Large Language Models as Source Planner for Personalized Knowledge-grounded Dialogues

Hongru Wang, Minda Hu, Yang Deng, Fei Mi, Weichao Wang

Yasheng Wang, Wai-chung Kwan, Irwin King, Kam-Fai Wong

MoE Key Laboratory of High Confidence and Software Technologies

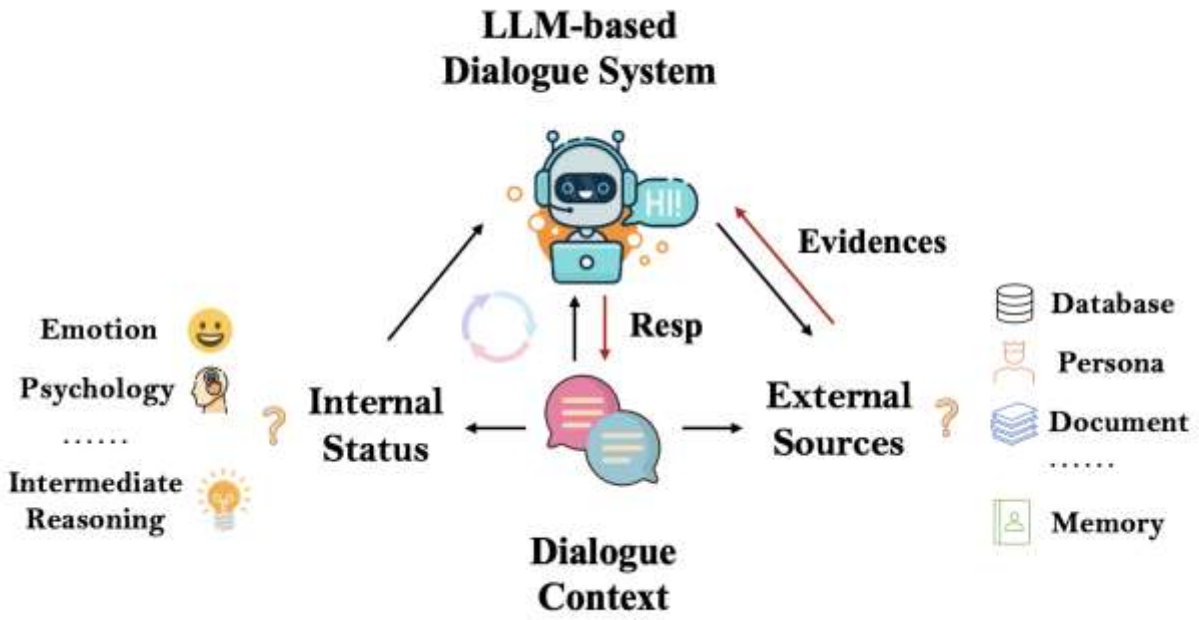
Department of Systems Engineering and Engineering Management

Department of Computer Science and Engineering

The Chinese University of Hong Kong

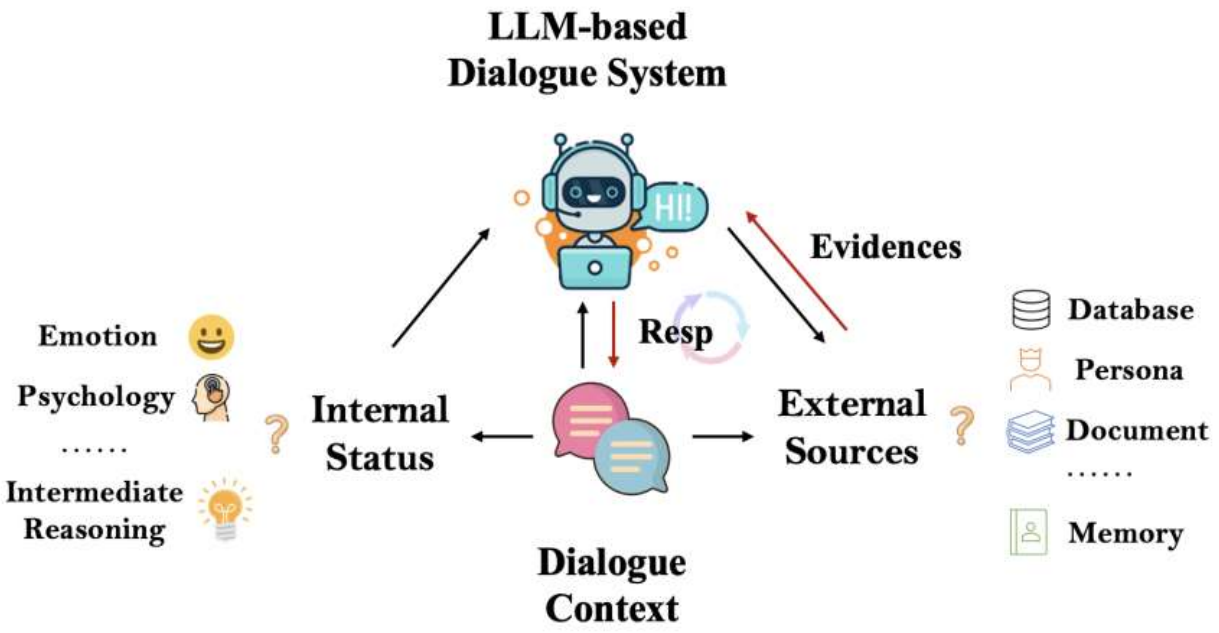
➢ What's LLM-based Dialogue System ?

- Internal Reasoning
  - Prompting engineering
    - Linguistic cues
      - Emotion
      - Psychology
      - ……
    - Intermediate reasoning
    - ……



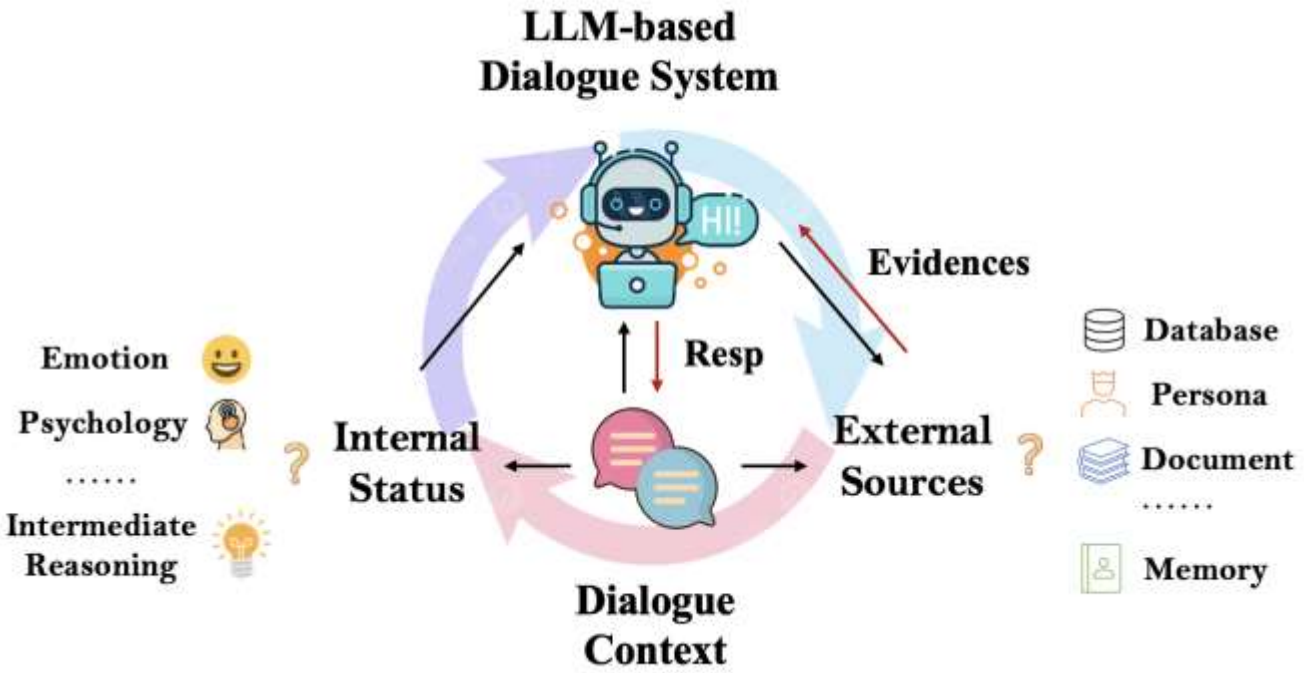
➢ What's LLM-based Dialogue System ?

- Internal Reasoning
  - Prompting engineering
    - Linguistic cues
      - Emotion
      - Psychology
      - ……
    - Intermediate reasoning
    - ……
- External Acting
  - Planning the actions/interactions
    - Generate tokens --- most fine-grained action
    - ……



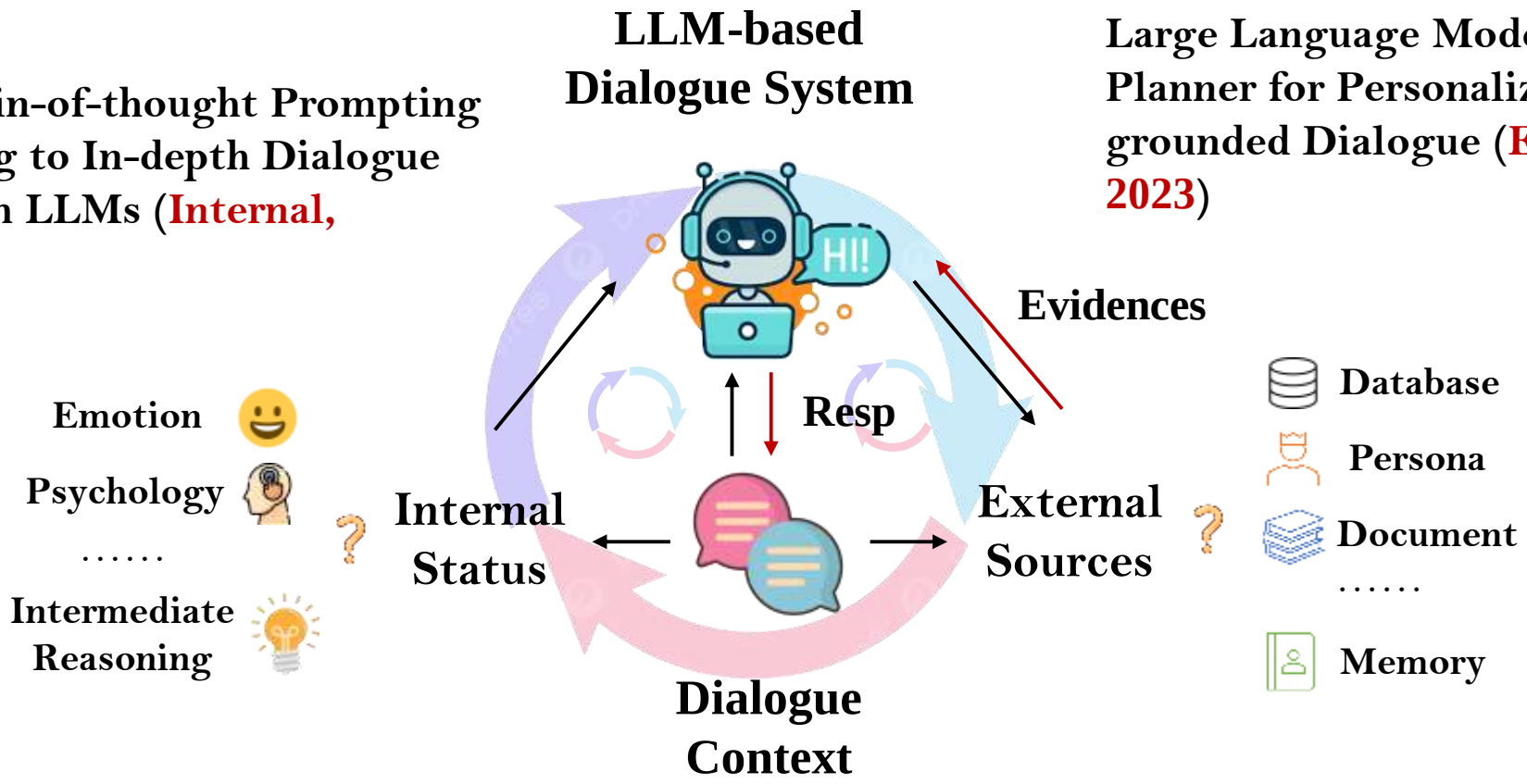
➢ What's LLM-based Dialogue System ?

- Internal Reasoning
  - Prompting engineering
    - Linguistic cues
      - Emotion
      - Psychology
      - ……
    - Intermediate reasoning
    - ……
- External Acting
  - Planning the actions/interactions
    - Generate tokens –– most fine-grained action
    - ……
- Reasoning + Acting
  - TPE multi-persona collaboration framework
    - Thinker (Reasoning)
    - Planner (Planning)
    - Executor (Acting)



➢ What's LLM-based Dialogue System ?

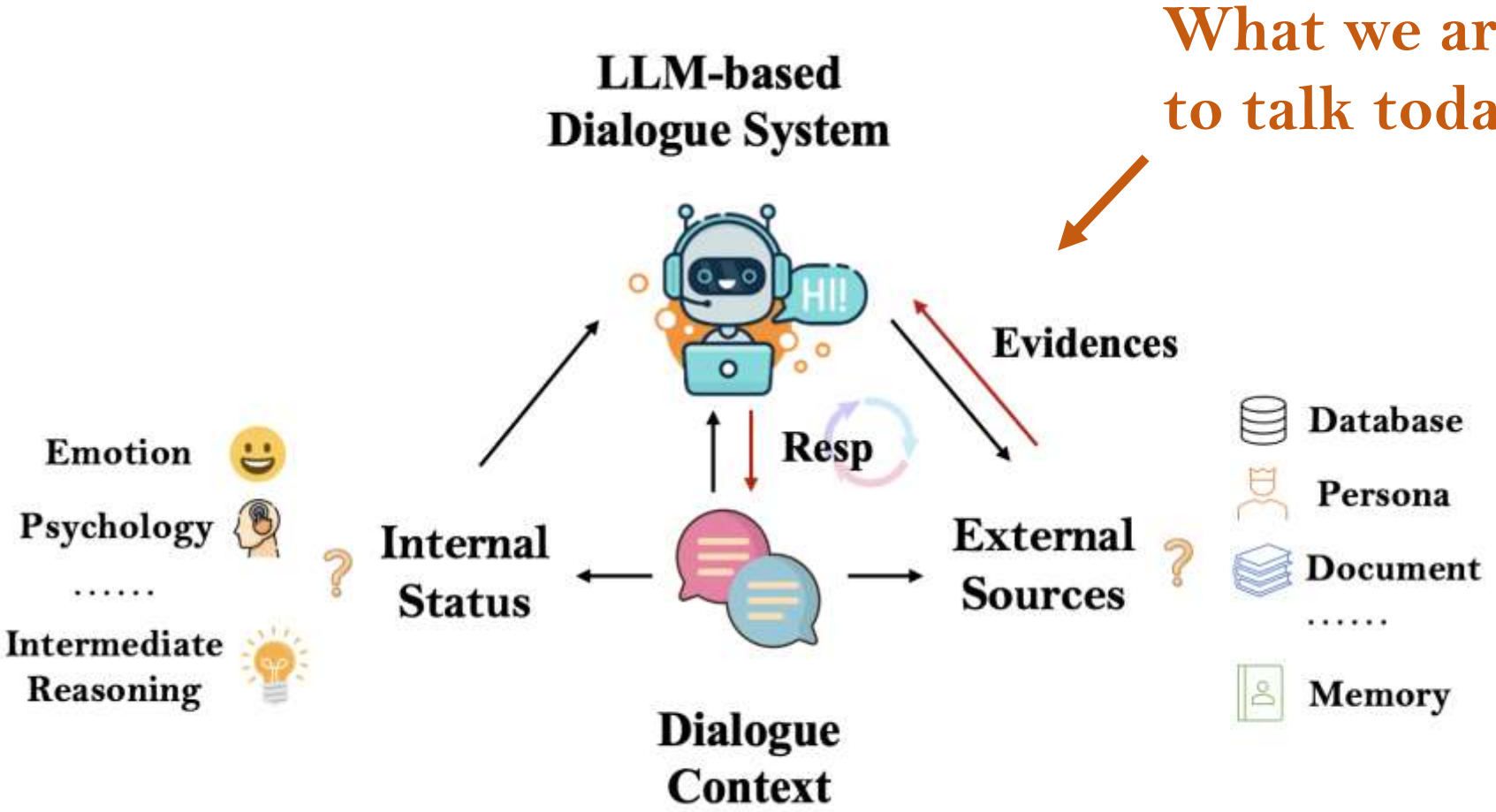
Cue-CoT: Chain-of-thought Prompting for Responding to In-depth Dialogue Questions with LLMs (Internal, EMNLP 2023)



Large Language Models as Source Planner for Personalized Knowledge-grounded Dialogue (External, EMNLP 2023)

TPE: Towards Better Compositional Reasoning over Conceptual Tools with Multi-persona Collaboration (Internal with External)

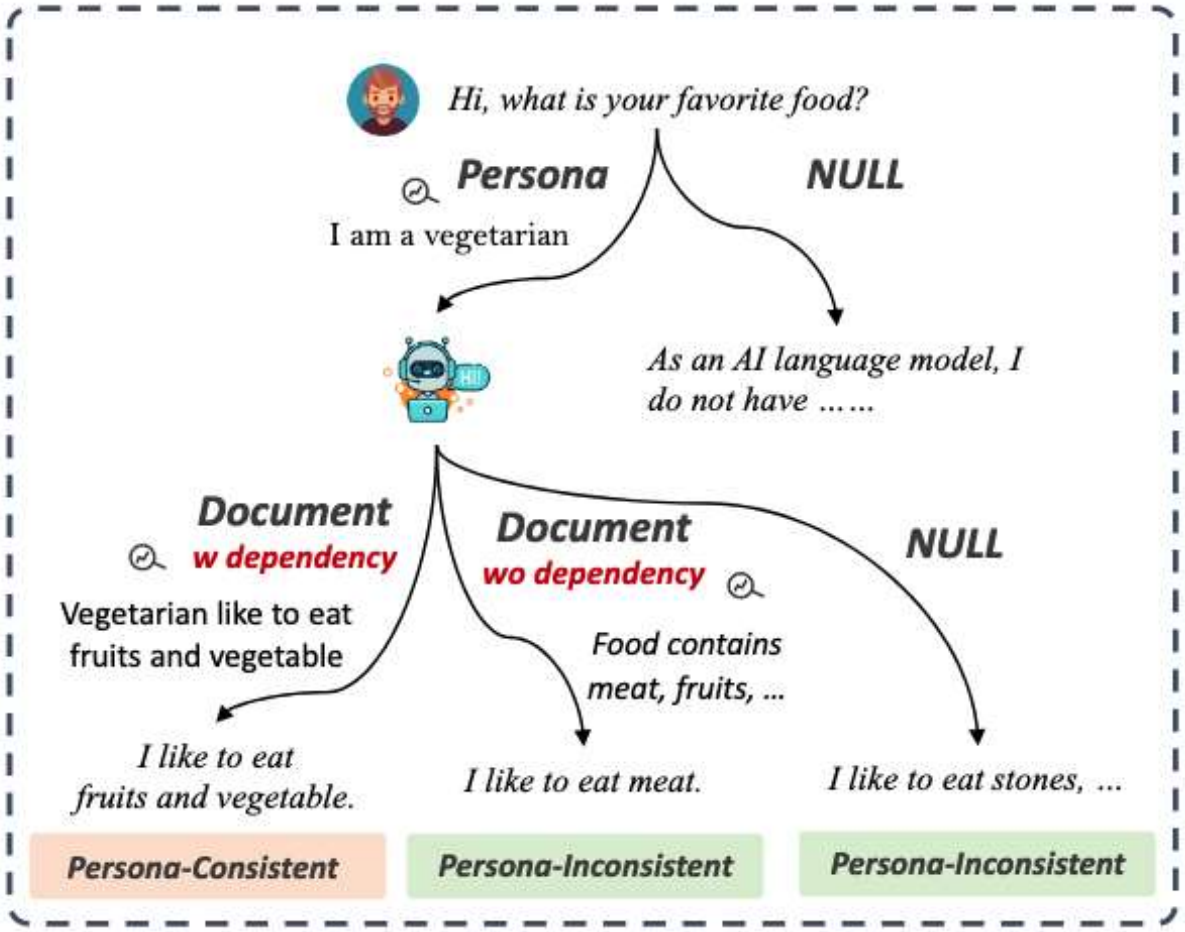
➢ The External Capability of LLM-based Dialogue System



➢ What is the external capability ?

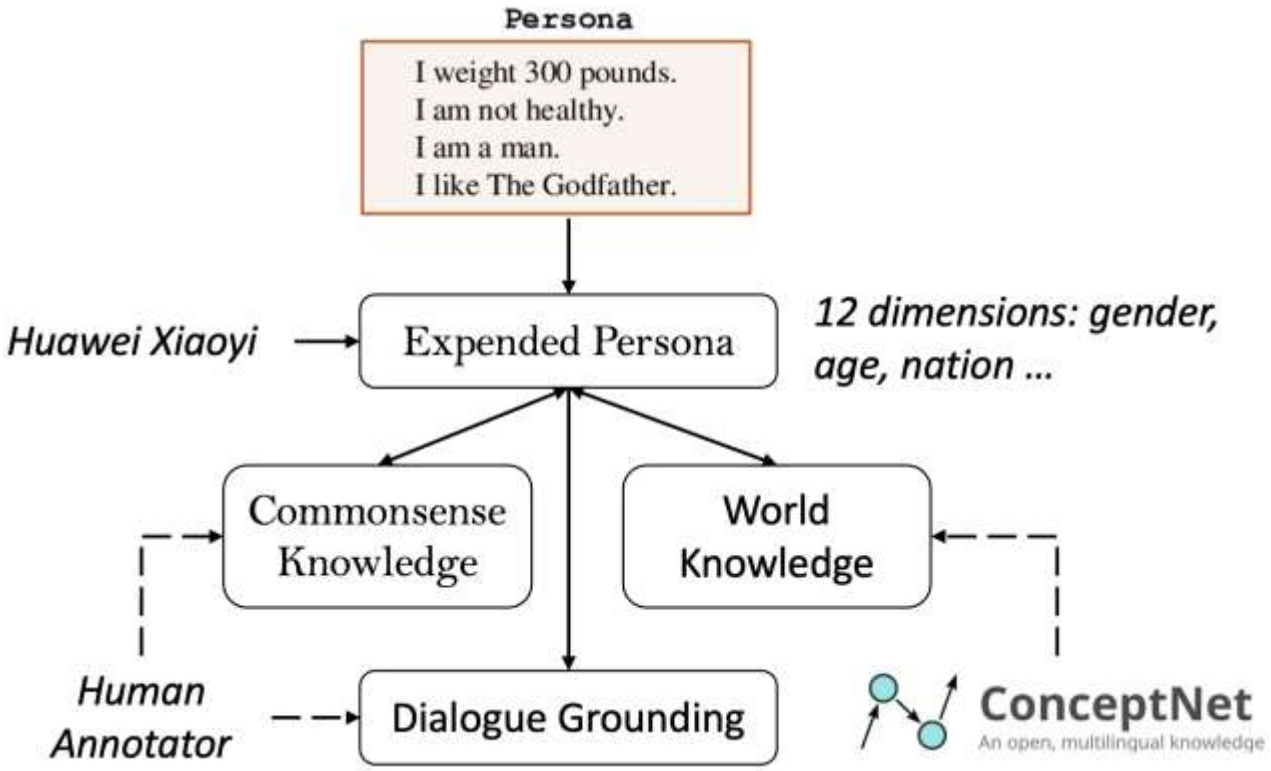
- Open-domain Dialogue System requires access to various external knowledge sources to deliver **reliable, informative, personalized, and helpful responses**, depending on which sources are invoked.
- Indiscriminately incorporating all sources bring **unnecessary computing cost**, and sometimes it does not require external knowledge.
- The **interdependence** between different external sources brings new challenges, while ignoring the complex relationship between different sources, leads to sub-optimal performance.

➢ A case of dependency between different knowledge sources



Dependency between Multiple Sources

➢ Knowledge Behind Persona (KBP) Dataset

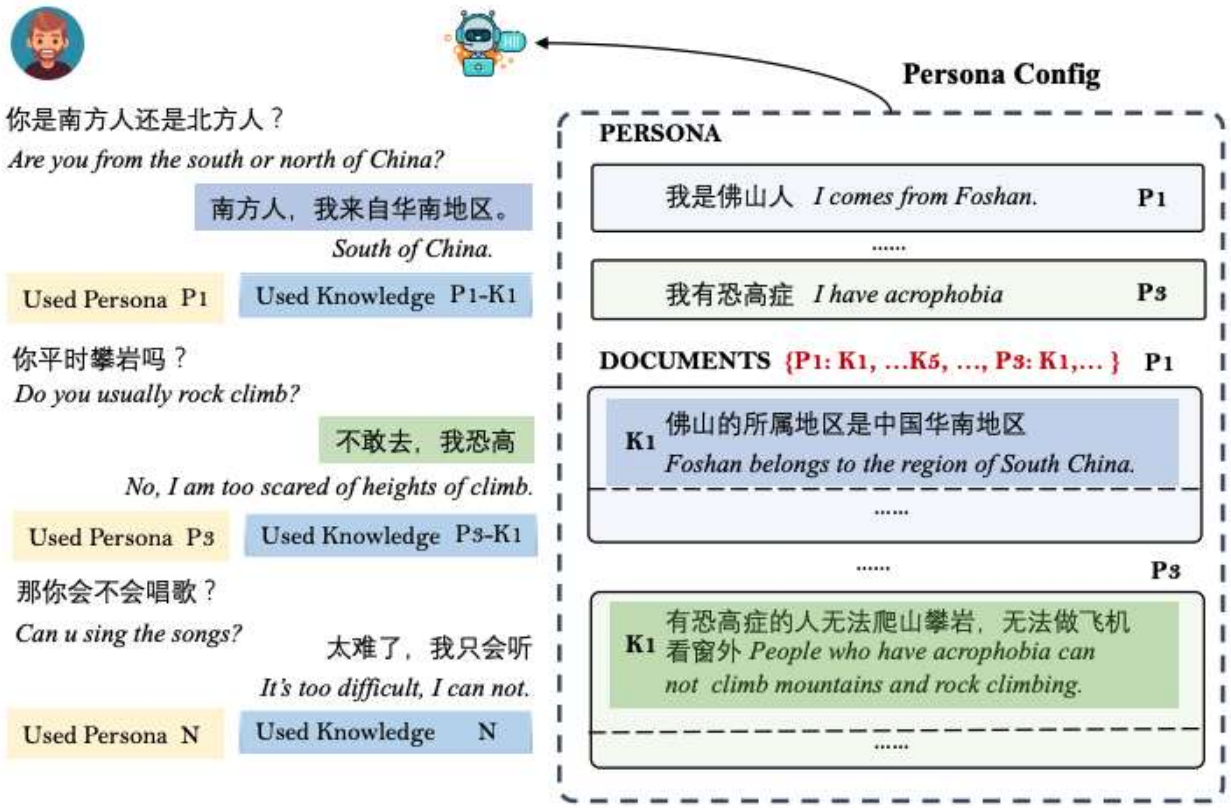


- **Step 1**: Persona and Knowledge Acquisition
- **Step 2**: Dialogue Collection

➢ Knowledge Behind Persona (KBP) Dataset

There are three situations in KBP:

- Do not need any external knowledge → NULL
- Only require PERSONA source of knowledge → PERSONA
- Require both PERSONA and DOCUMENT sources of knowledge → PERSONA DOCUMENT



➢ SAFARI Framework

Figure 1: An unified framework of the source-augmented dialogue system, where the response generation requires various sources of knowledge: persona, knowledge, and memory. **Planning**, **Retrieval** and **Assemble** steps are divided by dashed lines.

- **Planning:** make **a series of decision** to determine whether or not use knowledge, which and when.

$$\mathcal{M} : \boldsymbol{c} \rightarrow K_i, K_j, ..., K_n \quad or \quad \texttt{NULL}, \quad (1)$$

- **Retrieval:** retrieve *top-n* results from local databases according to the decided used source knowledge

$$\mathcal{R} : K_i, K_j, ..., K_n \rightarrow k_i^j, ..., k_n^m \qquad (2)$$

- **Assembling:** incorporate all retrieved middle results into the final response generation

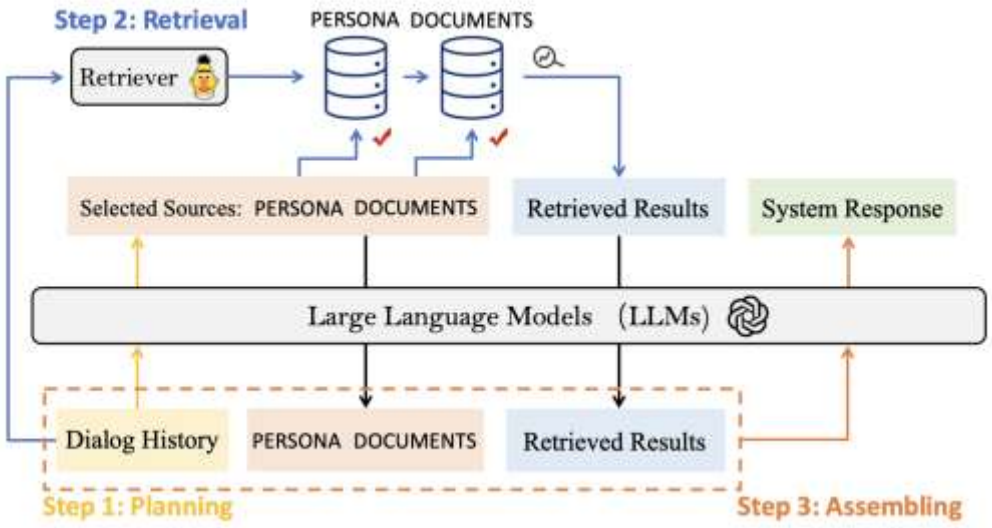
$$\mathcal{M} : Inp \rightarrow s_t, \qquad (3)$$

where $Inp = \{\boldsymbol{c}$ [SOURCE] $K_i, ..., K_n$ [EOS] [MIDDLE] $k_i^j, ..., k_n^m$ [EOM] $\}$.

!!! SAFARI has a great scalability and flexibility, such as more sources of knowledge situation.

➢ SAFARI Framework

- Supervised SAFARI **(End-to-end Training, LoRA)**
- Unsupervised SAFARI



Supervised **SAFARI**

There are two knowledge bases storing relevant information:
PERSONA: {PERSONA_DESC}
DOCUMENT: {DOCUMENT_DESC}
There exists a dependency between these knowledge bases. The invocation of DOCUMENT relies on the results from PERSONA. Please ensure the correct order of invoking them.
Here is the dialogue between the user and the system:
{DIALOGUE}
Based on the user's last question, please determine if it requires invoking the corresponding knowledge base. If the invocation is necessary, output the names of the knowledge bases in the order they should be invoked. If no invocation is needed, output "NULL".

Table 2: The zero-shot prompt of unsupervised SAFARI at planning step (translated from Chinese to English)

The dialogue is as follows:
{DIALOGUE}

Here is the system's persona and related domain knowledge:
{MIDDLE_RESULTS}

Please play the role of the system and generate a reply according to the context of the dialogue and given knowledge. Please make sure your reply is consistent with the given persona and related domain knowledge. If the provided knowledge is NULL, generate a response solely based on the dialogue context.
System:

Table 3: The zero-shot prompt of unsupervised SAFARI at assembly step (translated from Chinese to English)

Unsupervised **SAFARI**

➢ SAFARI Framework

- Performance of **Planning**
  - **Supervised ChatGLM > Supervised BELLE > Unsupervised ChatGPT > Others**
- Performance of **Retrieval**
  - **DPR > RocketQAv2 > BM25**
- Performance of **Assembling**
  - **Supervised BELLE > Supervised ChatGLM**

Model	NULL	Persona	Both
Supervised			
BELLE-LLAMA-7B-2M	42.67 (194)	14.08 (17)	83.77 (1018)
CHATGLM-6B	**47.10** (129)	**31.96** (69)	**86.59** (1031)
Unsupervised			
Zero-shot			
BELLE-LLAMA-7B-2M	**28.55** (940)	8.94 (54)	32.47 (235)
CHATGLM-6B	25.60 (1225)	0.0 (0)	0.43 (4)
CHATGPT	11.45 (116)	**20.67** (233)	**74.88** (880)
In-context			
BELLE-LLAMA-7B-2M	9.22 (36)	18.21 (1193)	0.0 (0)
CHATGLM-6B	25.67 (1190)	1.49 (9)	4.62 (30)
CHATGPT	**27.95** (699)	**23.14** (238)	**41.98** (292)

Table 4: The F1 of different decisions in **Planning** of different LLMs under supervised/unsupervised settings. We also report the frequency of different decisions in the bracket. There are 181 NULL, 125 PERSONA and 923 PERSONA, and DOCUMENTS in the ground planning.

Model	Persona	Both		
		PERSONA	DOCUMENTS	DOCUMENTS†
BM25	36.80	48.97	15.05	11.37
RocketQAv2	80.00	92.31	50.49	35.75
DPR	**83.20**	**93.07**	**51.67**	**39.33**

Table 5: The performance of **Retrieval** of different types of retrievers. There are 125 examples that only require PERSONA and 923 require both PERSONA and KNOWLEDGE. We also report the Recall@1 of DOCUMENTS without dependency (DOCUMENTS†).

Model	BLEU1	Rouge-L	P.C	K.C
Supervised Setting				
BELLE-LLAMA-7B-2M	**30.48**	**34.61**	75.34	**46.62**
CHATGLM-6B	23.81	26.70	**76.99**	42.39
Unsupervised Setting				
Zero-shot				
BELLE-LLAMA-7B-2M	11.84	19.24	30.59	27.34
CHATGLM-6B	6.18	14.50	14.73	24.73
CHATGPT	12.06	24.44	**73.47**	**38.00**
In-context				
BELLE-LLAMA-7B-2M	**19.51**	22.25	72.98	24.89
CHATGLM-6B	13.74	19.69	16.92	24.89
CHATGPT	16.03	**25.62**	46.38	35.56

Table 6: The performance of **Assembling** under supervised/unsupervised settings.

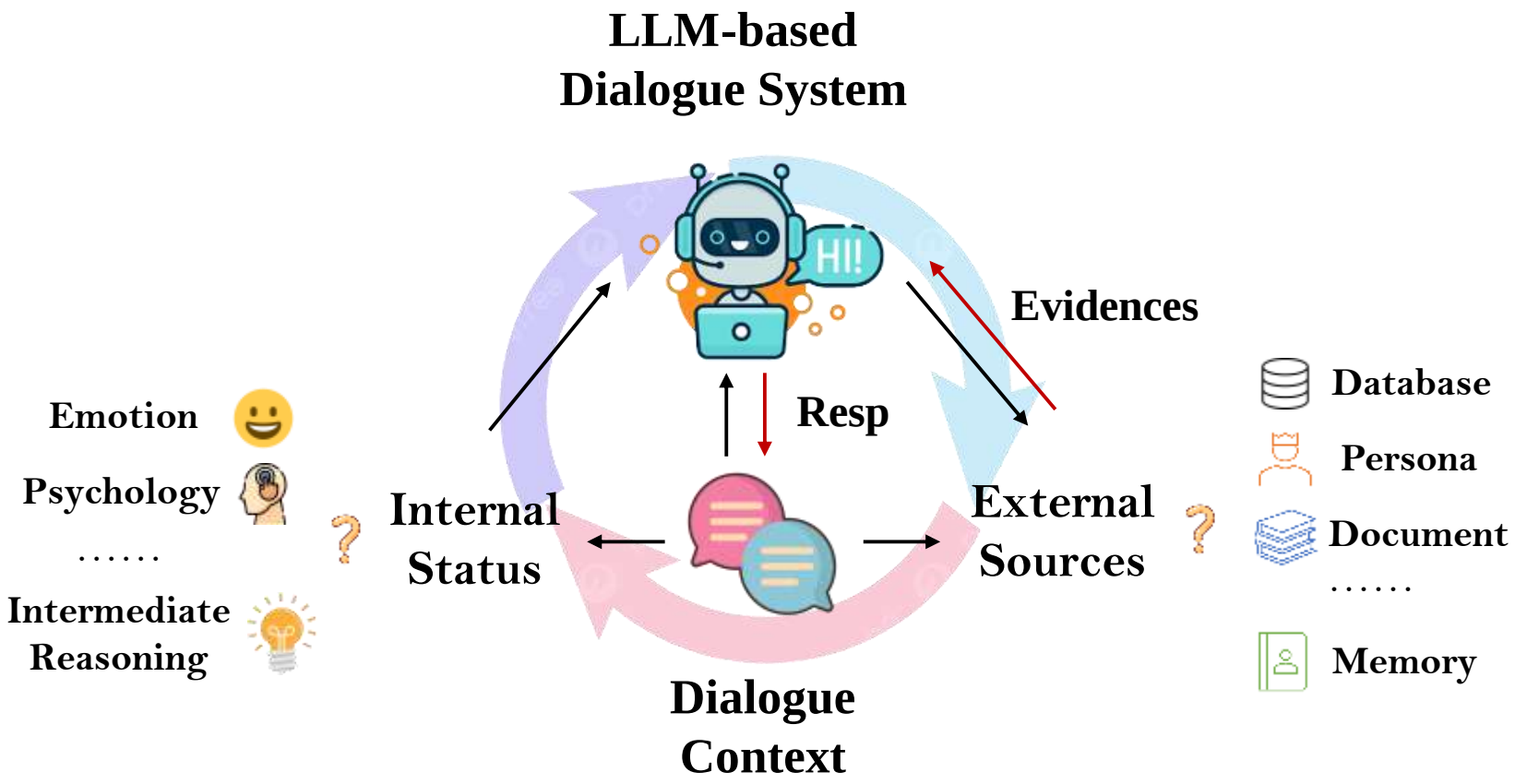
➢ Conclusions

- We are the first to augment the dialogue system to plan and incorporate multiple sources of knowledge into responses **(e.g., decide whether or not require knowledge, which source to call, and when to call).**

- We build a personalized knowledge-grounded dialogue dataset, KBP , where the responses are conditioned on multiple sources of knowledge with **dependency relationship.**

- We conduct lots of experiments and analysis on latest LLMs. More **ablation studies** can be found in the paper.

➢ Future Work



Check our latest paper!!!

[TPE: Towards Better Compositional Reasoning over Conceptual Tools with Multi-persona Collaboration]

Thanks.

Hongru WANG

https://rulegreen.github.io/

Code & Benchmark

Homepage

PaperWeekly Report