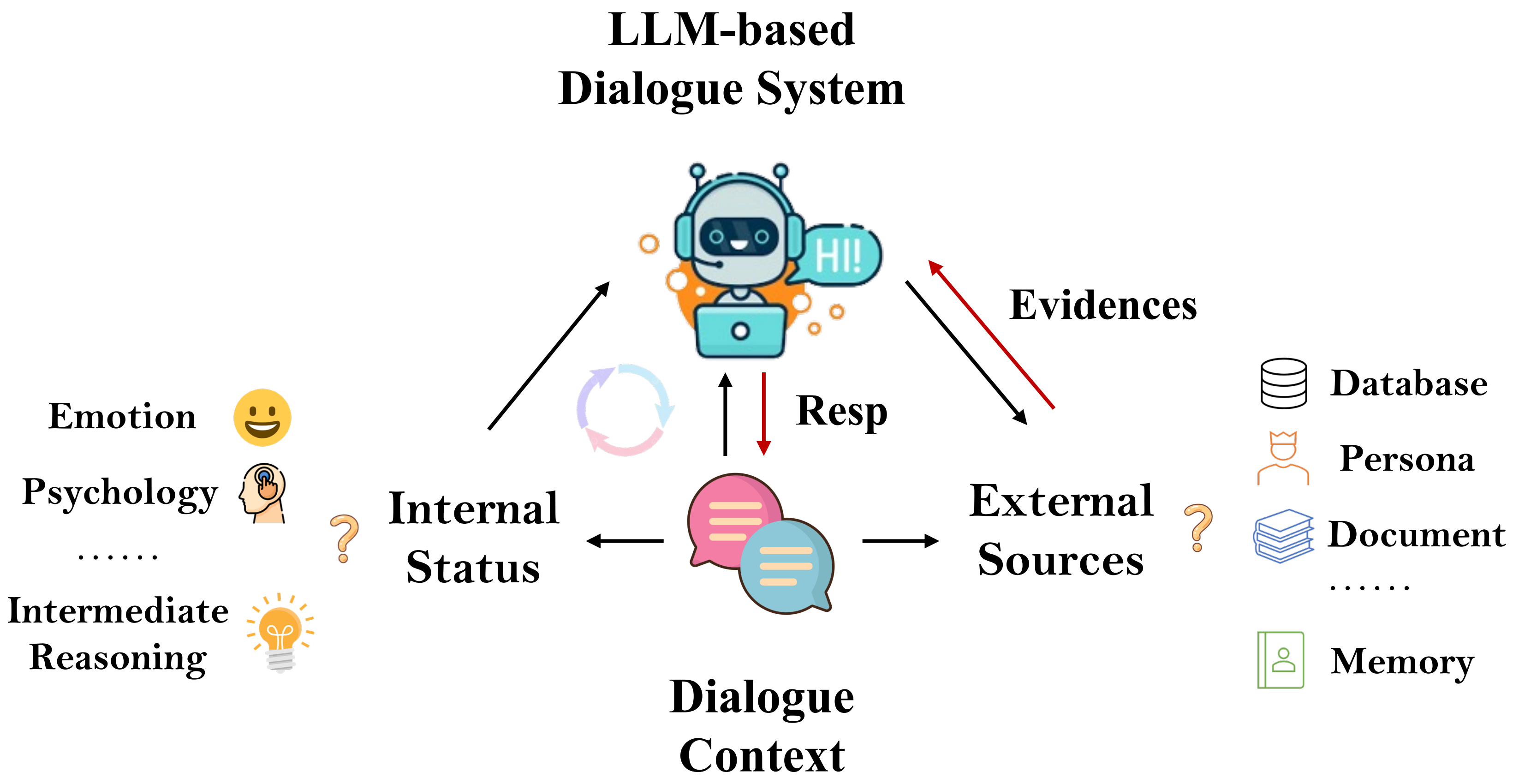
Cue-CoT: Chain-of-thought Prompting for Responding to In-depth Dialogue Questions with LLMs

Hongru Wang*, Rui Wang*, Fei Mi, Yang Deng, Zezhong Wang
Bin Liang, Ruifeng Xu, Kam-Fai Wong
MoE Key Laboratory of High Confidence and Software Technologies
The Chinese University of Hong Kong

➢ The Internal Capability of LLM-based Dialogue System



LLM-based
Dialogue System

Evidences

Resp

Emotion

Psychology

……

Intermediate
Reasoning

Internal
Status

External
Sources

Database

Persona

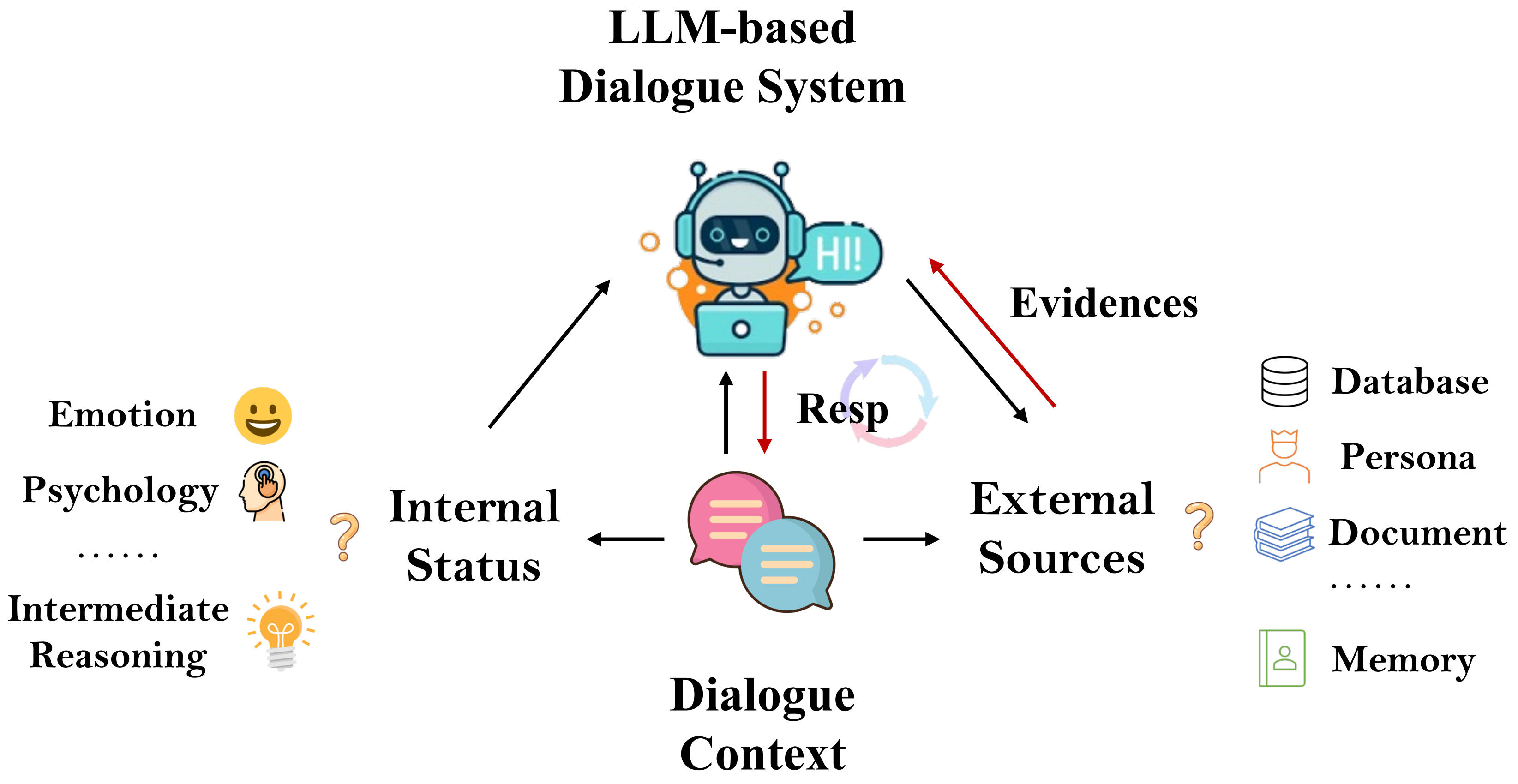
Document

……

Memory

Dialogue
Context

➢ The **External Capability** of LLM-based Dialogue System



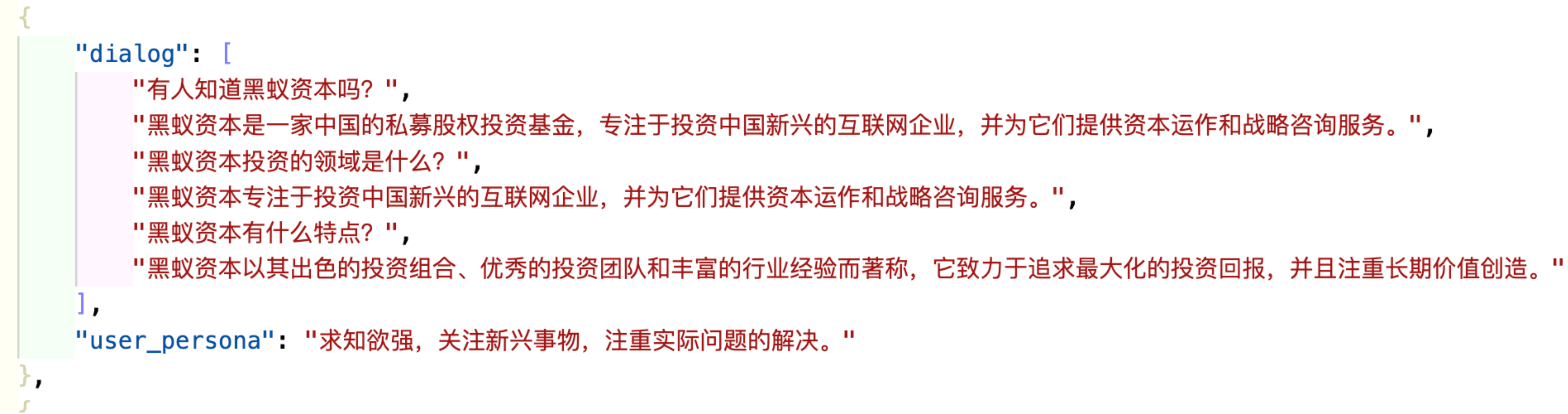
Check our another EMNLP 2023 paper

[Large Language Models as Source Planner for Personalized Knowledge-grounded Dialogues]()

➢ What is the internal capability ?

- **Large Language Models (LLMs)** possesses exceptional natural ***language understanding*** and ***generation*** capabilities, making it an ideal foundation for a conversational system that provides a great user experience and significant convenience.

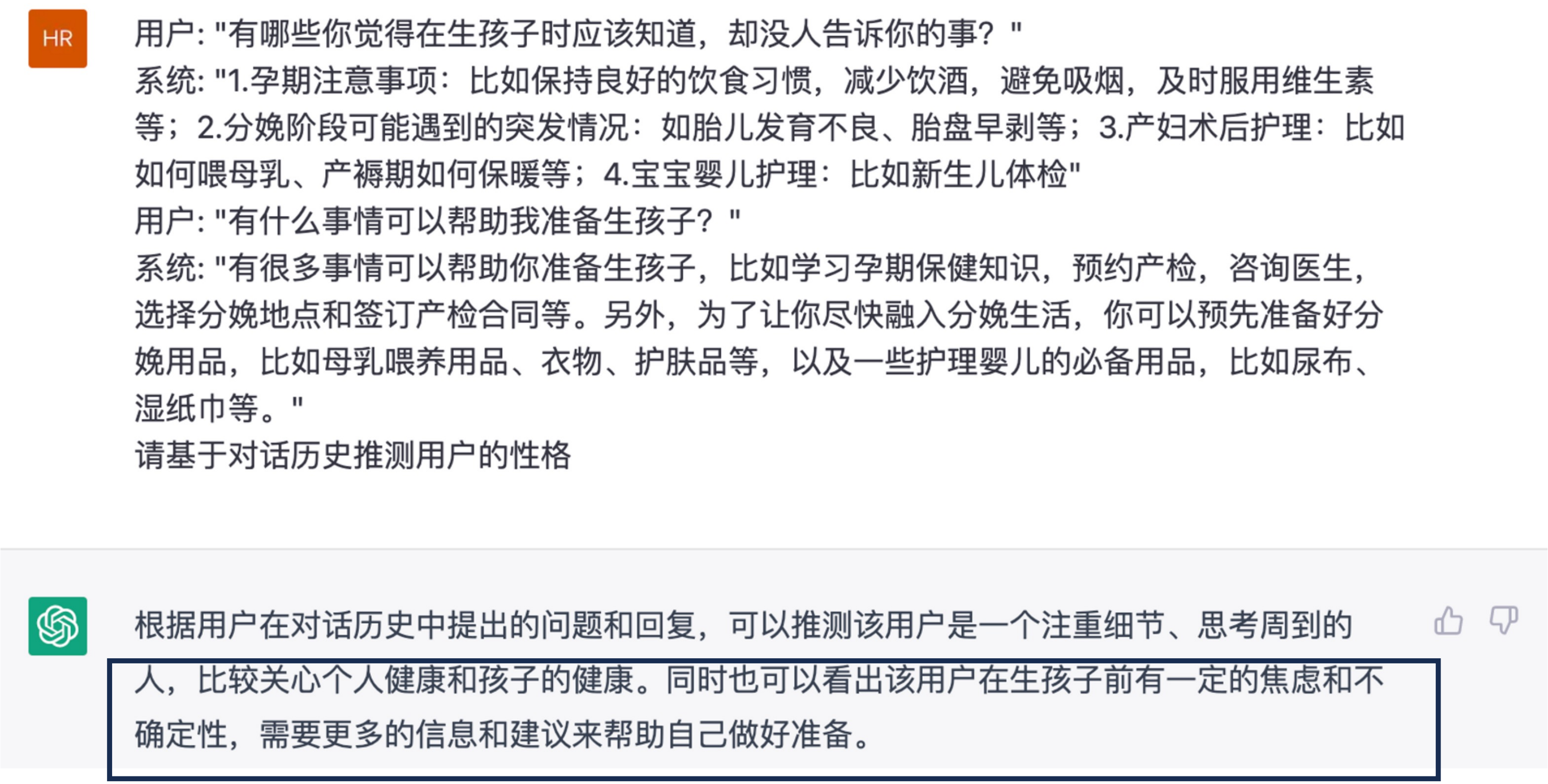
- **But:**
  - Most of them **directly mapping the dialogue context to the response**
  - Ignoring the linguistic cues underneath dialogue context such as user status, and thus lacks the ability to employ corresponding language styles and knowledge systems when dealing with users of different personalities, emotions, and psychological states.



```
{
    "dialog": [
        "有人知道黑蚁资本吗？",
        "黑蚁资本是一家中国的私募股权投资基金，专注于投资中国新兴的互联网企业，并为它们提供资本运作和战略咨询服务。",
        "黑蚁资本投资的领域是什么？",
        "黑蚁资本专注于投资中国新兴的互联网企业，并为它们提供资本运作和战略咨询服务。",
        "黑蚁资本有什么特点？",
        "黑蚁资本以其出色的投资组合、优秀的投资团队和丰富的行业经验而著称，它致力于追求最大化的投资回报，并且注重长期价值创造。"
    ],
    "user_persona": "求知欲强，关注新兴事物，注重实际问题的解决。"
},
{
```

➢ Multi-step Reasoning

- We formulate it as a **multi-step reasoning problem**, can be alleviated by CoT, which is **not explored** at dialogue.
  - **First, infer and trace the core of the user's question or concern.**
  - Then, plan a response that takes into account their status like emotional state and psychological well-being. (optional)
  - …… (optional)
  - Finally, generate a personalized dialogue response.

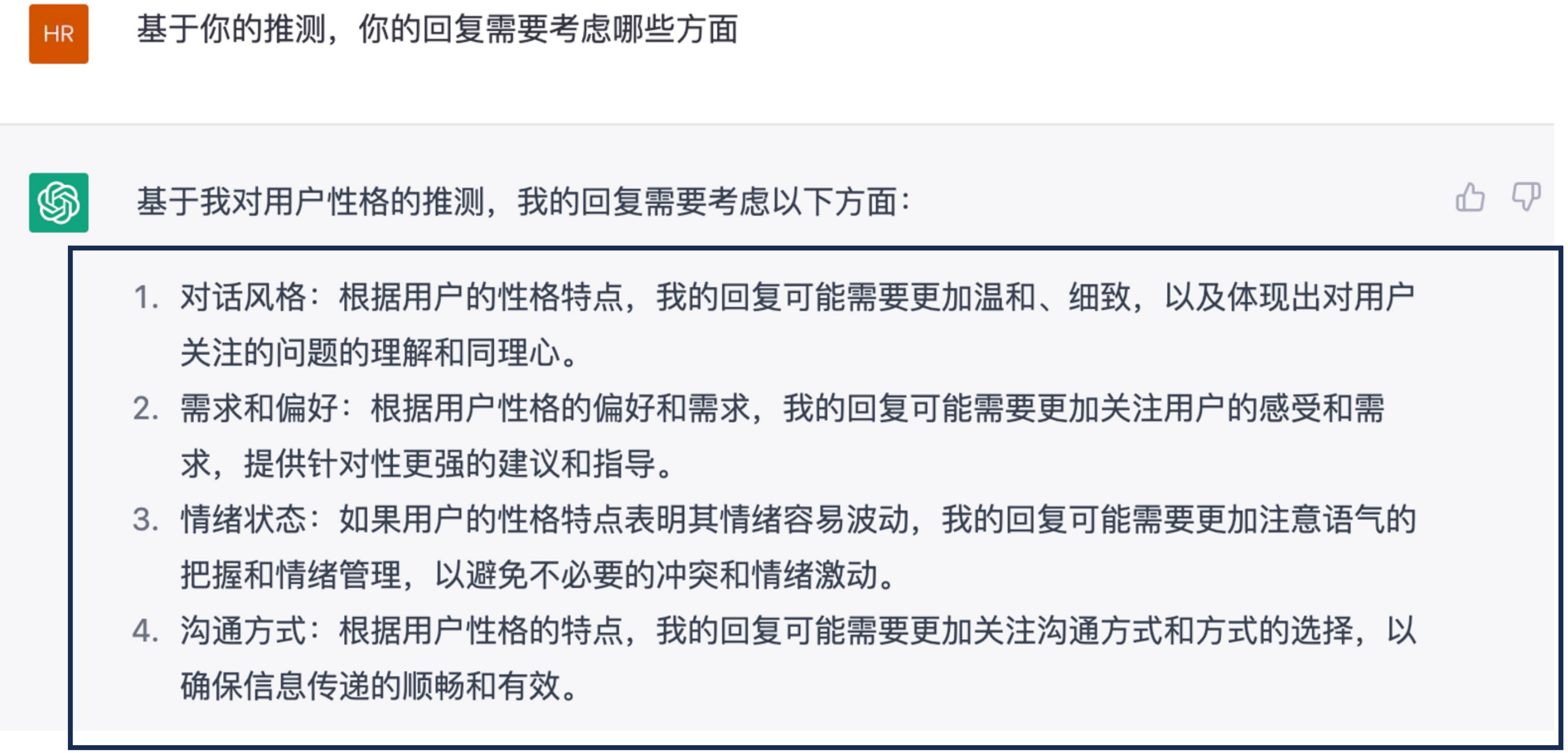


用户: "有哪些你觉得在生孩子时应该知道，却没人告诉你的事？ "
系统: "1.孕期注意事项：比如保持良好的饮食习惯，减少饮酒，避免吸烟，及时服用维生素等；2.分娩阶段可能遇到的突发情况：如胎儿发育不良、胎盘早剥等；3.产妇术后护理：比如如何喂母乳、产褥期如何保暖等；4.宝宝婴儿护理：比如新生儿体检"
用户: "有什么事情可以帮助我准备生孩子？ "
系统: "有很多事情可以帮助你准备生孩子，比如学习孕期保健知识，预约产检，咨询医生，选择分娩地点和签订产检合同等。另外，为了让你尽快融入分娩生活，你可以预先准备好分娩用品，比如母乳喂养用品、衣物、护肤品等，以及一些护理婴儿的必备用品，比如尿布、湿纸巾等。"
请基于对话历史推测用户的性格

根据用户在对话历史中提出的问题和回复，可以推测该用户是一个注重细节、思考周到的人，比较关心个人健康和孩子的健康。同时也可以看出该用户在生孩子前有一定的焦虑和不确定性，需要更多的信息和建议来帮助自己做好准备。

➢ Multi-step Reasoning

- We formulate it as a **multi-step reasoning problem**, can be alleviated by CoT, which is **not explored** at dialogue.
  - First, infer and trace the core of the user's question or concern.
  - **Then, plan a response that takes into account their status like emotional state and psychological well-being** (optional)
  - …… (optional)
  - Finally, generate a personalized dialogue response.

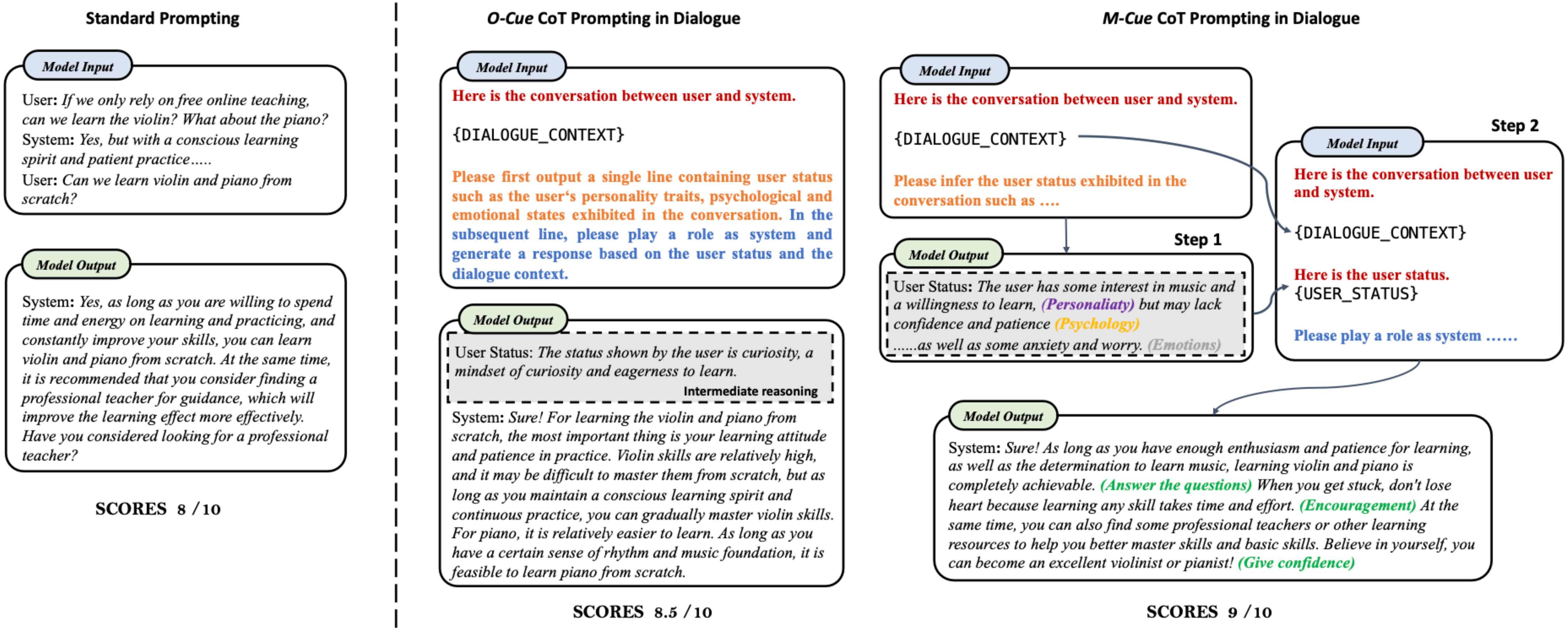


基于你的推测，你的回复需要考虑哪些方面

基于我对用户性格的推测，我的回复需要考虑以下方面：

1. 对话风格：根据用户的性格特点，我的回复可能需要更加温和、细致，以及体现出对用户关注的问题的理解和同理心。
2. 需求和偏好：根据用户性格的偏好和需求，我的回复可能需要更加关注用户的感受和需求，提供针对性更强的建议和指导。
3. 情绪状态：如果用户的性格特点表明其情绪容易波动，我的回复可能需要更加注意语气的把握和情绪管理，以避免不必要的冲突和情绪激动。
4. 沟通方式：根据用户性格的特点，我的回复可能需要更加关注沟通方式和方式的选择，以确保信息传递的顺畅和有效。

➢ Cue-CoT



Standard Prompting

Model Input

User: *If we only rely on free online teaching, can we learn the violin? What about the piano?*
System: *Yes, but with a conscious learning spirit and patient practice.....*
User: *Can we learn violin and piano from scratch?*

Model Output

System: *Yes, as long as you are willing to spend time and energy on learning and practicing, and constantly improve your skills, you can learn violin and piano from scratch. At the same time, it is recommended that you consider finding a professional teacher for guidance, which will improve the learning effect more effectively. Have you considered looking for a professional teacher?*

SCORES 8 /10

***O-Cue* CoT Prompting in Dialogue**

Model Input

Here is the conversation between user and system.

`{DIALOGUE_CONTEXT}`

Please first output a single line containing user status such as the user‘s personality traits, psychological and emotional states exhibited in the conversation. In the subsequent line, please play a role as system and generate a response based on the user status and the dialogue context.

Model Output

User Status: *The status shown by the user is curiosity, a mindset of curiosity and eagerness to learn.*

Intermediate reasoning

System: *Sure! For learning the violin and piano from scratch, the most important thing is your learning attitude and patience in practice. Violin skills are relatively high, and it may be difficult to master them from scratch, but as long as you maintain a conscious learning spirit and continuous practice, you can gradually master violin skills. For piano, it is relatively easier to learn. As long as you have a certain sense of rhythm and music foundation, it is feasible to learn piano from scratch.*

SCORES 8.5 /10

***M-Cue* CoT Prompting in Dialogue**

Model Input

Here is the conversation between user and system.

`{DIALOGUE_CONTEXT}`

Please infer the user status exhibited in the conversation such as ….

Step 1

Model Output

User Status: *The user has some interest in music and a willingness to learn, **(Personaliaty)** but may lack confidence and patience **(Psychology)** ......as well as some anxiety and worry. (Emotions)*

Step 2

Model Input

Here is the conversation between user and system.

`{DIALOGUE_CONTEXT}`

Here is the user status.
`{USER_STATUS}`

Please play a role as system ……

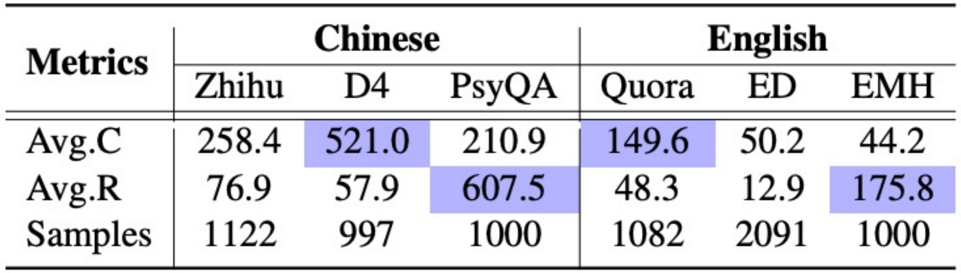
Model Output

System: *Sure! As long as you have enough enthusiasm and patience for learning, as well as the determination to learn music, learning violin and piano is completely achievable. **(Answer the questions)** When you get stuck, don't lose heart because learning any skill takes time and effort. **(Encouragement)** At the same time, you can also find some professional teachers or other learning resources to help you better master skills and basic skills. Believe in yourself, you can become an excellent violinist or pianist! **(Give confidence)***

SCORES 9 /10

Figure 1: An example of different prompting for responding to in-depth dialog questions with LLMs, including standard prompting, *O-Cue* CoT, and *M-Cue* CoT. We shadow the intermediate reasoning results, *i.e.*, the personality, empathy, and psychological status of the user, and highlight the instructions at the input and indicate the roles of different parts of the response (in green) in *M-Cue* CoT.

Benchmark



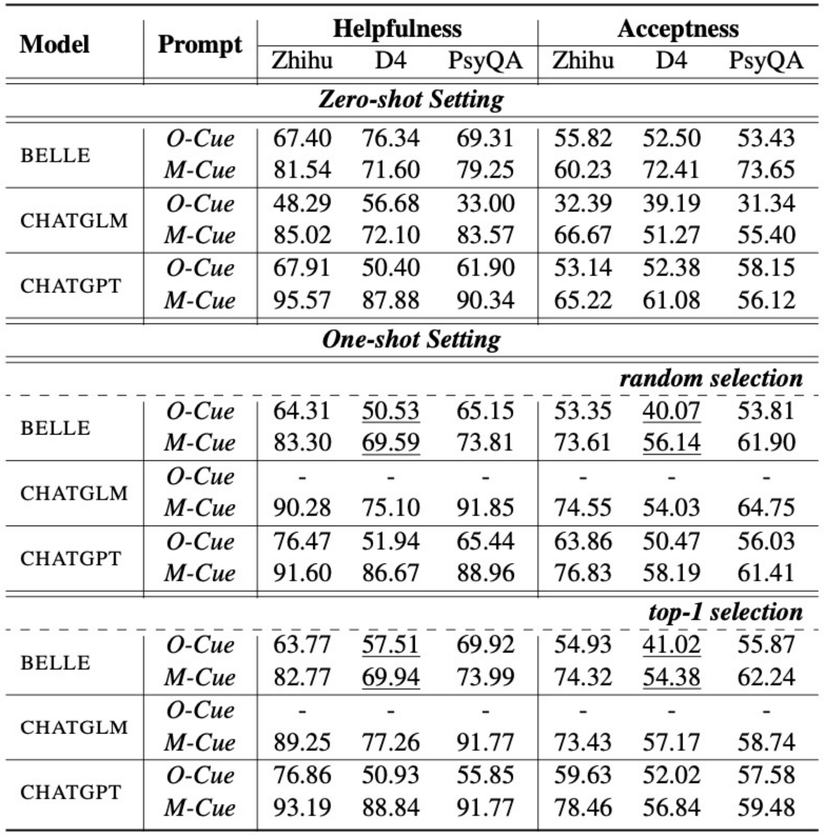
Metrics	Chinese			English		
	Zhihu	D4	PsyQA	Quora	ED	EMH
Avg.C	258.4	521.0	210.9	149.6	50.2	44.2
Avg.R	76.9	57.9	607.5	48.3	12.9	175.8
Samples	1122	997	1000	1082	2091	1000

Three major linguistic cues are considered:

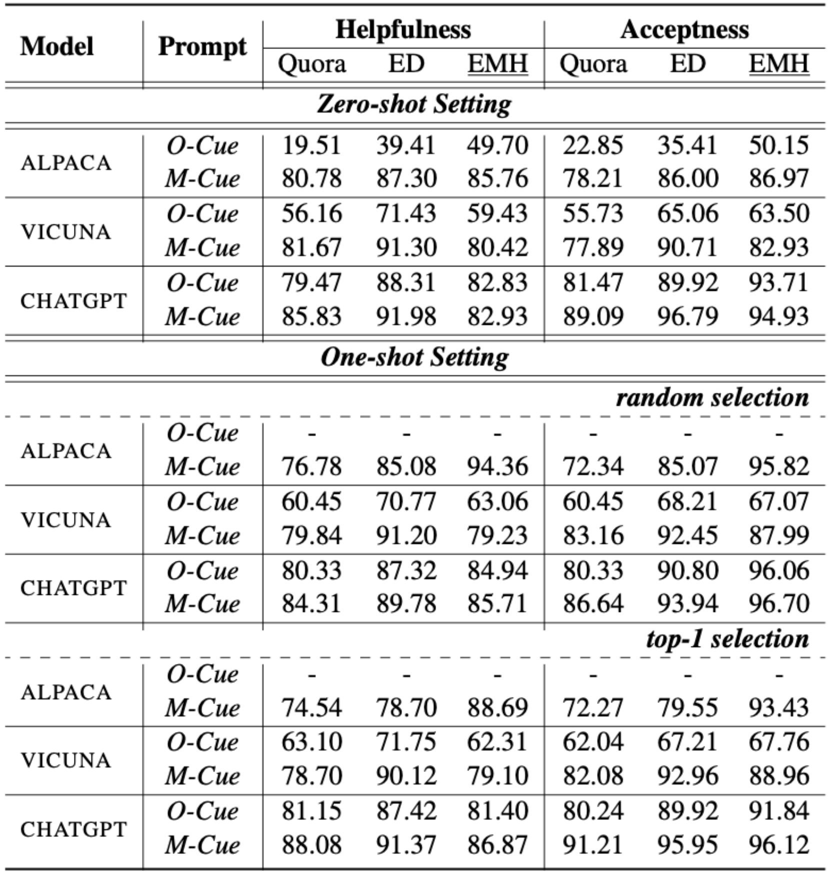
- ✓ **Personality:** Zhihu & Quora (Self-built)
- ✓ **Emotion:** D4 & ED
- ✓ **Psychology:** PsyQA & EMH

➢ Experiment

- O-Cue and M-Cue both are better than Standard Prompting, and M-Cue is more general and robust
- Chinese LLMs
  - BELLE： low **long-context understanding ability**; middle **instruction-following ability**
  - ChatGLM: middle long-context understanding ability; low instruction-following ability
  - ChatGPT: both high
- English LLMs
  - …….

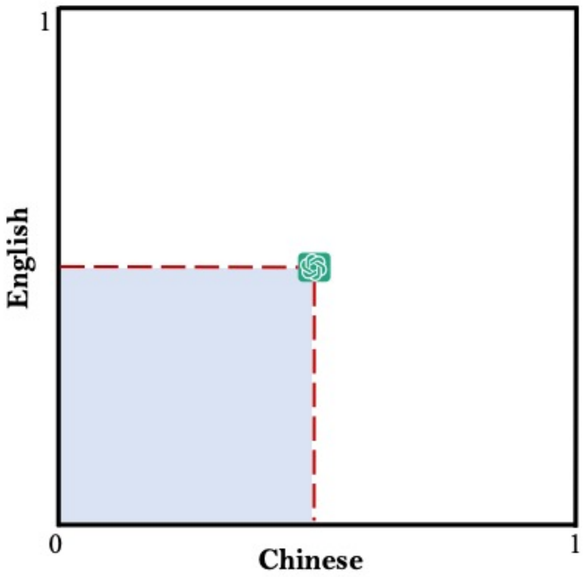


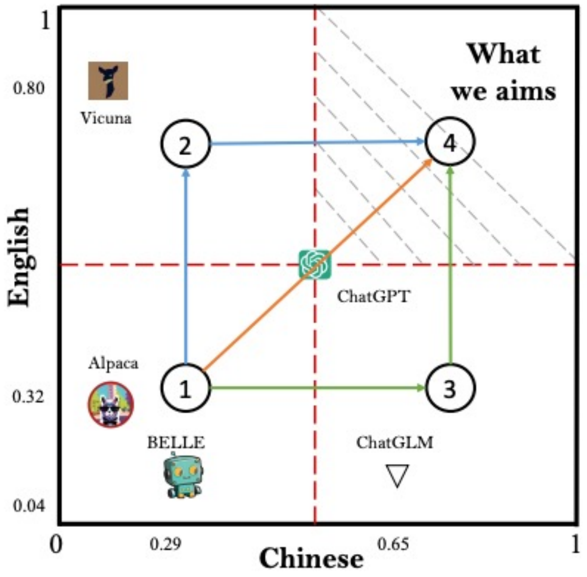
Model	**Prompt**	**Helpfulness**			**Acceptness**		
		Zhihu	D4	PsyQA	Zhihu	D4	PsyQA
Zero-shot Setting							
BELLE	*O-Cue*	67.40	76.34	69.31	55.82	52.50	53.43
	M-Cue	81.54	71.60	79.25	60.23	72.41	73.65
CHATGLM	*O-Cue*	48.29	56.68	33.00	32.39	39.19	31.34
	M-Cue	85.02	72.10	83.57	66.67	51.27	55.40
CHATGPT	*O-Cue*	67.91	50.40	61.90	53.14	52.38	58.15
	M-Cue	95.57	87.88	90.34	65.22	61.08	56.12
One-shot Setting							
random selection							
BELLE	*O-Cue*	64.31	50.53	65.15	53.35	40.07	53.81
	M-Cue	83.30	69.59	73.81	73.61	56.14	61.90
CHATGLM	*O-Cue*	-	-	-	-	-	-
	M-Cue	90.28	75.10	91.85	74.55	54.03	64.75
CHATGPT	*O-Cue*	76.47	51.94	65.44	63.86	50.47	56.03
	M-Cue	91.60	86.67	88.96	76.83	58.19	61.41
top-1 selection							
BELLE	*O-Cue*	63.77	57.51	69.92	54.93	41.02	55.87
	M-Cue	82.77	69.94	73.99	74.32	54.38	62.24
CHATGLM	*O-Cue*	-	-	-	-	-	-
	M-Cue	89.25	77.26	91.77	73.43	57.17	58.74
CHATGPT	*O-Cue*	76.86	50.93	55.85	59.63	52.02	57.58
	M-Cue	93.19	88.84	91.77	78.46	56.84	59.48



Model	**Prompt**	**Helpfulness**			**Acceptness**		
		Quora	ED	EMH	Quora	ED	EMH
Zero-shot Setting							
ALPACA	*O-Cue*	19.51	39.41	49.70	22.85	35.41	50.15
	M-Cue	80.78	87.30	85.76	78.21	86.00	86.97
VICUNA	*O-Cue*	56.16	71.43	59.43	55.73	65.06	63.50
	M-Cue	81.67	91.30	80.42	77.89	90.71	82.93
CHATGPT	*O-Cue*	79.47	88.31	82.83	81.47	89.92	93.71
	M-Cue	85.83	91.98	82.93	89.09	96.79	94.93
One-shot Setting							
random selection							
ALPACA	*O-Cue*	-	-	-	-	-	-
	M-Cue	76.78	85.08	94.36	72.34	85.07	95.82
VICUNA	*O-Cue*	60.45	70.77	63.06	60.45	68.21	67.07
	M-Cue	79.84	91.20	79.23	83.16	92.45	87.99
CHATGPT	*O-Cue*	80.33	87.32	84.94	80.33	90.80	96.06
	M-Cue	84.31	89.78	85.71	86.64	93.94	96.70
top-1 selection							
ALPACA	*O-Cue*	-	-	-	-	-	-
	M-Cue	74.54	78.70	88.69	72.27	79.55	93.43
VICUNA	*O-Cue*	63.10	71.75	62.31	62.04	67.21	67.76
	M-Cue	78.70	90.12	79.10	82.08	92.96	88.96
CHATGPT	*O-Cue*	81.15	87.42	81.40	80.24	89.92	91.84
	M-Cue	88.08	91.37	86.87	91.21	95.95	96.12

➢ Discussion & Future Work





- ✓ **Path 1 (orange):** Directly extend the capability of ChatGPT. **GPT-4, Claude, Bard, ……**
- ✓ **Path 2 (green):** Continually train current Chinese LLMs on more multi-lingual corpus.
- ✓ **Path 3 (blue):** Continually train current English LLMs on more multi-lingual corpus. **Chinese-LLaMA, Chinese-Vicuna, …..**

More analysis, case studies can be found in the paper!
Demo can be accessed in the Github!!!

Conclusion

- We build a benchmark to evaluate the **helpfulness** and **acceptability** of responses generated by current LLMs, considering **three major linguistic cues** of user statuses.

- We leverage the strong reasoning capability of LLMs by proposing a **Cue-CoT** to trace the status of users, decomposing the response generation into multiple reasoning steps. It can be applied into different downstream tasks and applications with **multiple intermediate results** as demonstration selection criterion.

- Experimental results demonstrate the superior performance of our method on **6 datasets under both zero-shot and one-shot settings**. We hope the release of our work can shed some light on the evaluation and development of LLMs.

Thanks.

Hongru WANG

https://rulegreen.github.io/

Code & Benchmark **Homepage** **PaperWeekly Report**