Multi-party Goal Tracking with LLMs: Comparing Pre-training, Fine-tuning, and Prompt Engineering

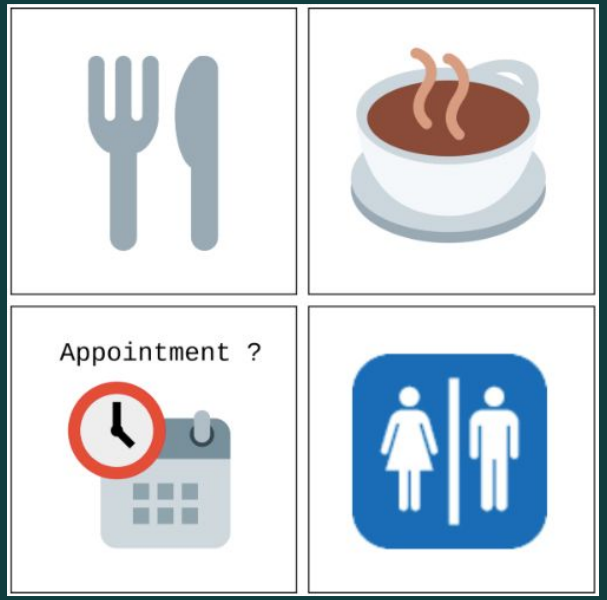
Angus Addlesee, Weronika Sieińska, Nancie Gunson, Daniel Hernández Garcia, Christian Dondrup, and Oliver Lemon

HERIOT WATT UNIVERSITY

The EU's SPRING H2020 Project

We are working on the dialogue system embedded within an ARI robot in a hospital memory clinic.

Data Collection



(Novikova et al., 2016)

- A patient and their companion are given their goals via pictograms (to encourage lexical diversity).
- Goals are arranged to elicit various scenarios, e.g. agreement, disagreement, nervousness, etc... (Addlesee et al., 2023).

Multi-Party Challenges

Who is speaking?

1: I think it is London... 2: Yeah, London

Multi-Party Challenges

Who is speaking? 1: I think it is London... 2: Yeah, London

Addressee? Are they addressing you, their friend, or the group?

Multi-Party Challenges

Who is speaking? 1: I think it is London… 2: Yeah, London

Addressee? Are they addressing you, their friend, or the group?

Response? Depends on above, not just dialogue content

Multi-Party Challenges

Who is speaking? 1: I think it is London… 2: Yeah, London

Addressee? Are they addressing you, their friend, or the group?

Response? Depends on above, not just dialogue content

Intent-slots? Agreement, split-utterances, etc…

Multi-Party Challenges

Who is speaking? 1: I think it is London… 2: Yeah, London

Addressee? Are they addressing you, their friend, or the group?

Response? Depends on above, not just dialogue content

Intent-slots? Agreement, split-utterances, etc…

Goal tracking? Share goals, answer each other

Example Dialogue 1 (more in paper)

1	**ARI**: Hello, my name is ARI. How can I help you?	
2	**User 1**: Hello ARI, we have to find a lift to get to the next floor. Where is the lift in this building?	
3	**ARI**: The lifts are at the entrance on the left.	
4	**User 1**: Ok, and we have time to have a coffee before going upstairs, is that right?	
5	**ARI**: Yes, absolutely.	

Example Dialogue 2 (more in paper)

	User 1: What time was our appointment?	1
	User 2: So we have an appointment at 10.30 PM.	2
	User 1: Ok.	3
	User 2: It's 10 PM just now so we have plenty of time.	4
	User 1: We'll walk slowly.	5

29 Dialogues Used Here

Wizard of Oz setup

29 Dialogues Used Here

Wizard of Oz setup

~80 Dialogues Collected

And counting!!
Now with live system

DialogLED by Microsoft (Zhong et al., 2022)

Noise Type	Original Dialogue	Noisy Dialogue
Speaker Mask	Tom: The weather is good today!	[MASK]: The weather is good today!
Turn Splitting	Tom: The weather is good today! Do you have any plans? How about we go to play basketball?	Tom: The weather is good today! [MASK]:Do you have any plans? [MASK]:How about we go to play basketball?
Turn Merging	Tom: The weather is good today! Do you have any plans? Bob: I still have homework to do today. I'm afraid I can't go out to play.	Tom: The weather is good today! Do you have any plans? I still have homework to do today. I'm afraid I can't go out to play.
Text Infilling	Tom: The weather is good today! Do you have any plans? How about we go to play basketball?	Tom: The weather is [MASK] Do you have [MASK] any plans? [MASK] we go to play basketball?
Turn Permutation	Tom: Do you have any plans? Bob: How about we go to play basketball? Sam: I still have homework to do today. I'm afraid I can't go out to play.	Sam: I still have homework to do today. I'm afraid I can't go out to play. Tom: Do you have any plans? Bob: How about we go to play basketball?

DialogLED - Our Modification

This window-based denoising defined our task.

	Original Dialogue	Noisy Dialogue
Goal Masking	Alex: Yes, we need to go to room 17 for treatment. @ G(LC+RP, go-to(room_17))	Alex: Yes, we need to go to room 17 for treatment. **[MASK]**
Intent-slot Masking	Morgan: Um, hello. Could you help me? # greet(); request(help)	Morgan: Um, hello. Could you help me? **[MASK]**

T5

Included BERT-style de-masking in training, very key for our task!

GPT 3.5-turbo

Prompt ‘styles’

Basic

Simply tells the model what we want as output

Prompt ‘styles’

Basic

Simply tells the model what we want as output

Specific

Added lots of detail to the basic prompt

Prompt ‘styles’

Basic

Simply tells the model what we want as output

Specific

Added lots of detail to the basic prompt

Annotation

We provide annotation instructions

Prompt ‘styles’

Basic

Simply tells the model what we want as output

Specific

Added lots of detail to the basic prompt

Annotation

We provide annotation instructions

Story

Present the task as a fictional tale

Prompt ‘styles’

Basic

Simply tells the model what we want as output

Specific

Added lots of detail to the basic prompt

Annotation

We provide annotation instructions

Role-play

Tell the model to be a character doing a task

Story

Present the task as a fictional tale

Prompt ‘styles’

Basic

Simply tells the model what we want as output

Specific

Added lots of detail to the basic prompt

Annotation

We provide annotation instructions

Reasoning

Explain why we annotate particular things with reasons

Role-play

Tell the model to be a character doing a task

Story

Present the task as a fictional tale

Big Tables in Paper

Model	train/test %	Prompt Style	Exact %	Correct %	Partial %
T5	0/100	-	0	0	0
T5	20/80	-	0 ± 0	0 ± 0	0 ± 0
T5	50/50	-	0 ± 0	0 ± 0	0 ± 0
T5	80/20	-	0 ± 0	0 ± 0	0 ± 0
DialogLED	0/100	-	0	0	0
DialogLED	20/80	-	0 ± 0	0 ± 0	5.80 ± 1.45
DialogLED	50/50	-	0 ± 0	2.38 ± 2.38	1.19 ± 0.63
DialogLED	80/20	-	0 ± 0	0 ± 0	20 ± 11.55
GPT 3.5-turbo	0/100	Basic	0	3.45	31.03
GPT 3.5-turbo	0/100	Specific	0	3.45	24.14
GPT 3.5-turbo	0/100	Annotation	0	6.90	44.83
GPT 3.5-turbo	0/100	Story	0	0	0
GPT 3.5-turbo	0/100	Role-play	0	0	6.90
GPT 3.5-turbo	0/100	Reasoning	3.45	34.48	79.31
GPT 3.5-turbo	7/80*	Basic	11.59 ± 3.83	30.43 ± 10.94	86.96 ± 6.64
GPT 3.5-turbo	7/80*	Specific	20.29 ± 3.83	43.48 ± 9.05	92.75 ± 2.90
GPT 3.5-turbo	7/80*	Annotation	14.49 ± 5.80	28.99 ± 3.83	82.61 ± 4.35
GPT 3.5-turbo	7/80*	Story	17.39 ± 6.64	36.23 ± 13.83	86.96 ± 4.35
GPT 3.5-turbo	7/80*	Role-play	18.84 ± 7.25	46.38 ± 12.38	92.75 ± 5.22
GPT 3.5-turbo	7/80*	Reasoning	**27.54 ± 1.45**	**62.32 ± 9.50**	**94.20 ± 5.80**

Goal Tracking: GPT 3.5-turbo, few-shot

Prompt	Correct
Basic	30.43
Specific	43.48
Annotation	28.99
Story	36.23
Role-play	46.38
Reasoning	**62.32**

Intent-slot Recognition: GPT, few-shot

Prompt	Correct
Basic	36.23
Specific	60.87
Annotation	40.58
Story	47.83
Role-play	49.27
Reasoning	**69.57**

Future Work

Experiment with open LLM's

We cannot use GPT-4 with real patients for obvious data security reasons.

Keep collecting data

We are continuing this collection, and I have another focused on people with dementia

Knowledge Grounding

If a social robot was really deployed in a hospital, it could not ever lie to patients!

Thanks!

Do you have any questions?

a.addlesee@hw.ac.uk | addlesee.co.uk

Angus Addlesee (@addlesee_ai) on: