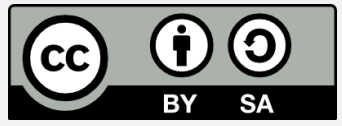
OD's LLM trials

NPFL140 LLMs

11/4/2024

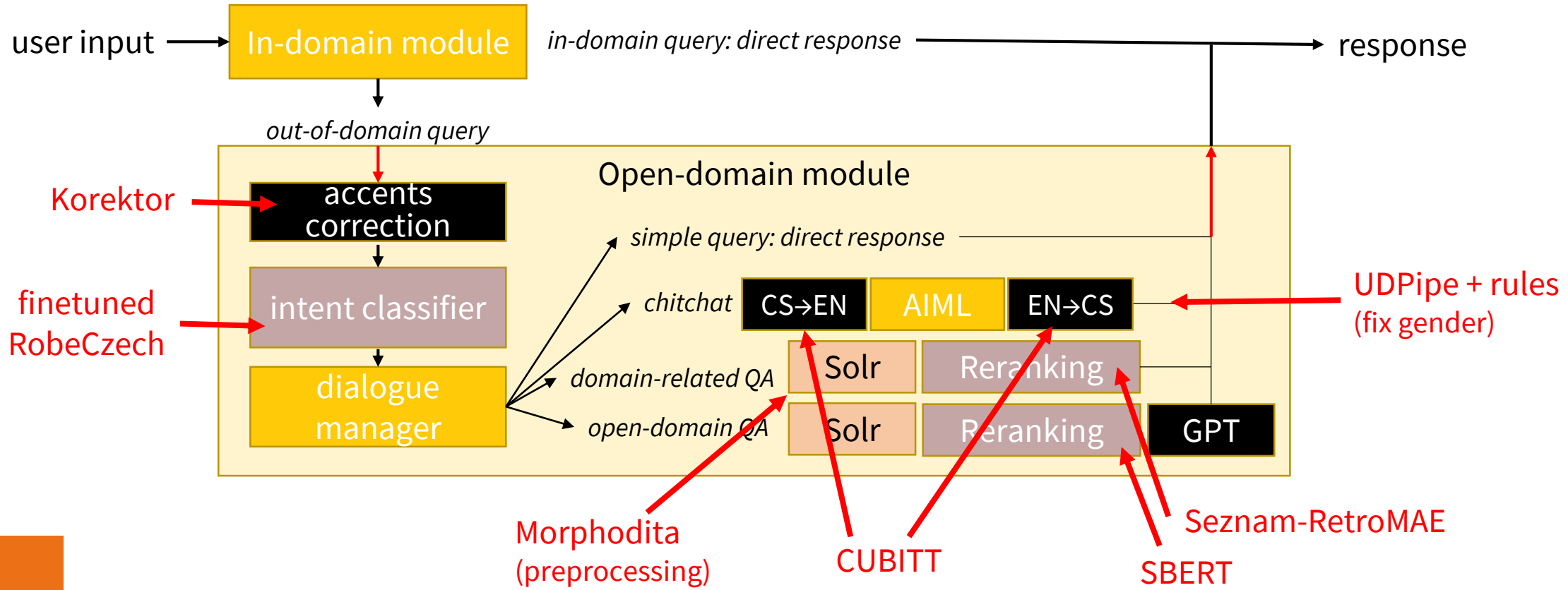
Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics



EDU-AI: Chatbot for high-school students

http://edu-ai.eu/

- Only non-rule-based stuff passed to us (course training handled elsewhere)
- Chitchat: we switched from BlenderBot LM to AIML rules (no data, must have control)
- QA on Wikipedia: using retrieval-augmented generation with GPT3.5
  - lots of tinkering needed: query filtering, abbrev expansion, lemmatizing…



EDU-AI Czech Wikipedia QA

1. Retrieve Wiki articles
   - dump of 1st paragraphs, Solr IR
2. Rerank
   - choose best article via SBERT
3. Article & question → **GPT3/3.5/4**
   - option to tell it's irrelevant
   - if irrelevant: just ask, no article (~hallucinate)
4. Return GPT's response

- Better than other QA models
  - 60-70% token F1
  - vs. 10-20% for T5s etc.
  - SQuAD QA is a toy problem, retrieving relevant article is the issue
  - "hallucination saves us"

```
Q: Jak se jmenuje nejvyšší hora České republiky?

Solr Retrieval:
    1. Nejvyšší hora Dominikánské republiky/9.390833
    2. Nejvyšší hora Jihoafrické republiky/9.218349
    3. Nejvyšší hora Burundi/9.035112
    4. Nejvyšší hora Středoafrické republiky/8.894738
    5. Nejvyšší hora Republiky Kongo/8.675601
    6. Nejvyšší hora Konžské republiky/8.675601
    7. Venkovský dům čp. 176 (Andělská Hora)/8.565605
    8. Nejvyšší soud České republiky/8.144595
    9. Kostel svatého Vavřince (Černá Hora)/8.08573
   10. Morový sloup (Kutná Hora)/8.043259

SBERT Reranking:
    1. Venkovský dům čp. 176 (Andělská Hora)/0.3883
    2. Morový sloup (Kutná Hora)/0.4335
    3. Nejvyšší hora Republiky Kongo/0.5180
    4. Nejvyšší hora Konžské republiky/0.5180
    5. Kostel svatého Vavřince (Černá Hora)/0.5253
    6. Nejvyšší hora Středoafrické republiky/0.5368
    7. Nejvyšší soud České republiky/0.5881
    8. Nejvyšší hora Jihoafrické republiky/0.5894
    9. Nejvyšší hora Dominikánské republiky/0.6010
   10. Nejvyšší hora Burundi/0.6519

GPT3 prompt: Odpověz na otázku s využitím kontextu.
    Využij pouze informace z kontextu, kopíruj text co nejvíc je to
    možné. Buď stručný a odpověz maximálně jednou větou. Nepoužívej
    více vět.
    Kontext: Venkovský dům čp. 176 stojí na katastrálním území
             Andělská Hora ve Slezsku. Je představitelem lidové
             kultury východosudetského…
    Otázka: Jak se jmenuje nejvyšší hora České republiky?
    Odpověď:

Response: Myslím, že Sněžka je nejvyšší hora České republiky.
(Zdroj: https://cs.wikipedia.org/wiki/Venkovský_dům_čp._176_…
```

older prompt version, now we explicitly ask about relevance

Task-oriented Dialogue with LLMs

- **Assistant:** fulfill user requests (book a hotel/restaurant/taxi etc.)
  - need to **connect to external DB & provide correct info**
  - MultiWOZ: benchmark for this, 10k dialogues

(Budzianowski et al., 2018)
https://aclanthology.org/D18-1547/

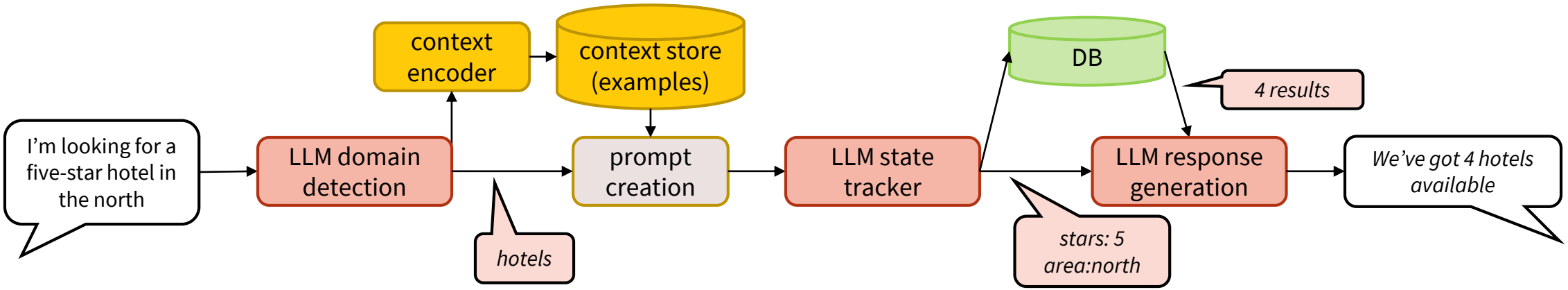
1. user: I am looking for a train from Cambridge to London Kinks Cross.
   state: *{train {departure = cambridge, destination = london kings cross}}*
   DB: *{train (70) {…}}*
   system: There are **[count]** 70 trains departing from Cambridge to **[departure]** London Kings Cross **[destination]**. What day would you like to travel?
2. user: I would like to leave on Saturday after 18:45.
   state: *{train {day = saturday, departure = cambridge, destination = london kings cross, leave at = 18:45}}*
   DB: *{train (3) {arrive by = 19:51,21:51,23:51; id = TR0427,TR0925,TR4898; leave at = 19:00,21:00,23:00; … }}*
   system: **[id]** TR0427 leaves at **[leave_at]** 19:00 and arrives by **[arrive_by]** 19:51 . Would you like to book this train?
3. user: Yes, I would like to book it for eight people.
   system: I have booked it for you. Your reference number is **[reference]** 00000057. Is there anything else I can help you with?
4. user: I am also looking for an expensive restaurant in the centre.
   belief: *{restaurant {area = centre, price range = expensive} train {...}}*
   DB: *{restaurant (33) {area = centre (33); name=Curry Garden, ...; ...}, ...}*
   system: There are **[count]** 33 **[price_range]** expensive restaurants in the **[area]** centre. Is there a particular type of food you would like?

Task-oriented Dialogue with LLMs

- How good are LLMs if we require structure?
  - slots / DB are given
  - no finetuning ~ **prompting only**
  - ChatGPT, Tk-Instruct, Alpaca… (7-20B params)
- A few examples in prompt (context store)
  - wide application potential
- Still the same idea: **context** → **state** → **DB** → **response**
  - additional step needed: domain detection

instruction	Definition: Capture values from a conversation about hotels. Capture pairs "entity:value" separated by colon and no spaces in between. Separate the "entity:value" pairs by hyphens. Values that should be captured are:
domain description	- "pricerange": the price of the hotel - "area": the location of the hotel …
examples	--- Example 1 --- … ---
dial. history	Assistant: "Hello, how can I help you?" …
user input	Customer: "I am looking for a five-star hotel in the north"

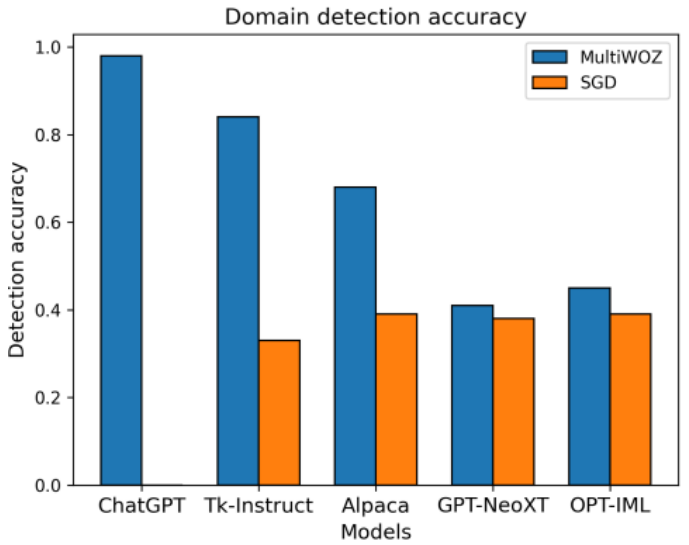
(Hudeček & Dušek, 2023)
https://aclanthology.org/2023.sigdial-1.21



Task-Oriented Dialogue Results

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

- Evaluation on MultiWOZ, SGD sets (w/o ChatGPT)
- Domain detection accuracy: pretty good
  - Alpaca & TkInstruct: >70%
  - ChatGPT: >95%
  - good enough to get relevant examples & prompts
- Belief tracking – not great
  - much worse than SotA (would be >80% slot F1)
  - ChatGPT best (~50-60% slot F1)
    TkInstruct bearable
    others fail



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

Task-Oriented Dialogue Results

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

- Responses: OKish – especially if using gold belief state
  - 1-step corpus success rate (checking for correct slot 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