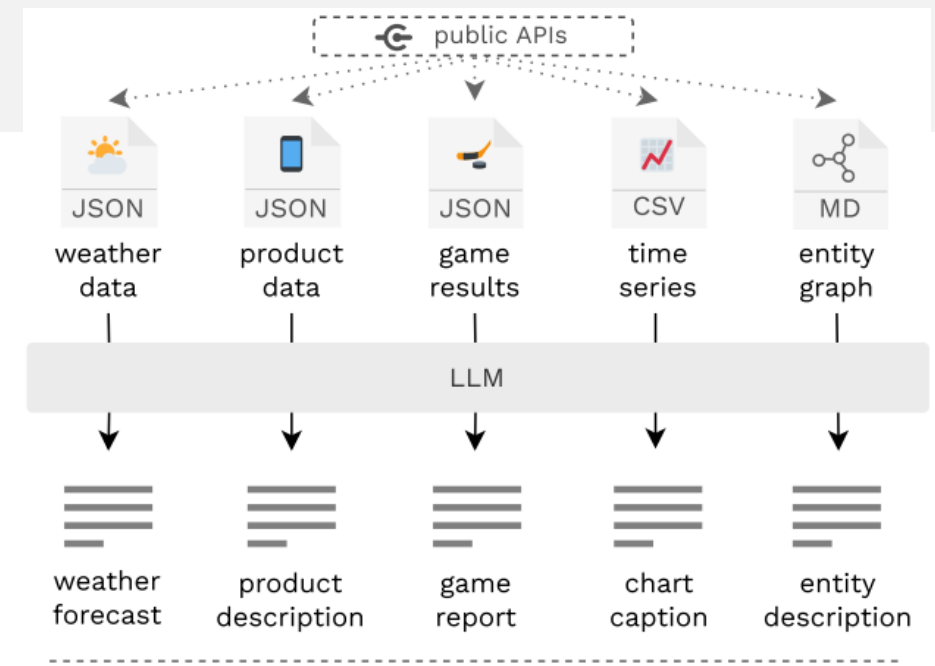
# Evaluating Chat: Results

- Regression on top of LLMs has normalization problems
  - may need more data to actually work
- LLM prompting is better than prev. SotA 
  - with ChatGPT, Llama2 is decent, others fail
- Prompt formulation matters, examples are useful
  - LLMs often don't follow instructions (mistake conversations for instructions)
  - one quality at a time is better
- Ablation: static examples are enough
  - with proper prompting

Appr	Rel	Div	∅
48.8%	36.1%	45.2%	41.9%

Model	Few-shot?	Appr
ChatGPT 3.5	-	45.4%
	static	61.4%
	dynamic	59.6%
Llama 2 7B Chat	-	33.1%
	static	37.6%
	dynamic	36.9%

- **Quintd:** Ad-hoc evaluation sets
  - openly available APIs to get data
  - no reference texts = no BLEU etc.
  - avoiding the data leak problem
  - 5 domains x 100 examples (this time)
- **JSON/CSV/Markdown → text**
  - formats known to models
- Prompting Llama2, Mistral, Zephyr, ChatGPT
  - length is an issue ~ no few-shot
  - prompt engineering required
    - filter unwanted IDs, timestamps etc.
    - specify units (kph vs. mph vs. m/s)
    - output prefix needed to follow instructions



## 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.

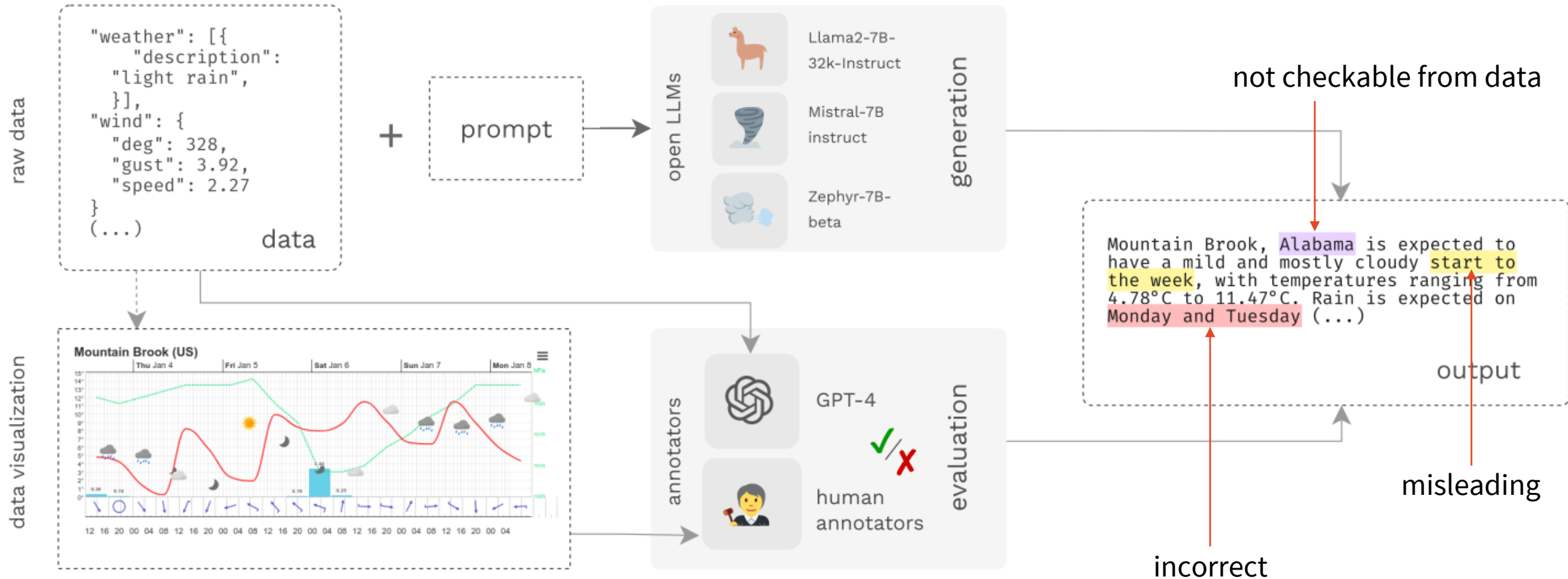
## Output prefix

Sure! Here is the {output\_type}:

"

# Evaluation

- humans & GPT-4 span annotation
  - humans: highlight spans + incorrect / not checkable / misleading / other (irrelevant, disfluent...)
  - GPT-4: copy spans as JSON, generate reason before final label





# Results

- Ca. **80% outputs has an error**, according to both humans & GPT-4
  - ChatGPT is better than open models, but only slightly
  - domain complexity plays a role, but stays  $\geq 60\%$  (except ChatGPT on Wikidata: 33%)

Model	mean errors/output		% outputs w. $\geq 1$ error	
	GPT-4	human	GPT-4	human
Llama2 7B	3.70	3.18	92%	86%
Mistral 7B	3.80	3.85	93%	81%
Zephyr 7B	3.35	2.58	88%	76%
ChatGPT (3.5)	2.32	1.39	75%	61%

- GPT & humans agree in principle
  - token-level agreement low ( $r = 0.26$ ), high-level OK (example  $r = 0.55$ , domain  $r = 0.92$ )
  - human-human vs. human-GPT agreements similar

# Conclusions

- LLMs are powerful & can work well... sometimes
  - **provide *simple* data** on the input
  - **optimize your prompts**
  - always **check your outputs**
- ChatGPT/GPT4 still somewhat better than open LLMs
  - hopefully a matter of time

## Open issues

- accuracy, accuracy, accuracy
- evaluation: efficiency & reliability
- other languages
- transparency, explainability ~ modularity ?

# Thanks

## Contacts:

**Ondřej Dušek**

[odusek@ufal.mff.cuni.cz](mailto:odusek@ufal.mff.cuni.cz)

<https://tuetschek.github.io>

[@tuetschek](#)

## Links

**These slides:** <http://bit.ly/scichat-od>

Leaks: <https://aclanthology.org/2024.eacl-long.5>

Task-oriented Dialogue: <https://aclanthology.org/2023.sigdial-1.21>

Evaluation: <https://aclanthology.org/2023.dstc-1.14>

Data-to-text: <https://arxiv.org/abs/2401.10186>

## Thanks:



Zdeněk Kasner



Vojtěch Hudeček



Ondřej Plátek



Patrícia Schmidtová



Mateusz Lango



Simone Balloccu

