LREC-COLING 2024

BP4ER: Bootstrap Prompting for Explicit Reasoning in Medical Dialogue Generation

Yuhong He[1,6], Yongqi Zhang[2], Shizhu He[3,4] and Jun Wan[1,3,5,*]
[1] Macau University of Science and Technology, Macao, China
[2] 4Paradigm Inc., Beijing, China
[3] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[4] The Lab of Cognition and Decision Intelligence for Complex Systems, CASIA, China
[5] MAIS, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[6] Zhongkai University of Agriculture and Engineering, Guangzhou, China
yuhonghe.ai@gmail.com, yzhangee@connect.ust.hk, shizhu.he@nlpr.ia.ac.cn, jun.wan@ia.ac.cn

Contents

Background

1. What is medical dialogue generation?

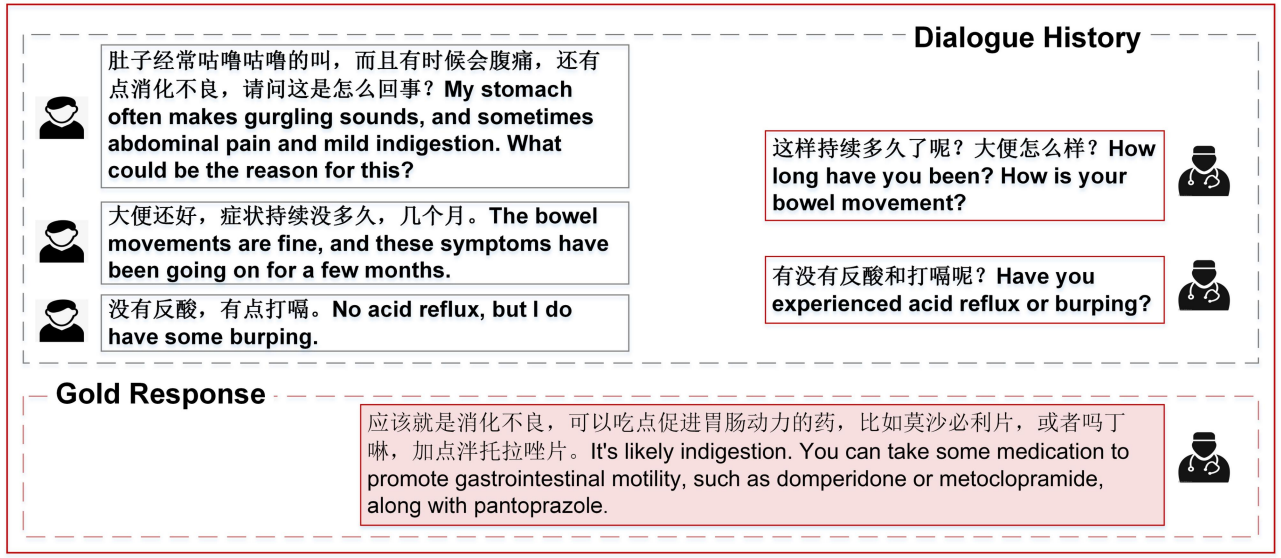


Figure 1 Medical dialogue sample.

Medical dialogue system (MDS): offering accessible medical services such as health consultations, diagnosis, and prescription, to a broader population.

Medical dialogue generation (MDG): crucial part in MDS, generating accurate medical responses based on given dialogue histories.

Background

2. Previous Works

Seq2Seq models: process dialogue context and annotated medical entities, utilizing pre-trained text encoders and decoders to generate medical responses.

Challenges (Motivation)

(1) Lack of process explanation.
(2) Requirement for large-scale annotations.

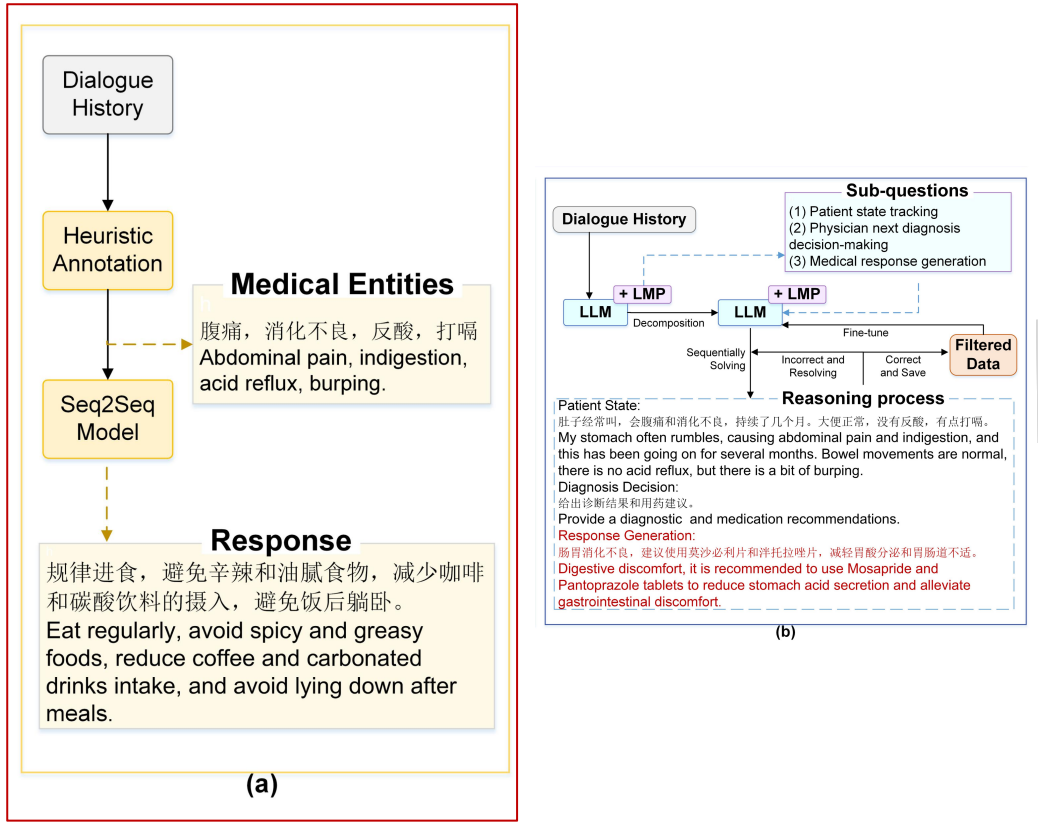


Figure 2 Seq2Seq models (a) and our model (b)

Background

3. Our Model (BP4ER)

Lack of process explanation

Requirement for large-scale annotations

Treating MDG as a multi-step reasoning problem eliminates the necessity for entity annotation by explicitly depicting its reasoning process.

Breaking MDG into interrelated sub-questions;

Facilitates the reasoning and corrects intermediate errors autonomously through the iterative approach

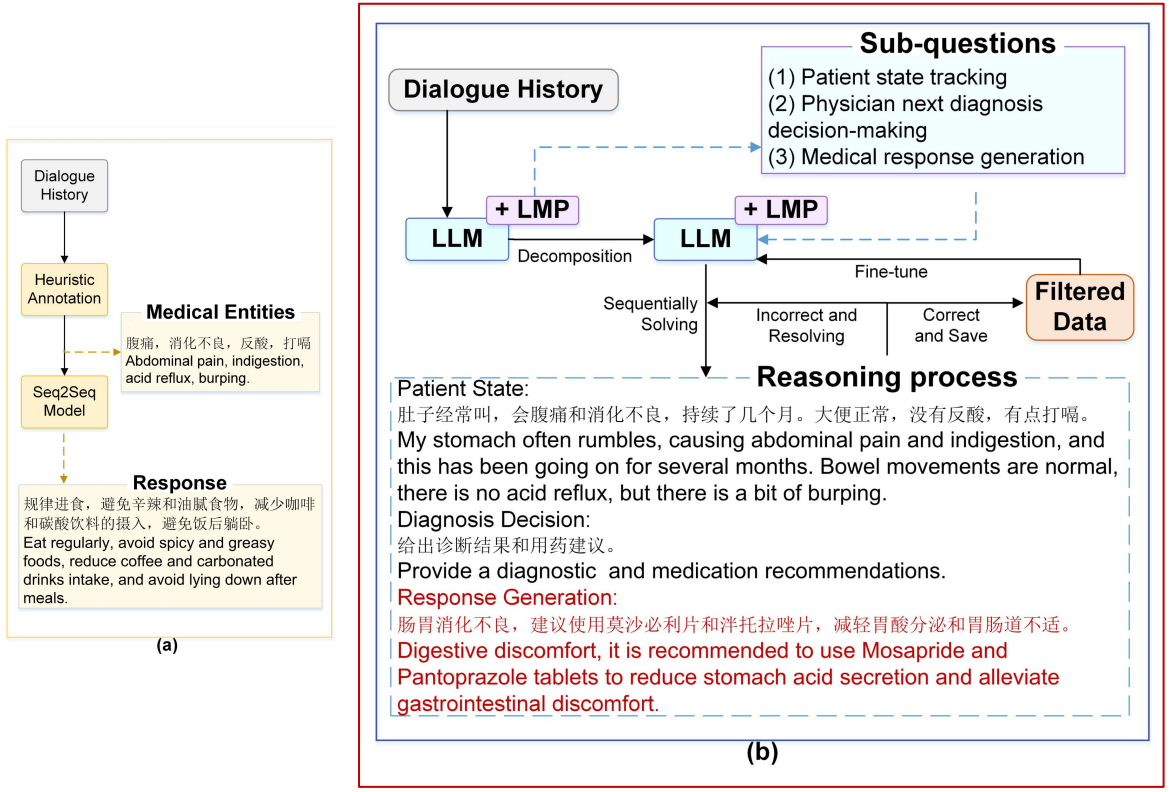


Figure 2 Seq2Seq models (a) and our model (b)

BP4ER

1. Multi-step Reasoning

2. Explicit Reasoning Process

3. Bootstrap Prompting
   - Answer-Providing Bootstrapping
   - Prompt-Revising Bootstrapping

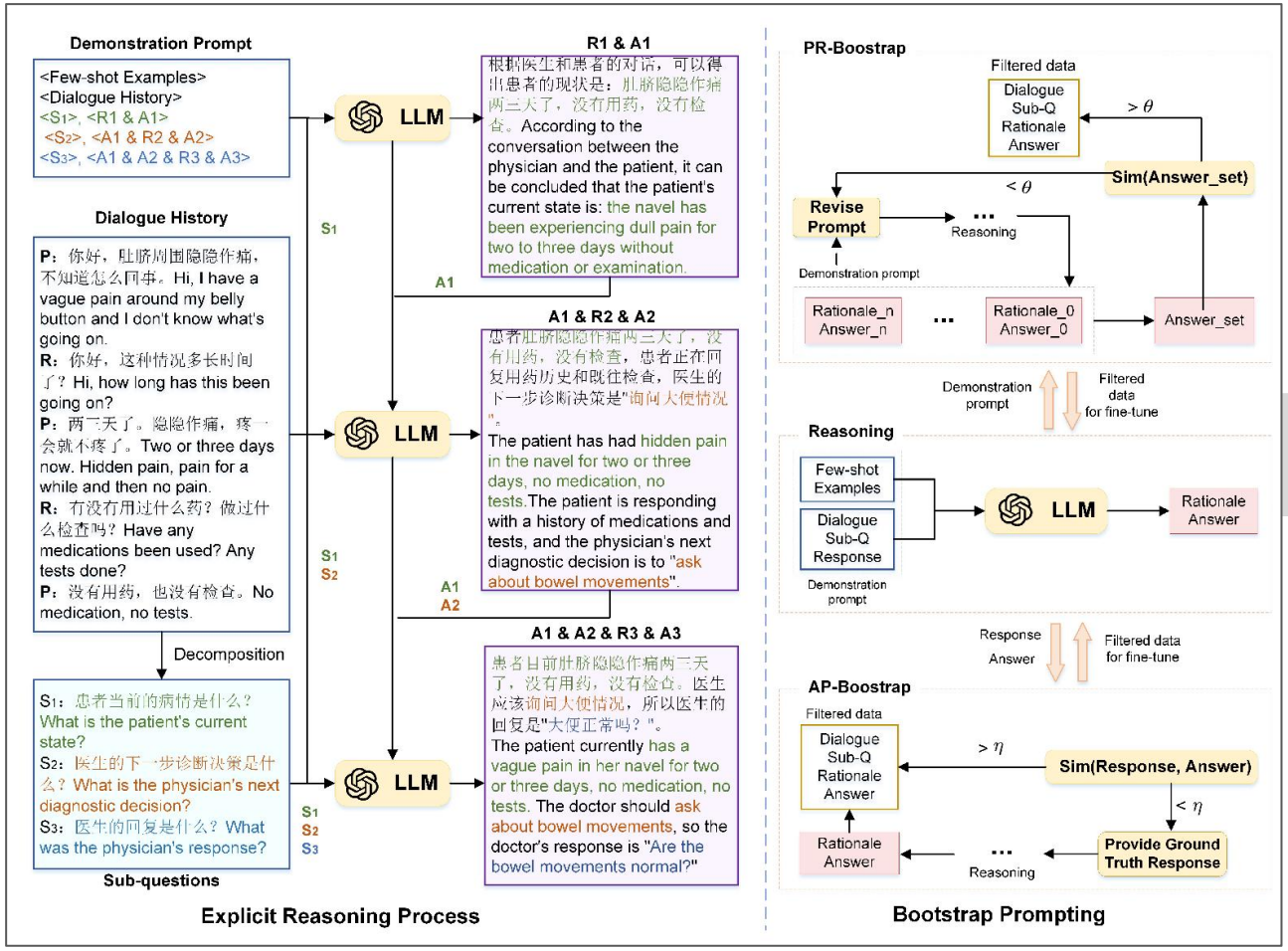


Figure 3 Our model BP4ER

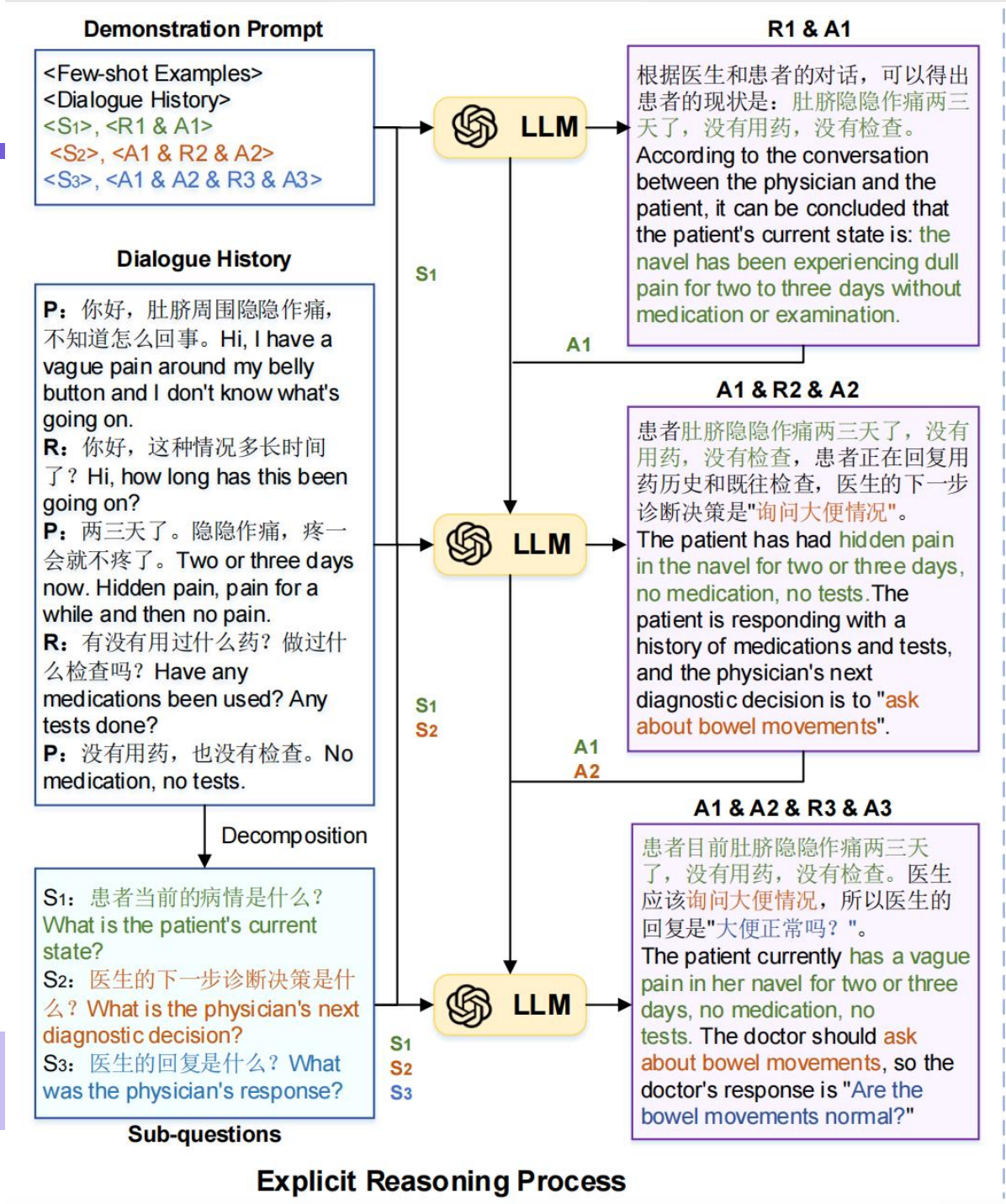
BP4ER

1. Multi-step Reasoning

- Patient State Tracking
- Next Diagnosis Decision-making
- Medical Response Generation:

2. Explicit Reasoning Process

- S1: What's the patient's current state?
- S2: What's the physician's next decision?
- S3: What's the physician's response?



BP4ER

3. Bootstrap Prompting

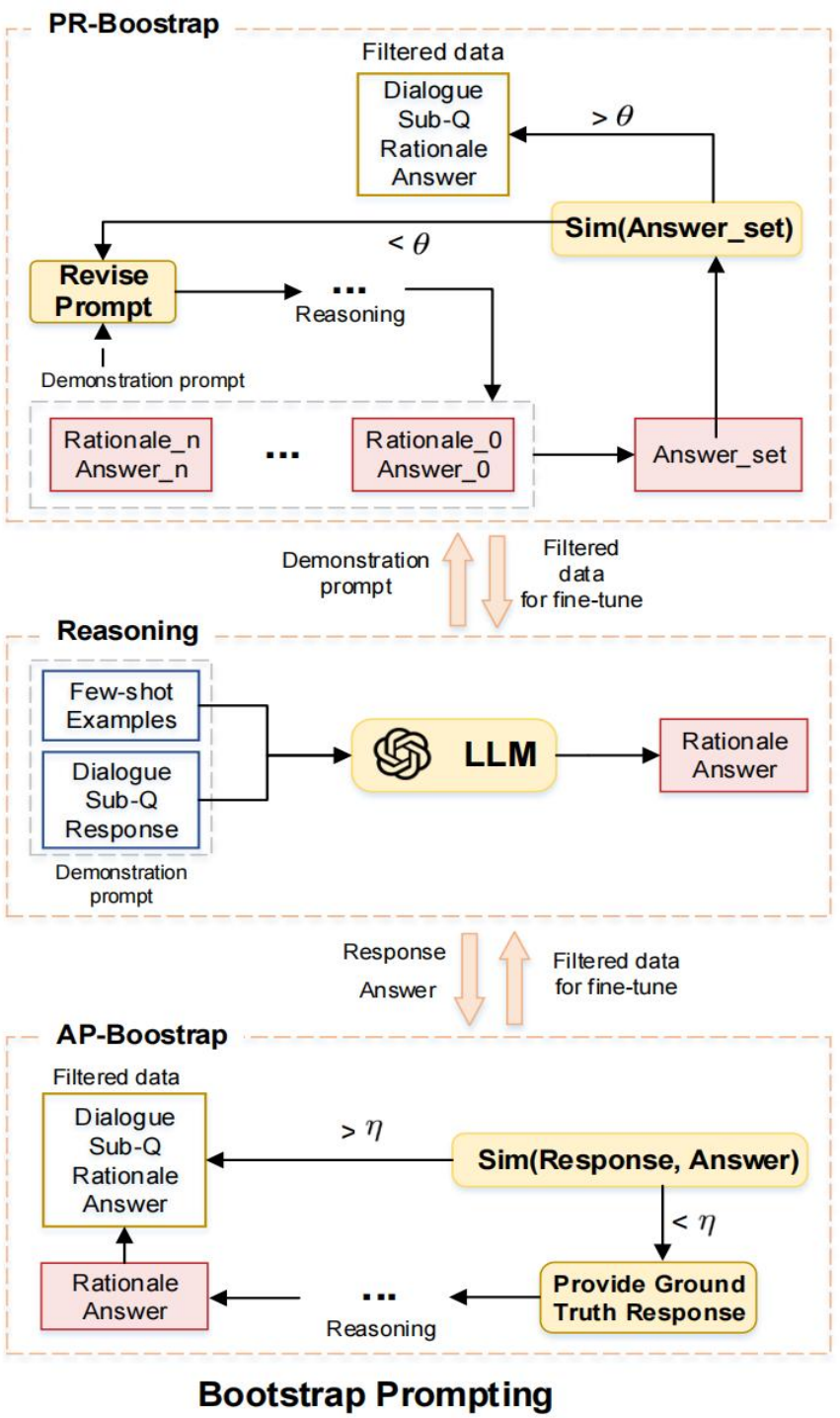
- Answer-Providing Bootstrapping

$$\mathcal{J}(\mathcal{M}, H, R) = \sum_i \mathbb{E}_{\hat{Q}_i, A_i \sim p_M(\cdot|H_i)} f_I(\cdot)$$

$$\nabla \mathcal{J}(\mathcal{M}, H, R) = \sum_i \mathbb{E}_{\hat{Q}_i, A_i \sim p_M(\cdot|H_i)} [f_I(\cdot) \cdot \nabla log p_M = (A_i, \hat{Q}_i | H_i)]$$

- Prompt-Revising Bootstrapping

iterate the prompt revision process to explore diverse reasoning paths and generate alternative answers until reliable answers are obtained for all provided data.



Bootstrap Prompting

Experiments

1. Main Evaluation

(1) Automatic Evaluation

Dataset	Model	B@1	B@2	B@4	R@1	R@2	D@2
MedDG	Seq2Seq (Sutskever et al., 2014)	28.55	22.85	15.45	25.61	11.24	/
	HRED (Serban et al., 2016)	31.61	25.22	17.05	24.17	9.79	/
	DialoGPT (Zhang et al., 2019)	32.77†	26.93†	17.96†	27.11†	11.34†	79.26†
	GPT-2 (Radford et al., 2019)	35.27	28.19	19.16	28.74	13.61	/
	VRBot (Li et al., 2021)	29.69	23.9	16.34	24.69	11.23	/
	MedPIR (Zhao et al., 2022)	38.72†	27.64†	18.14†	25.72†	10.30†	82.77†
	DFMed (Xu et al., 2023)	42.56	33.34	22.53	29.31	14.21	/
	ChatGLM-6B (Du et al., 2022)	37.96	24.22	15.37	18.05	10.53	89.81
	BP4ER (ours)	**44.78**	**33.80**	**23.76**	**41.47**	**22.47**	**89.93**
	Improvement	+2.22	+0.46	+1.23	+12.16	+8.26	+0.12
KaMed	Seq2Seq (Sutskever et al., 2014)	23.52	18.56	12.13	23.56	8.67	/
	HRED (Serban et al., 2016)	26.75	21.08	16.36	18.71	7.28	/
	DialoGPT (Zhang et al., 2019)	30.17†	25.53†	17.09†	24.30†	9.79†	80.27†
	GPT-2 (Radford et al., 2019)	33.76	26.58	17.82	26.8	10.59	/
	VRBot (Li et al., 2021)	30.04	23.76	16.36	18.71	7.28	/
	MedPIR (Zhao et al., 2022)	29.42†	21.60†	16.47†	20.69†	9.27†	83.75†
	DFMed (Xu et al., 2023)	40.20	30.97	20.76	28.28	11.54	/
	ChatGLM-6B (Du et al., 2022)	38.70	27.19	16.38	33.86	20.21	85.70
	BP4ER (ours)	**41.89**	**31.74**	**20.81**	**35.76**	**21.19**	**86.83**
	Improvement	+1.69	+0.77	+0.05	+1.90	+0.98	+1.07

Figure 5 Automatic Evaluation

Observation:

- BP4ER enhances response quality and ensures semantic consistency
- BP4ER performs better with smaller, focused datasets but faces challenges with larger, diverse datasets.
- Compared to traditional seq2seq-based models, LLM-based models good at capturing richness and diversity in MDG, making them ideal for tasks requiring comprehensive and diverse outputs.

Experiments

1. Main Evaluation

(2) Human Evaluation

Model	Fluency	Cohe.	Correct.
DialoGPT	3.11	2.56	2.89
MedPIR	3.34	3.07	3.23
BP4ER	4.00	3.50	3.52
Gold	4.32	4.17	4.41

Figure 6 Human Evaluation

Observation:

- consistently outperformed other models across all three manual evaluation indicators

Experiments

2. Ablation Study

Fine-Tune	Exp. Rea.	AP-Boots.	PR-Boots.	MedDG B@1	MedDG R@1	MedDG D@1	MedDG D@2	KaMed B@1	KaMed R@1	KaMed D@1	KaMed D@2
✓	✓	✓	✓	44.78	41.47	91.20	89.93	41.89	35.76	89.10	86.83
✓	✓	✓		42.27	37.64	89.76	88.73	40.69	35.01	87.97	85.94
✓	✓			40.75	36.63	90.14	88.90	39.68	34.99	88.47	86.07
✓				39.41	27.38	88.54	89.81	39.13	33.97	87.34	85.83

Figure 7 Human Evaluation

Confirm the effectiveness of each module on model performance. The results indicate that all these modules contribute positively to our approach, underscoring their importance in achieving superior performance in MDG tasks.

Experiments

3. CaseStudy

- Highlights BP4ER's ability to produce responses that are more aligned with the context of the dialogue.
- BP4ER predicts the next diagnostic decision and generates a response to inquire about the patients drug history
- BP4ER still falls short when compared to the gold standard response

P: I have morning and evening stomach pain, which eases after eating, with normal bowel movements and no nausea.
R: How long? Have any prior stomach issues?
P: Three or four days. No stomach issues, just these two days.
Gold Response: Probably peptic ulcer. You can take Daxi and omeprazole .
MedPIR: Stomach indigestion; eat small, frequent meals.
ChatGLM: Gastric ulcer, avoid greasy, spicy foods and caffeine.
BP4ER:
Patient's State: Stomach pain in the morning and evening, eases after eating, normal stools, no nausea, lasts for three or four days, no stomach issues.
Diagnostic Decision: Give diagnostic results and drug advises.
Response: Gastric ulcer, suggest to take anti-inflammatory drugs, such as omeprazole.

P: Is the pain on the outer front of the knee rheumatism?
Gold Response: How long? Does it hurt when pressed down?
MePIR: The knee pain can have many causes, not necessarily related to rheumatism.
BP4ER: How long has the pain been lasting? How does it pain?

P: What medication or injection for brucellosis? I'm a sheep farmer, and I get the disease with no energy and cold.
R: Hi, have you been diagnosed? How long have you been sick?
P: It has been two or three months since the diagnosis.
Gold Response: Rifampicin 600-900mg/day plus tetracycline 200mg/day for 6 weeks. Is there any medication? It needs a sufficient course of treatment.
ChatGLM: Brucellosis is an infectious disease caused by Brucella, that may require antibiotics to treat.
BP4ER: Have any medicine? Antibiotics are usually used for brucellosis, like tetracycline.

Figure 4 Case Study

Conclusion

- BP4ER is the first model to systematically deconstruct MDG into an intermediate reasoning chain, which notably enhances the interpretability of the MDG process.
- Introduces LLM with the LMP strategy to explicit reasoning
- Proposed an iterative approach to bootstrap the prompting process for augmenting the LLM’s reasoning abiolities

Limitations and Future Works

1. Demonstration prompts are designed by human.

→ Guide LLMs to automatically generate several sub-questions for the current dialogue (in progress)

2. Lack of specialized medical knowledge results in ncorrect or nonsensical responses during the process of reasoning.

→ Construct a disease-oriented knowledge graph, transform the knowledge into text to be added to the prompt for incorporation into the LLM (in progress)

THANK YOU