Metamo: Empowering Large Language Models with Psychological Distortion Detection for Cognition-aware Coaching

Hajime Hotta[1], **Huu-Loi Le**[2], **Manh-Cuong Phan**[3], **Minh-Tien Nguyen**[3]*
[1] Hajime Institute, Kuala Lumpur, Malaysia.
hotta@hajime.institute
[2] AI Academy Vietnam, Vietnam.
loilh@aiacademy.edu.vn
[3] Hung Yen University of Technology and Education, Hung Yen, Vietnam.
cuongpm@spkt.edu.vn; tiennm@utehy.edu.vn

Abstract

We demonstrate Metamo, a browser-based dialogue system that transforms an off-the-shelf large language model into an empathetic coach for everyday workplace concerns. Metamo introduces a light, single-pass wrapper that first identifies the cognitive distortion behind an emotion, then recognizes the user's emotion, and finally produces a question-centered reply that invites reflection, all within one model call. The wrapper keeps the response time below two seconds in the API, yet enriches the feedback with cognitively grounded insight. A front-end web interface renders the detected emotion as an animated avatar and shows distortion badges in real time, whereas a safety layer blocks medical advice and redirects crisis language to human hotlines. Empirical tests on public corpora confirmed that the proposed design improved emotion-recognition quality and response diversity without sacrificing latency. A small user study with company staff reported higher perceived empathy and usability than a latency-matched baseline. Metamo is model-agnostic, illustrating a practical path toward cognition-aware coaching tools.

1 Introduction

Conversational coaching tools have begun to accompany performance reviews, career planning, and daily self-reflection in the workplace (Mitchell et al., 2022; Beinema et al., 2023; Arakawa and Yakura, 2024; Ong et al., 2024). Built on large language models (LLMs), such systems can deliver fluent encouragement; however, they typically operate with a single prompt that folds perception and advice into one turn. It leads to two limitations. First, the generated guidance is often generic, without empathy, because the prompt lacks an explicit representation of *why* the user feels distressed. Second, multi-step prompting or tool calling pushes response latency beyond the five-second window that humans perceive as natural.

Psychological research attributes many negative emotions to recurring thought patterns, known as *cognitive distortions* (Beck, 2020). Recent NLP studies have shown that prompting LLMs to label such distortions can sharpen downstream reasoning (Chen et al., 2023; Wang et al., 2024; Lim et al., 2024), but existing pipelines are usually designed with two separate steps: emotion recognition with shallow textual representation, and response generation. This makes ill-suited for interactive and empathetic coaching (Liong and Patel, 2024).

We argue that conversational coaching systems should be empowered by a deeper understanding of human thinking that fosters empathetic effective communication between humans and AI models. To address this argument, this study presents **Metamo**, a demonstration system that is empowered by psychological distortion detection. The system uses a compact prompt to first yield a ten-way distortion label, then emotion labels, and finally, a dialogue move that combines a Socratic follow-up question with a brief reframing suggestion. The design preserves latency while making cognitive analysis transparent: the web client maps the emotion to an avatar's facial expression and displays the distortion as a color-coded badge.

We evaluated Metamo under zero-shot conditions on EMPATHETICDIALOGUES (Rashkin et al., 2019) and a subset of EDOS (Welivita et al., 2021). The wrapper improves emotion recognition accuracy and response diversity over strong baselines, yet maintains real-time speed on both GPT-3.5 via API and a local Qwen-7B model. Human evaluation with corporate employees further indicated higher System Usability Scale scores and perceived empathy. The system is model-agnostic and requires no fine-tuning, offering a concrete blueprint for deploying cognition-aware coaching without compromising speed, safety, or user experience.

*Corresponding Author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 862–872

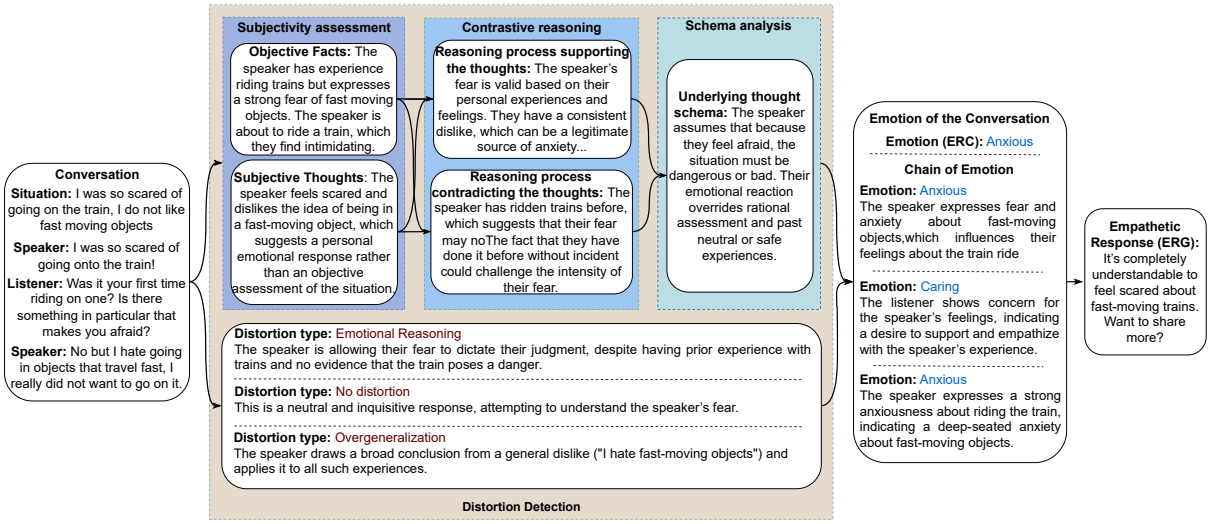

Figure 1: The Metamo model receives a conversation and its situation, and predicts distortions. It next recognizes emotions using cognitive indicators. It finally uses emotions and cognitive indicators for generation.

2 System Design

The system includes three main steps: distortion detection, emotion recognition in conversations, and empathetic response generation.

2.1 Distortion Detection

Cognitive distortion detection serves as a foundational procedure in cognitive behavioral therapy (CBT) (Beck, 2020) and has recently been investigated to enhance the reasoning capabilities of LLMs (Chen et al., 2023; Wang et al., 2024; Lim et al., 2024). According to CBT, mental disorders are closely associated to maladaptive thought patterns that manifest through emotional responses. For instance, individuals experiencing depression or anxiety often show negative thoughts, leading to negative emotions. Inspired by DoT (Dianogsis-of-Thought) (Chen et al., 2023), we introduce distortion detection to detect cognitive distortions of participants in a conversation. The detection includes three steps: subjectivity assessment, contrastive reasoning, and schema analysis. Distortion detection is a core module that establishes a cognitive model for subsequent reasoning processes.

Subjectivity assessment To detect cognitive distortions, the initial step involves evaluating the subjectivity of the speaker (or listener). This is because their utterances typically contain a mixture of objective information (factual content) and subjective interpretations or opinions. The subjectivity assessment is responsible for identifying and differentiating between these objective facts and subjective thoughts given the situation. This crucial step clarifies factual statements from personal opinions, providing clearer evidence for further analysis.

Contrastive reasoning Once the assessment has been established, contrastive reasoning offers two distinct reasoning processes to evaluate subjective thoughts using objective evidence. The first process establishes support for the thoughts, while the second constructs arguments that contradict them. By contrasting these opposing interpretations derived from the same set of facts, the method can more precisely uncover and characterize the speaker's underlying thought schemas.

Schema analysis This step elucidates the reasons behind the speaker's formation of specific reasoning processes. Drawing on psychological principles, we refer a "schema" as a cognitive structure that integrates an individual's knowledge, beliefs, and expectations to construct a coherent cognitive model. By identifying and analyzing these schemas, the method can effectively build a comprehensive cognitive model of the conversation to support subsequent reasoning tasks.

Given a situation of a conversation and the current utterance, the detection classifies distortions of the utterance into 10 types: personalization, mind reading, overgeneralization, all-or-nothing thinking, emotional reasoning, labeling, magnification, mental filter, should statements, and fortune-telling (Beck, 2020; Shreevastava and Foltz, 2021). These types are injected into the prompt and LLMs are required to predict the distortion which has the highest probability given an input conversation.

2.2 Emotion Recognition

Once the cognitive model has been built, the method recognizes emotions of a conversation. Given the situation and cognitive indicators, the method predicts the emotion of the conversation (or utterances) by prompting LLMs. Different from sentiment analysis that only uses shallow textual features, our method takes into account cognitive modeling in the form of distortions that are in a deeper level than the shallow textual level.

The chain of emotions We observe that emotional states have interconnections with distortions in utterances of a conversation. This is because the chain of emotions illustrates the emotional states engagement of both the speaker and listener in a conversation. Constructing this chain involves two important perspectives. From a psychological perspective, it reflects the transformation from cognitive modeling at a deeper level to emotional expression at a shallow level. By capturing individuals' thoughts throughout the dialogue, the chain of emotions enables the method to perform more nuanced and fine-grained reasoning. Regarding emotion expression, the overall emotion conveyed by a conversation should align with this emotional progression. For instance, in Figure 1, the conversation's emotion can be inferred as "Anxious" based on the chain "Anxious-Caring-Anxious". This chain of emotions is used for response generation.

2.3 Empathetic Response Generation

The final step is to generate empathetic responses using information from the situation, cognitive indicators, and the chain of emotions. In practice, the method uses a prompt that includes three components for emotion recognition and response generation. The empathy aspect is enhanced by using constraints in the prompt that guide LLMs to generate more empathetic responses.

3 Evaluation

The Metamo was applied to two tasks: emotion recognition in conversations (ERC) and empathetic response generation (ERG), which are critical components of an empathetic coaching system.

3.1 Settings

Datasets The evaluation uses two datasets. `EmpatheticDialogues` (Rashkin et al., 2019) is a text-only corpus. Each conversation contains a situation and a conversation created by the interaction between the speaker and listener. `EDOS` is a large-scale dataset for empathetic response generation (Welivita et al., 2021). It includes 1M dialogues from movie subtitles. Due to the limited budget, 500 dialogues from EDOS were randomly sampled. Table 1 shows the statistics of the datasets.

Table 1: Statistics of ERC and ERG datasets.

Dataset	train	dev	test	# labels
EmpatheticDialogues	19,533	2,770	2,547	18, 32
EDOS (500)	—	—	500	32

Baselines Baselines were selected with two categories: LLMs and task-specific models. For LLMs, the proposed method was tested with **GPT-3.5** (gpt-3.5-turbo-0125), **GPT-o3-mini**, **GPT-4o-mini** because of promising results in the zero-shot setting (Nori et al., 2023; Peng et al., 2023; Mozikov et al., 2024; Katz et al., 2024). We did not use GPT-4 due to the high cost. The method was further confirmed with **Qwen-7B** (Yang et al., 2024) and **Mistral-7B** (Jiang et al., 2023), small open models showing competitive results in various tasks. We also reimplemented **InsideOut** (Mozikov et al., 2024).

For task-specific models, **Transformer** uses the encoder-decoder architecture (Waswani et al., 2017) for ERG. **MoLE** (Lin et al., 2019) handles both ERC and ERG tasks by first detecting user emotions and then generating empathetic responses. **EmpDG** (Li et al., 2020) utilizes coarse-grained dialogue and token-level emotions. **CEM** (Sabour et al., 2022) takes into account the understanding of users' emotions and cognition for ERC and ERG. **KEMP** (Li et al., 2022) combines ConceptNet and emotional lexicons to enhance the representation of implicit emotions. **text-CNN**, **bcLSTM**, and **DiaRNN** uses the text modality for training ERC models (Poria et al., 2019). **MM-DFN** is a multimodal model, but we report the results using the text modality (Hu et al., 2022). **CFEG** (Chen et al., 2024) is a cause-aware ERG approach that uses Chain-of-Though prompts for fine-tuning LLMs.

Evaluation metrics Following previous studies (Cai et al., 2023; Mozikov et al., 2024), the evaluation uses accuracy for ERC; and ROUGE (Rouge, 2004), BLEU (Papineni et al., 2002), Distinct-1 (Li et al., 2016), and human evaluation for ERG.

3.2 Results and Discussion

3.2.1 Performance Comparison

Emotion recognition in conversations Table 2 summarizes the accuracy of LLM-based methods

for the ERC task. As observed, the proposed multi-agent framework achieves the best results in most of all cases. More importantly, the proposed method is significantly better than the baselines that directly use LLMs. A possible reason is that cognitive modeling in the form of distortion detection helps the method to simulate the thinking process and then to understand deeply the reasoning mechanism of humans. Interestingly, the gap between GPT-o3-mini and the proposed method is small, in which GPT-o3-mini outputs better performance on EmpatheticDialogues. A possible reason is that GPT-o3-mini is good at reasoning,[1] so adding more complex prompts seems to be unnecessary. The InsideOut method achieves low performance, as our argument in Section 1, in which InsideOut uses a shallow emotion analysis while the proposed model uses a deeper level of cognitive modeling.

Table 2: The performance of ERC. **Bold** text is the best. 18 classes were shared from Mozikov et al. (2024).

# Classes	LLMs	Baseline	InsideOut	Metamo
EmpatheticDialog				
32	GPT-3.5	35.89	36.59	**45.66**
	GPT-o3-mini	**48.49**	40.11	48.29
	GPT-4o-mini	42.87	40.28	**48.48**
	Qwen	39.07	30.43	**42.51**
	Mistral	37.38	34.35	**42.30**
18	GPT-3.5	55.71	39.97	**56.61**
	GPT-o3-mini	**62.35**	41.46	61.80
	GPT-4o-mini	56.69	45.70	**62.36**
	Qwen	45.94	40.16	**56.73**
	Mistral	46.09	43.35	**56.17**
EDOS				
32	GPT-3.5	24.60	27.40	**28.60**
	GPT-o3-mini	27.60	25.80	**28.40**
	GPT-4o-mini	25.80	25.05	**28.80**
	Qwen	25.40	22.00	**28.00**
	Mistral	21.40	21.00	**25.65**

Table 3: Comparison with strong task-specific methods. The results are derived from Mozikov et al. (2024).

EmpatheticDialogues (32 classes)						
LLMs	**Metamo**	**MoEL**	**MINE**	**EmpDG**	**CEM**	**KEMP**
GPT-3.5	**45.66**					
GPT-o3-mini	**48.29**					
GPT-4o-mini	**48.48**	31.74	30.96	31.65	36.84	36.57
Qwen	**42.51**					
Mistral	**42.30**					

Table 3 shows the comparison of Metamo with task-specific models of ERC. The proposed method outperforms task-specific models. There are two possible reasons. First, the proposed framework uses LLMs that have shown promising scores for ERC (Mozikov et al., 2024). Second, LLMs are empowered by cognitive modeling by using multiple agents. It allows the framework to better understand the thinking process of humans, leading to the improvements of emotion prediction. We could not compare the method with strong models on EDOS due to different settings.

Empathetic response generation Table 4 reports the comparison between the proposed framework and strong methods of ERG.

Table 4: The performance of ERG.

LLMs	Method	B-1	B-2	R-1	R-2	Dist-1
EmpatheticDialogues						
GPT-3.5	Baseline	11.49	2.05	12.78	**1.89**	0.989
	InsideOut	11.11	1.56	11.43	1.39	0.983
	Metamo	**13.56**	**4.04**	**13.30**	1.77	**0.991**
GPT-o3-mini	Baseline	9.26	0.95	**9.91**	**0.85**	**0.976**
	InsideOut	6.98	0.27	5.98	0.23	0.974
	Metamo	**9.29**	**1.47**	9.18	0.56	0.961
GPT-4o-mini	Baseline	11.52	2.13	13.17	**1.87**	**0.983**
	InsideOut	8.97	1.23	11.09	1.33	0.982
	Metamo	**13.06**	**3.97**	**13.74**	1.78	0.982
Qwen	Baseline	10.47	1.68	**13.34**	**1.56**	0.985
	InsideOut	9.47	1.29	11.62	1.35	0.982
	Metamo	**12.54**	**3.12**	12.32	1.36	**0.988**
Mistral	Baseline	8.74	2.06	**14.73**	**1.93**	0.924
	InsideOut	8.47	1.83	13.56	1.77	**0.989**
	Metamo	**11.64**	**3.23**	14.19	1.65	0.944
EDOS						
GPT-3.5	Baseline	13.71	2.56	9.14	0.75	**0.981**
	InsideOut	11.42	2.44	8.06	0.73	0.976
	Metamo	**13.80**	**4.92**	**10.64**	**2.11**	0.959
GPT-o3-mini	Baseline	9.14	1.21	7.00	0.37	0.968
	InsideOut	8.74	1.16	5.26	0.23	0.982
	Metamo	**9.74**	**1.68**	**7.05**	**0.66**	**0.984**
GPT-4o-mini	Baseline	11.29	2.13	8.52	0.62	**0.975**
	InsideOut	10.06	1.93	7.90	0.60	0.970
	Metamo	**12.46**	**2.45**	**9.14**	**0.65**	0.965
Qwen	Baseline	11.48	2.72	7.67	0.75	0.990
	InsideOut	10.22	1.97	6.73	0.70	0.991
	Metamo	12.25	1.95	**8.38**	**0.63**	**0.994**
Mistral	Baseline	**9.93**	**2.11**	**9.32**	0.75	0.934
	InsideOut	9.36	1.65	9.22	**0.96**	**0.977**
	Metamo	9.32	1.91	6.03	0.47	0.963

The scores share the trend with the ERC task in Table 2, in which Metamo achieves better scores than the methods that only use LLMs in almost all cases. The improvements come from cognitive modeling and multiple agents for response generation. The performance of the proposed framework is better than that of InsideOut and baseline. It validates our assumption that modeling cognition benefits the reasoning process.

We also compared the method with strong ERG models. Table 5 shows that there are large gaps between LLM-based methods and task-specific fine-tuned models. This is because LLM-based methods

[1] https://openai.com/index/openai-o3-mini/

work in the zero-shot setting, while task-specific models are fine-tuned with training data of the downstream task. Also note that the performance of the proposed method can still be improved by using few-shot learning.

Table 5: Comparison with task-specific methods on EmpatheticDialogues. The results of task-specific methods are from Mozikov et al. (2024) and Zhang et al. (2024).

Method	B-1	B-2	R-1	R-2	Dist-1
Transformer	18.07	8.34	17.22	4.21	0.36
MoEL	18.07	8.30	18.24	**4.81**	0.59
MINE	18.60	8.39	17.08	4.05	0.47
EmpDG	**19.96**	9.11	**18.02**	4.43	0.46
CEM	16.12	7.29	15.77	4.50	0.62
KEMP	16.72	7.17	16.11	3.31	0.66
CFEG	—	**10.54**	—	—	**2.96**
GPT-3.5	**13.56**	**4.04**	**13.30**	1.77	**0.991**
GPT-o3-mini	9.29	1.47	9.18	0.56	0.961
GPT-4o-mini	**13.06**	**3.97**	**13.74**	1.78	0.982
Qwen	12.54	3.12	12.32	1.36	**0.988**
Mistral	**11.64**	**3.23**	14.19	1.65	0.944

We conducted human evaluation because generating tokens with empathy challenges automatic evaluation for both meaning and linguistics. We randomly selected 100 samples for five annotators with four metrics: Fluency (**F**), Identification (**I**), Empathy (**E**), Suggestion (**S**), and Overall Effectiveness (**OE**) (Mozikov et al., 2024). Fluency measures the coherency and fluency of the generated response. Identification assesses the accurate identification of the problem that the speaker needs support from the listener. Empathy estimates the understanding empathy of the response of the listener given the emotion and feeling of the speaker. Suggestion measures the appropriateness and usefulness of advice or suggestions from the listener. Overall Effectiveness estimates effectiveness in providing information of the listener that supports the speaker's emotions. The annotators were trained with a guideline in three steps: (i) reading the situation of the conversation, (ii) understanding the conversation and the gold reference (the final utterance of the listener), and (iii) giving the score of the generated response based on steps (i) and (ii). After training, each annotator gave a score, ranging from 1 to 10, for each metric.[2]

The results in Table 6 confirm that the proposed framework obtains the best scores in almost all cases, followed by the baseline. This is because Metamo incorporates cognitive modeling in the form of multiple agents. Among the metrics, the fluency scores are the highest, followed by the empathy and identification scores. It is understandable that LLMs can generate fluency responses. The gap score for empathy between the proposed framework and the baselines is quite large, which confirms the contribution of cognitive modeling to enhance empathy in LLMs.

[2]The evaluation link: https://tinyurl.com/yupkp395

Table 6: Human evaluation of ERG.

LLMs	Method	F	I	E	S	OF	Average
EmpatheticDialogues							
GPT-3.5	Baseline	8.21	7.04	7.53	6.37	6.89	7.20
	InsideOut	8.13	6.88	7.10	6.76	6.30	7.03
	Metamo	**9.30**	**8.27**	**9.15**	**7.14**	**8.47**	**8.46**
GPT-o3- mini	Baseline	8.44	7.76	7.40	5.14	7.00	7.14
	InsideOut	7.99	7.46	7.12	5.46	6.53	6.77
	Metamo	**9.03**	**7.92**	**8.09**	**7.11**	**8.10**	**8.08**
GPT-4o-mini	Baseline	9.09	7.08	8.04	5.38	7.51	7.42
	InsideOut	8.09	7.87	7.15	6.64	8.06	7.56
	Metamo	**9.19**	**8.21**	**9.12**	**7.23**	**8.51**	**8.45**
Qwen	Baseline	7.78	7.11	7.03	5.42	6.60	6.70
	InsideOut	7.49	6.90	6.57	5.82	6.05	6.56
	Metamo	**9.11**	**8.15**	**8.08**	**7.23**	**8.00**	**8.11**
Mistral	Baseline	7.08	6.86	6.04	6.50	6.63	6.56
	InsideOut	7.63	6.94	6.56	5.95	6.45	6.70
	Metamo	**9.00**	**8.09**	**8.04**	**7.10**	**7.99**	**8.04**

3.2.2 Prompt Sensitivity

To assess the robustness of our framework to prompt design, we conducted a series of sensitivity experiments in which the original prompt was systematically perturbed. We considered three settings: **Ablation**, where a key portion of the instruction was omitted by removing two guiding questions in the diagnosis-of-thought step; **Terminology**, where domain-specific terminology was replaced with semantically similar but less technical expressions (Table 12); and **Rephrased**, where ChatGPT was instructed to rewrite the original prompt in its own words. The observation was done with 100 samples on EmpatheticDialogues.

Table 7: Accuracy of different prompt settings on the ERC (18 classes) task in EmpatheticDialogues.

Model	Metamo			
	Original	Ablation	Terminology	Rephrased
GPT-3.5	59.00	60.00	**61.00**	54.00
GPT-4o-mini	67.00	65.00	**68.00**	57.00
GPT-o3-mini	**62.00**	61.00	59.00	**62.00**

Table 7 reports ERC performance across models. The results indicate that all three settings yield comparable accuracy to the original prompt, with deviations of only one to three percentage points. Notably, Terminology did not substantially degrade accuracy, suggesting that the models can tolerate terminological shifts as long as semantic intent is preserved. In contrast, Rephrased showed a clearer drop in performance for GPT-3.5 and GPT-4o-mini. This suggests that when the model is asked to freely

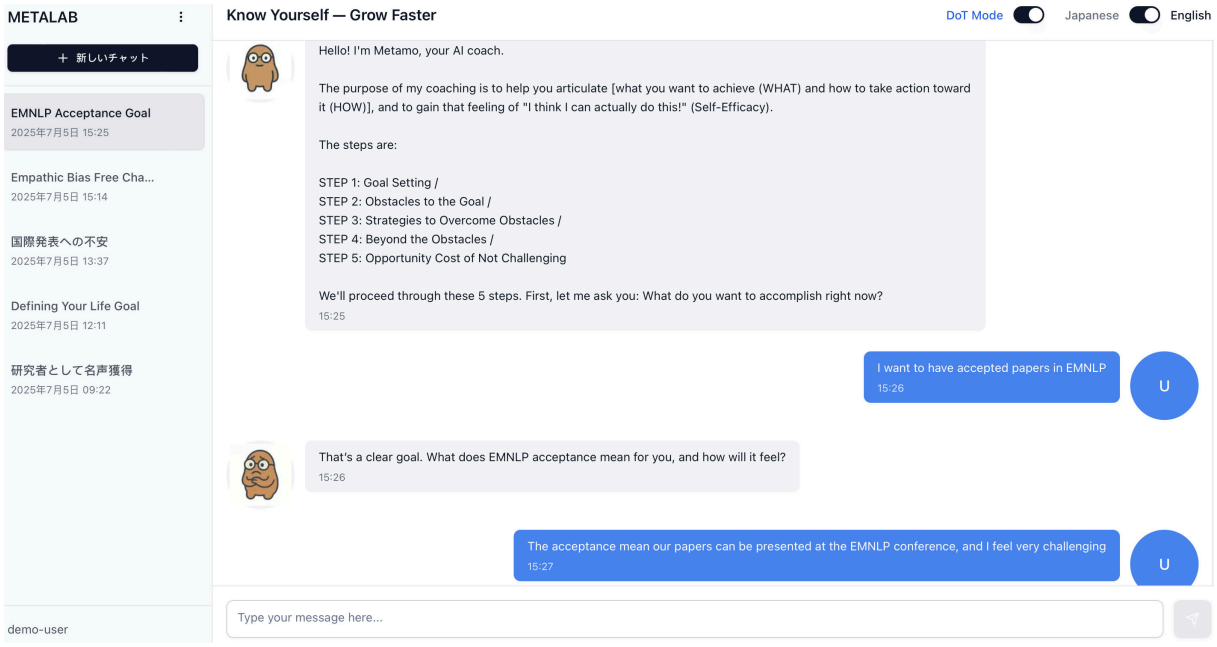

Figure 2: Metamo: the coaching system using Metamo. The system interacts with the user in the case that the user wants to have an accepted paper at EMNLP but he is not so confidence due to a very competitive conference.

rewrite the prompt, it tends to omit or soften certain instructions that are crucial but not obvious on the surface. As a result, the new prompt loses part of the structure that originally guided the reasoning process, leading to weaker results.

Table 8: Performance of different prompt settings on the ERG task in EmpatheticDialogues.

Model	Metamo	B-1	B-2	R-1	R-2	Dist-1
GPT-3.5	Original	13.91	3.35	13.14	**1.46**	**0.991**
	Ablation	14.07	2.87	13.51	0.71	**0.991**
	Terminology	13.63	3.31	13.51	1.12	0.986
	Rephrased	**14.15**	**3.37**	13.13	1.37	0.986
GPT-4o mini	Original	13.27	**4.18**	14.18	**2.04**	0.984
	Ablation	**14.52**	3.98	**14.98**	1.54	0.987
	Terminology	13.20	3.93	14.75	1.83	0.983
	Rephrased	12.17	3.27	12.21	1.54	**0.993**
GPT-o3 mini	Original	**8.44**	**1.99**	9.85	**0.99**	0.947
	Ablation	6.58	0.97	**10.79**	0.31	0.948
	Terminology	8.16	0.91	9.25	0.64	0.956
	Rephrased	7.77	1.51	9.68	0.62	**0.958**

Table 8 reports the performance of different prompt settings on the ERG task. Terminology appears relatively stable, with scores close to the original prompt, confirming that semantic intent rather than precise wording drives performance. Ablation yields mixed outcomes: for some models it slightly improves BLEU-1 or ROUGE-1, but at the cost of lower BLEU-2 and diversity, suggesting that removing diagnostic questions narrows responses to simpler patterns. In contrast, Rephrased consistently degrades results for GPT-4o-mini and GPT-o3-mini, while only marginally affecting GPT-3.5. This indicates that free-form rewriting can strip away the implicit structure of the original prompt, leading to less coherent and less cognitively grounded outputs. Overall, the results highlight that empathetic response generation is more sensitive to prompt design than emotion recognition, and that carefully engineered instructions remain crucial for sustaining quality.

3.2.3 Running Time and Cost

This section provides the investigation of running time and cost which are critical for actual applications. To do that, 100 samples were randomly selected from the EmpatheticDialogues dataset. Table 9 shows the average numbers computed in three runs. The baseline only using GPT-4o-mini is the best in terms of cost. Our method follows the baseline on the cost. For running time, the ratio also shows that our method is faster for ERG but slower for ERC. This is because ERG bases on ERC, in which ERC outputs distortion detection and chain of emotions for generating empathetic responses. In contrast, the baseline uses two different prompts for the two tasks. The InsideOut method takes a lot of time and cost due to the collaboration of multi-agents. The observation shows that Metamo is practical for actual cases.

4 Demonstration Scenario

This section shows the demonstration scenario of Metamo in two parts: system demonstration and human evaluation.

Table 9: The running time and cost of 100 samples. The time ratio was computed over the beseline.

Method	Task	Time (s)	Time ratio (%)	Cost (USD)
Baseline	ERC	0.966	—	0.002
	ERG	2.370	—	0.008
InsideOut	ERC	5.912	617.798	0.056
	ERG	9.094	403.998	0.102
Metamo	ERC	1.011	105,712	0.015
	ERG	1.787	79.3869	0.033

4.1 Metamo

The proposed prompt (Figure 1) was applied to build Metamo, an empathetic conversational coaching system that makes a huge impact (Mitchell et al., 2022; Arakawa and Yakura, 2024). We refer to "know yourself" in terms of understanding distortions and emotions. The system can be easily to be adapted to education or healthcare support.

Figure 2 shows the Metamo[3] system,[4] which offers two languages: Japanese and English. The central panel shows the main screen of conversations between users and the system, and the configuration. When users access the system and create a new conversation, the main screen in the center shows the guideline to setup the coaching in five steps shown in Figure 2. Metamo guides users to follow these steps to make the final advice. In conversations, Metamo detects distortions of users, predicts their emotions, shows appropriate avatars, and generates empathetic responses. It also utilizes a Socratic follow-up question combined with a brief reframing suggestion to interact with users.

4.2 Human Evaluation

Tasks We asked 38 people (English and Japanese, 32-57 years old) to judge the user experience of the system in three tasks: career change, English learning, and team relations. The scenario seed shown to the participants (summarized) is as follows. For career change, the scenario is "You feel anxious about a job transition." For English learning, the scenario is "You aim to raise your TOEIC score in 3 months." For team relation, the scenario is "You face conflict with a team-mate at work."

Settings Participants rate the system in three settings: (1) bare minimal prompt, no coaching steps, no avatar using ChatGPT, (2) the system with 5-step strategy prompt, neutral avatars, and (3) the proposed Metamo system.

[3]System: https://dot.hajime-institute.com/

[4]Video: https://www.youtube.com/watch?v=ui1NuSMFoxw

Procedure Each participant completed nine sections (three conditions: emotion classification, generation quality, and overall satisfaction) with the score of 1-5 (larger is better). The free chat is larger than or equal five round-trip within six minutes. Table 10 shows the instrumentation.

Table 10: The instrumentation of human evaluation.

Construction	Items	Timing
Facial empathy	3	post-session
Working alliance	10	post-session
Session satisfaction	8	post-session

The instrumentation is shown in Appendix A.2.

Human evaluation results Table 11 shows the results of human evaluation of Metamo. It indicates that our method obtains the highest scores of user satisfaction thanks to cognitive distortion detection. It is understandable that Metamo integrates cognitive distortion into the reasoning process, leading to more human-like responses. The STEP method obtains the second-best overall score. The possible reason is the use of neutral avatars compared to the dynamic facial expressions of Metamo. ChatGPT produces the lowest scores because the simple prompt does not use cognition-aware coaching.

Table 11: User experience of the system.

Method	Classification	Generation	Overall
Bare (ChatGPT)	3.16	3.73	3.82
STEP without Bare	3.15	4.24	4.57
Ours	**4.34**	**4.33**	**4.79**

5 Conclusions

This paper introduces Metamo, an empathetic conversational coaching system. Empowered by psychological distortion detection, Metamo can analyze thinking patterns of users in a deeper level, and then predict their emotions for empathetic response generation. The dialog move is also combined with a Socratic follow-up question with a brief reframing suggestion. Metamo is evaluated on two tasks: emotion recognition in conversations and empathetic response generation on two benchmark datasets. Promising experimental results leverage Metamo for actual coaching scenarios.

Future work will apply Metamo to other domains, e.g. education, and utilize multimodal data to improve the model's performance. In addition, human evaluation with larger, more diverse populations and longitudinal deployments will be important to further validate these findings.

Limitations

Although achieving promising results, the proposed method has the following limitations. First, it models cognition using distortion detection. It allows us to deeply understand the thought of the participants in a conversation. As a result, the framework obtains competitive performance when the conversation has distortion indicators. In other cases, understanding emotion states requires other cognitive aspects. Note that in mental health or health care treatment, distortion detection should be confirmed by psychologists. Second, the method is designed to handle text-only data; therefore, multimodal datasets may challenge the method and require other designs. However, the method can be easily extended for working with multimodal data by adding new agents, e.g., visual agents. Finally, Metamo requires more guardrails for actual user deployment. This helps Metamo avoid LLM jailbreaking by using prompt-based attacks.

Human evaluation, while valuable, is based on a relatively small number of samples and participants. The human evaluation of empathetic response generation in Section 3.2.1 only uses 100 samples with five annotators. The participant pool in Section 4.2 (38 users, English and Japanese, aged 32–57) is relatively narrow, which may limit the generalizability of the results to broader populations. In addition, the scenarios (career change, English learning, team conflict) and short interaction time (six minutes) simplify real-world conditions, and thus may not fully reflect long-term or diverse use cases. Furthermore, the study relies on self-reported satisfaction and alliance measures, which, while standard in user studies, can be influenced by expectations or novelty. Despite these constraints, the consistent improvements across different tasks and settings suggest that the observed benefits of cognition-aware responses are not accidental but stem from systematic design choices.

Ethics Statement

The authors confirm that this study does not have ethical issues. Thanks to the sharing of prompts from InsiteOut and DoT authors, we can successfully run the two methods. People participated in the human evaluation process with a clear guideline and observation to avoid human bias. We do not use any personal information from annotators for experiments. Code and datasets are collected from public GitHub links from original papers. All experiments do not have specific parameter tuning to maintain a fair comparison among methods.

References

Riku Arakawa and Hiromu Yakura. 2024. Coaching copilot: blended form of an llm-powered chatbot and a human coach to effectively support self-reflection for leadership growth. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, pages 1–14.

Judith S Beck. 2020. *Cognitive behavior therapy: Basics and beyond*. Guilford Publications.

Tessa Beinema, Harm Op den Akker, Hermie J Hermens, and Lex van Velsen. 2023. What to discuss?—a blueprint topic model for health coaching dialogues with conversational agents. *International Journal of Human–Computer Interaction*, 39(1):164–182.

Hua Cai, Xuli Shen, Qing Xu, Weilin Shen, Xiaomei Wang, Weifeng Ge, Xiaoqing Zheng, and Xiangyang Xue. 2023. Improving empathetic dialogue generation by dynamically infusing commonsense knowledge. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7858–7873.

Xinhao Chen, Chong Yang, Man Lan, Li Cai, Yang Chen, Tu Hu, Xinlin Zhuang, and Aimin Zhou. 2024. Cause-aware empathetic response generation via chain-of-thought fine-tuning. *arXiv preprint arXiv:2408.11599*.

Zhiyu Chen, Yujie Lu, and William Wang. 2023. Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4295–4304.

Dou Hu, Xiaolong Hou, Lingwei Wei, Lianxin Jiang, and Yang Mo. 2022. Mm-dfn: Multimodal dynamic fusion network for emotion recognition in conversations. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7037–7041. IEEE.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2024. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020. Empdg: Multi-resolution interactive empathetic dialogue generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4454–4466.

Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge bridging for empathetic dialogue generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 10993–11001.

Sehee Lim, Yejin Kim, Chi-Hyun Choi, Jy-yong Sohn, and Byung-Hoon Kim. 2024. Erd: A framework for improving llm reasoning for cognitive distortion classification. *arXiv preprint arXiv:2403.14255*.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 121–132.

Kyrin Liong and Nadya Shaznay Patel. 2024. Empathetic coaching conversations: Being emotionally present and connected with students. In *Coaching Students in Higher Education*, pages 53–69. Routledge.

Elliot Mitchell, Noemie Elhadad, and Lena Mamykina. 2022. Examining ai methods for micro-coaching dialogs. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–24.

Mikhail Mozikov, Nikita Severin, Maria Glushanina, Mikhail Baklashkin, Andrey Savchenko, and Ilya Makarov. 2024. Insideout: Unifying emotional llms to foster empathy. In *ECAI 2024*, pages 4499–4502. IOS Press.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Qi Chwen Ong, Chin-Siang Ang, Davidson Zun Yin Chee, Ashwini Lawate, Frederick Sundram, Mayank Dalakoti, Leonardo Pasalic, Daniel To, Tatiana Erlikh Fox, Iva Bojic, et al. 2024. Advancing health coaching: A comparative study of large language model and health coaches. *Artificial Intelligence in Medicine*, 157:103004.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381. Association for Computational Linguistics.

Lin CY Rouge. 2004. A package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization of ACL, Spain*, volume 5.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. Cem: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11229–11237.

Sagarika Shreevastava and Peter Foltz. 2021. Detecting cognitive distortions from patient-therapist interactions. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 151–158.

Ruiyi Wang, Stephanie Milani, Jamie Chiu, Jiayin Zhi, Shaun Eack, Travis Labrum, Samuel Murphy, Nev Jones, Kate Hardy, Hong Shen, et al. 2024. Patient-: Using large language models to simulate patients for training mental health professionals. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12772–12797.

A Waswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, A Gomez, L Kaiser, and I Polosukhin. 2017. Attention is all you need. In *NIPS*.

Anuradha Welivita, Yubo Xie, and Pearl Pu. 2021. A large-scale dataset for empathetic response generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1264.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Yiqun Zhang, Fanheng Kong, Peidong Wang, Shuang Sun, SWangLing SWangLing, Shi Feng, Daling Wang, Yifei Zhang, and Kaisong Song. 2024. Stickerconv: Generating multimodal empathetic responses from scratch. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7707–7733.

A Appendix

A.1 Prompt Sensitivity

Table 12 shows the Terminology setting used for the observation of prompt sensitivity.

Table 12: Replacement of technical terms in the Terminology setting.

Original term	Replacement term
Cognitive distortion	Thinking error
Diagnosis of thought	Thought analysis
Reasoning process	Logic pathway
Underlying cognition mode	Core belief
Objective facts	Verifiable facts
Subjective opinions	Personal beliefs

A.2 Instrumentation

This section shows the instrumentation in Tables 13, 14, and 15 for human evaluation in Section 4.

A.3 Prompt Details

Table 16 provides detailed information of the prompts used in the baseline (LLMs) and proposed method.

Table 13: Metamo facial-empathy short scale.

No.	Instrumentation
1	The AI coach's character expression accurately reflected how I felt.
2	The AI coach's expression made me feel understood and at ease.
3	The expression shown by the AI coach was appropriate for the situation.

Table 14: Working alliance inventory.

No.	Instrumentation
1	I believe the AI coach likes me.
2	The AI coach and I respect each other.
3	I feel confident working with the AI coach.
4	The AI coach and I trust each other.
5	The AI coach and I agree on the goals for this session.
6	The AI coach and I are working toward mutually agreed-upon goals.
7	The goals we set are important to me.
8	The AI coach and I agree on what tasks will help me reach my goals.
9	I understand what the AI coach wants me to do during the session.
10	I believe the tasks we worked on will help me.

Table 15: Session satisfaction questionnaire.

No.	Instrumentation
1	How would you rate the overall quality of this AI-coaching session?
2	Did you receive the kind of support you wanted from the AI coach?
3	To what extent did the session meet your needs?
4	If a friend had a similar problem, would you recommend this AI coach?
5	How satisfied are you with the amount of help you received?
6	Has the session helped you deal more effectively with your problem?
7	Overall, how satisfied are you with the AI coach's support?
8	If you needed help again, would you return to this AI coach?

Table 16: The prompts using in the proposed method.

Method	Prompt
Baseline	You will be provided with a conversation and the following situation with the emotion of the speaker and listener. Your task is to create a response to the conversation naturally, as if you were part of the conversation. Your response should reflect understanding, care, and emotional connection.
Metamo	Given a conversation, your task is to: 1) Finish a few diagnose of thought questions to analyze the thought patterns of the participants. Then based on the diagnose of thought analysis, 2) identify if there is cognitive distortion in the conversation. 3) Recognize the specific types of the cognitive distortion. These are the following diagnosis of thought questions: a) What is the situation? Find out the facts that are objective; what are the participants thinking or imagining? Find out the thoughts or opinions that are subjective. b) What makes the participants think the thought is true or is not true? Find out the reasoning processes that support and do not support these thoughts. c) Why do the participants come up with such reasoning processes supporting the thought? What's the underlying cognition mode of it? Based on the diagnosis of thought, for each speech made by the speaker and listener, identify if there is cognitive distortion, specify the type of distortion and provide an explanation. Here we consider the following common distortions: (followed by the descriptions and examples of all ten prompts included in the dataset metadata) <utterance: speaker or listener>: < speech >[<cognitive distortion type >or No distortion] <explanation > Based on the diagnosis of the speaker's and listener's thoughts and the situation of the conversation, your task is to determine the emotions of each utterance in the conversation. <utterance: speaker or listener>: <speech>[emotion] <explanation> The situation of the conversation is following: {situation} The conversation is as follows: {conversation} Based on the above analysis, generate an empathetic response.