LLMs are SOTA chat evaluators.
Baby steps on top of simple prompts.

Three Ways of Using LLMs to Evaluate Chat

Ondřej Plátek, Vojtěch Hudeček, Patricia Schmidtová,
Mateusz Lango, and Ondřej Dušek
{oplatek,hudecek,schmidtova,lango,odusek}@ufal.mff.cuni.cz

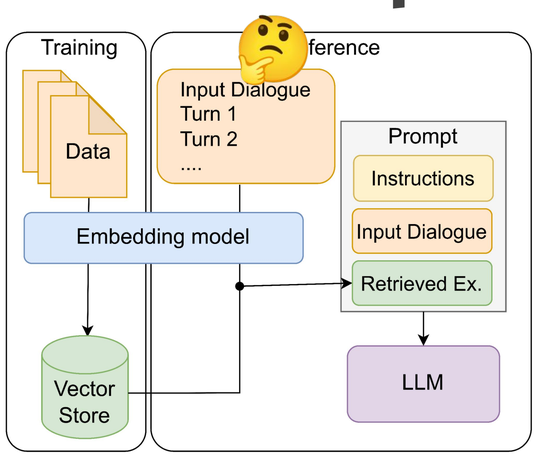
CHARLES UNIVERSITY

ÚFAL

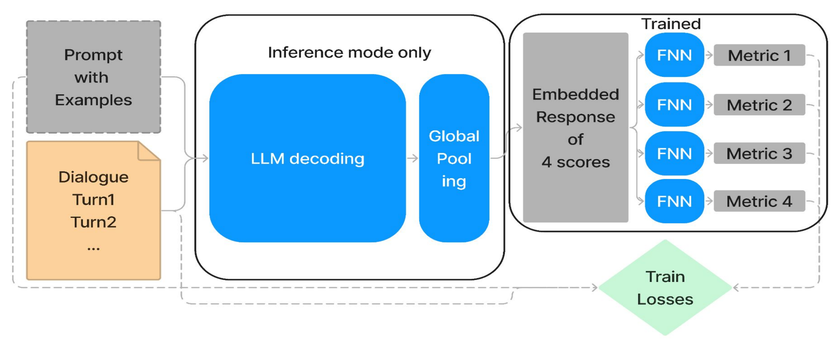
🦙 Prompts 🎉

- ❖ Trial and Error
- ❖ Single target metrics
- ❖ Examples of conversations
- ❖ Experiments:
  - *ChatGPT*
  - *Llama2*
  - *TkInstruct*
  - *OPT*
  - *GPT-NeoX*

Retrieve Examples?



LLM features? 😭



Ingredients

ChatEval DSTC11 Data

- ❖ **Target metrics** annotated **only for test set**
  Appropriatness, Content Richness, Grammaticality, Relevance
- ❖ Unclear target metrics for Training / Dev split
  - 👉 **Manual annotation** of target metrics for *156* rehearsal samples
  - 👉 **Heuristicly map existing datasets** in dev set to target metrics

ChatGPT1 Spearman

Dataset	Appr.	Rel.	Content
TEST-ALL	*0.488*	0.361	0.452
CHATGPT	*0.122*	0.060	0.181
DSTC10PERSONA	*0.803*	0.968	0.216
GPT3	*0.091*	0.007	0.242

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

LLM Systems

Submitted

- ❖ TkInstruct, GPT-X, OPT, etc. poor performance
- ❖ **ChatGPT 3.5**-turbo-0301 turn robustness **second place**
- ❖ Regressor on top of LLM global pooled logits

Ablation

- ❖ **Llama2,** ChatGPT 3.5-turbo-0613
- ❖ Fixed prompts

Experiments

Conclusions

- 👉 Prompted LLMs better than previous SOTA
- 👉 Prompts: fomatting matters *orig-> impr ->norm*
- 👉 Examples useful for format specifications.
- ❖ Llama2-7b-chat close to ChatGPT-turbo-0613
- ❖ Dynamic examples may be inferior.
  Future: Combine dynamic and preselected?
- ❖ Future: Add (QLoRa) finetuning?

Results

Best Submitted Avg. Spearman cor. coef.

system	Avg. Spearman
Baseline(Zhang et al., 2020)	0.3387
Winning submission (team4)	0.4890
ChatGPT + Vector Store (ours)	0.4190

Approprietness Spearman cor. coef.

system	Spearman	%fail
Llama2-Pnorm	0.3914	0.98%
Llama2-Pnorm-fix-2egs	0.3551	0.06%
Llama2-Pnorm-dyn-2egs	0.3756	0.65%
ChatGPT-Pnorm-dyn-2egs	0.5462	-
ChatGPT-Pimpr-fix-2egs	0.6136	0.00%
ChatGPT-Pimpr-dyn-2egs	0.5962	0.00%
ChatGPT-Porig-dyn-2egs	0.4880	-

Presented at DSTC 11 workshop, 2023, Prague.

https://github.com/oplatek/chateval-llm

Supported by Charles University projects GAUK 40222, SVV 260575 and the ERC grant 101039303 NG-NLG. It used resources provided by the LINDAT/CLARIAH-CZ Czech Ministry of Education, Youth and Sport LM20181101 grant. We thank Milan Fučík and Mateusz Krubiński for technical support and Zdeněk Kasner for the poster template.