Just ask? Exploring Knowledge Capabilities of Large Language Models in Mental Health

Wenwen Li

School of Management

Fudan University

Joint work with Haowen Shen (Fudan University)

Introduction

Depression

- A leading contributor to the global burden of disease (Whiteford et al., 2013)
- 5% of people worldwide have suffered from depression (World Health Organization, 2012)
- Poor help-seeking behavior among mentally ill individuals (Law et al., 2010)

Proactive identification and prevention

Introduction

Leveraging online postings for depression and emotional distress detection

- Beck's cognitive theory of depression (Beck, 1979)
- Cognitive biases and negative thoughts can be identified in one's writing (Pennebaker and King, 1999; Rude et al., 2004).

Machine learning approaches (Chau et al., 2020; Skaik and Inpen, 2020)

- Poor performance
- Poor generalizability

- Limited training data
- Context-specific training tasks

Introduction

- Anlyzing online postings for depression detection

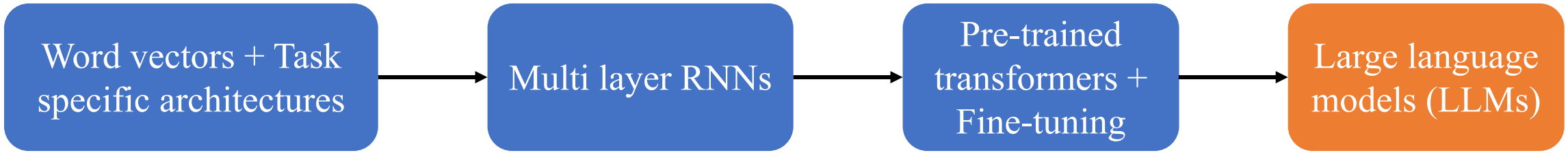
- Natural language processing (NLP) task

Shifting Paradigms in NLP

Introduction

- Limitations of pre-training → Fine-tuning
  - Needing large task-specific datasets for fine-tuning
  - Overfitting, poor generalizability

Shifting Paradigms in NLP



Introduction

Language models:

$$P_\theta(w_t = \text{“}Information\text{”} | w_{t-1}, w_{t-2}, \dots)$$

Large Language Models

- huge training sets
- θ is a large neural network with many *parameters*

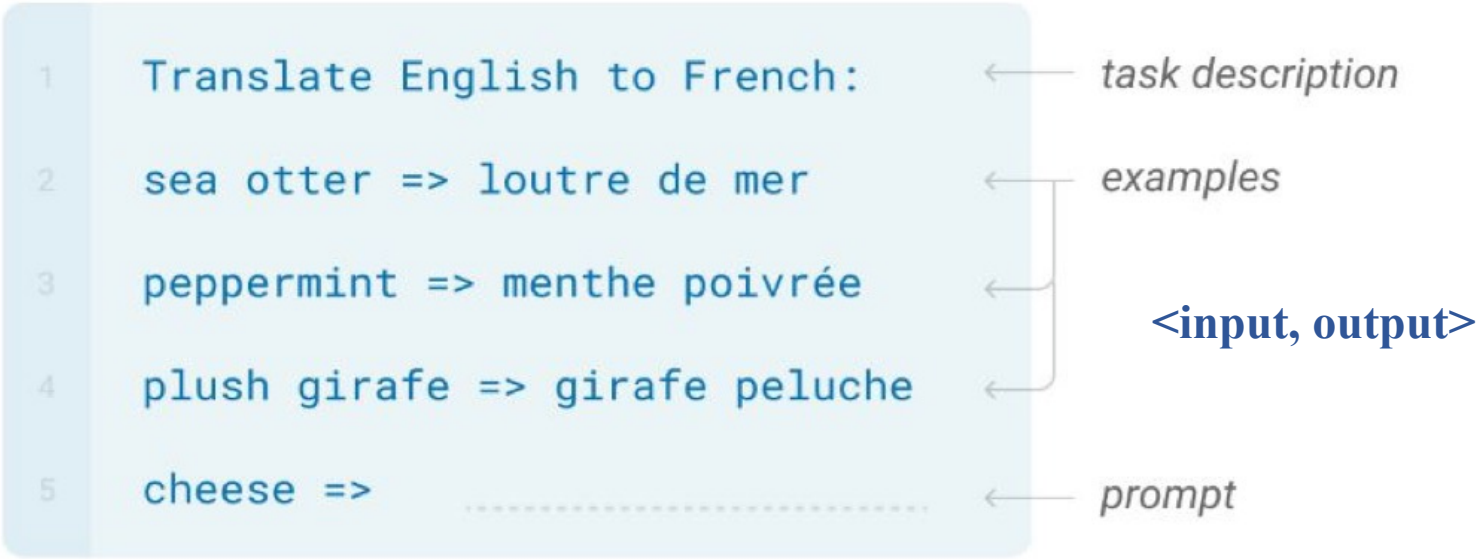
Examples:

- *GPT-3 & GPT-4 (from OpenAI)*
- *LLAMA (from Meta)*
- *BERT (from Google)*

Introduction

Why can LLMs perform better?

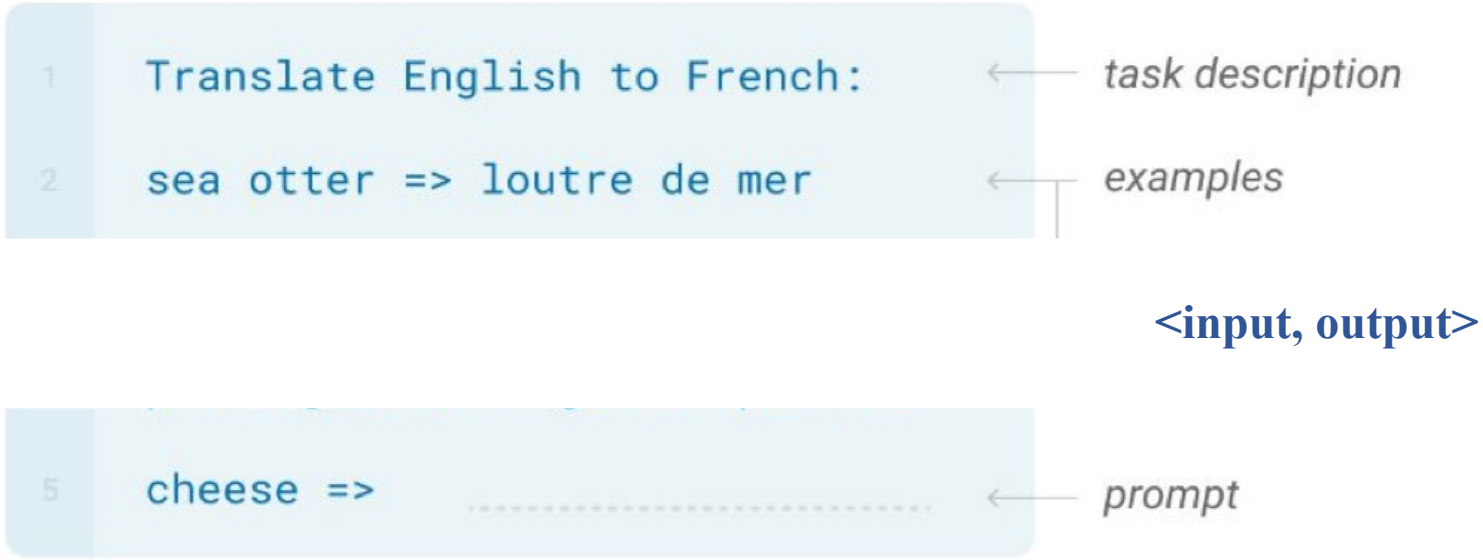
- Scaling up (Kaplan et al. 2020)
- In-context learning (Brown et al. 2020)
  - Learning how to learn
  - The model is only trained once. Then, weights are frozen.
  - Prompting the model at inference time using natural language instructions (prompts)
    - **Few-shot**



Introduction

Why can LLMs perform better?

- Scaling up (Kaplan et al. 2020)
- In-context learning (Brown et al. 2020)
  - Learning how to learn
  - The model is only trained once. Then, weights are frozen.
  - Prompting the model at inference time using natural language instructions
    - Few-shot
    - **One-shot**



Introduction

Why can LLMs perform better?

- Scaling up (Kaplan et al. 2020)
- In-context learning (Brown et al. 2020)
  - Learning how to learn
  - The model is only trained once. Then, weights are frozen.
  - Prompting the model at inference time using natural language instructions
    - Few-shot
    - One-shot
    - **Zero-shot**

```
Translate English to French:        ← task description

cheese =>                           ← prompt
```

Introduction

- E.g., GPTs
  - “Creating a tailored version of ChatGPT to be more helpful at **specific tasks**”
  - “Creating one is as easy as starting a conversation, **giving it instructions** and **extra knowledge**, and picking what it can do, like searching the web, making images or analyzing data”

Instructions/prompts?
Knowledge?

Introduction

Two open challenges for LLMs should be considered:

- **A significant drop in performance whenever the task is highly specialized and requires a large amount of expertise** (Zhu et al. 2023; Wu et al. 2023)
  - A **mismatch** between the domain-specific knowledge required for the task and the knowledge that LLMs use for problem-solving (Zhu et al. 2023; Talmor et al. 2020)
  - The **knowledge capabilities** of LLMs in downstream tasks are ambiguous.

- **Lack of interpretability may hinder preventative precautions against depression** (Pan et al. 2021).
  - Outputs or decisions generated by LLMs are not explainable to humans (Danilevsky et al. 2020).
  - LLMs represent knowledge implicitly in their parameters, which makes it difficult to validate and extract knowledge obtained by LLMs.

Introduction

Human-AI collaboration

Our observations:

- Depression detection currently relies heavily on experts' **tacit knowledge and experience** in practice due to the intricate nature of mental health.
  - Depression often manifests in diverse and subjective ways with various symptoms.
  - A nuanced understanding that may not be easily codified into explicit rules is required to interpret these symptoms.
  - Knowledge graphs/knowledge bases are not suitable.

- Incorporating **domain-specific knowledge** into LLMs through implicitly or explicitly means boosts LLMs' performance on downstream tasks (Pan et al. 2023; Talmor et al. 2020).

- LLMs store a huge amount of knowledge (Petroni et al. 2019).
  - The corpora used to pretrain LLMs are huge aggregations of information and data from the internet.

- The **reasoning capacity** of LLMs holds the potential to enhance both the interpretability of their outputs and their ability to match tasks with appropriate knowledge (Wei et al. 2022).
  - The ability to trace LLMs' inferences back to the input data
  - Reasoning capacity allows LLMs to engage in task-specific reasoning, enabling them to deduce relevant knowledge for a given task.

Introduction

Motivated by these observations, we adopt the design science paradigm (Gregor and Hevner, 2013) and propose **KePrompt** (Knowledge-enhanced Prompting) framework.

- Exploring the knowledge capabilities of LLMs in mental health based on the Known-Unknown quadrant

- Enhancing LLMs' performance on depression detection with a human-AI collaboration design
  - Incorporating domain-specific knowledge explicitly and implicitly with expert evaluation
  - Optimizing prompts with an expert-in-the-loop approach

Literature Review

Depression Detection

- Deep learning methods (Nadeem et al. 2022; Hasan and Ghane, 2022)
- Ignoring the form of depression diagnosis in the real world, which is of little help to professional doctors for instant diagnosis

Health IT in IS

- Online health communities (Agarwal et al. 2010)
- Electronic health records (EHR) and video-sharing sites (Lin et al. 2017; Liu et al. 2020)
- AI assistive systems (Zhu et al. 2021; Yu et al. 2022)

Literature Review

Table 1. The Classification of Prompting Methods

Prompt method	Type	Definition
Hard prompt	Task instruction prompting (Gu et al. 2023)	Providing prompts with instructions for a task to guide the LLM's behavior $x_{input} = f_{prompt}(x, I)$ Here, f_{prompt} denotes the prompting fuction that integrates a given input x with an instruction I.
	In-context learning (Brown et al. 2020)	Providing prompts with a few expamples in the context to make LLMs learn from analogy $x_{input} = f_{prompt}(x, s, I)$ Here, f_{prompt} denotes the prompting fuction that integrates a given input x with demonstration examples s and an optional instruction I.
	Chain-of-thought prompting (Wei et al. 2022)	Providing a series of intermediate reasoning steps At each step t, $x_{input} = f_{prompt}(x, C)$. C denotes the chain of thought.
Soft prompt	Prompt tuning (Lester et al. 2021)	Creating continuous vector representations as input hints
	Prefix tuning (Li and Liang, 2021)	Adding a sequence of continuous task-specific vectors to the input, while keeping the LLM parameters frozen

Literature Review

Prompt Optimization with Human Feedback

- **Soft prompts:** by adding trainable continuous prefixes to NLP tasks and optimizing prompts through methods such as backpropagation (Li and Liang, 2021; Lester et al. 2021; Liu et al. 2021, 2022b; Goswami et al. 2023).
  - Become inapplicable when there is only API access to the LLM

- **Hard prompts:** using human-understandable natural language as prompt templates and optimizing through methods such as automatically generating instructions or guiding searches (Yang et al. 2023; Zhou et al. 2022b; Shin et al. 2020).

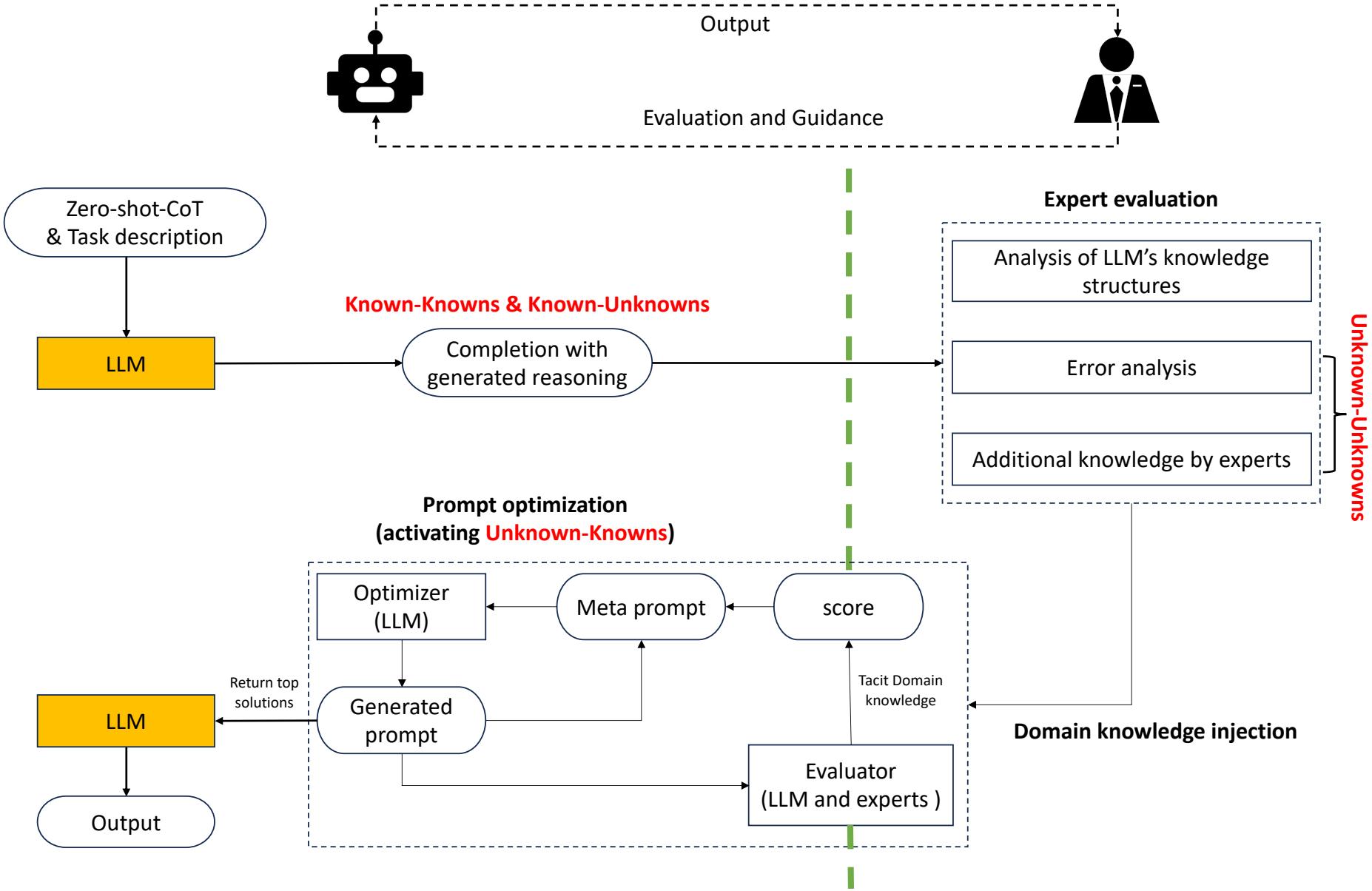
KePrompt Framework

復旦大學管理学院
SCHOOL OF MANAGEMENT
FUDAN UNIVERSITY

Addressing the following three questions:

- For downstream tasks, what kind of domain knowledge is missing from LLMs?
- How to incorporate domain knowledge into LLMs?
- How to improve LLMs' ability to match knowledge to tasks?

KePrompt Framework



Known-Unknown Quadrant

What do LLMs know?

Known-Unknowns • Knowledge unavaliable • Can identify knowledge-task pairs	**Unknown-Unknowns** • Knowledge unavaliable • Cannot identify knowledge-task pairs
Known-Knowns • Knowledge avaliable • Knowledge-task pairs avaliable	**Unknown-Knowns** • Knowledge available • Knowledge-task pairs unavaliable

Knowledge

Awareness

Known-Unknown Quadrant

Two perspectives:

- **Knowledge: the availability of knowledge**
  - measuring particular knowledge accomplishment of LLMs on target tasks
  - At the core of LLMs' functionality is their access to a vast repository of pre-existing human knowledge.
- **Awareness: the ability to match appropriate knowledge to the task**
  - knowledge-task pair matching

The "**Known**": domain knowledge and the ability of knowledge-task pair matching

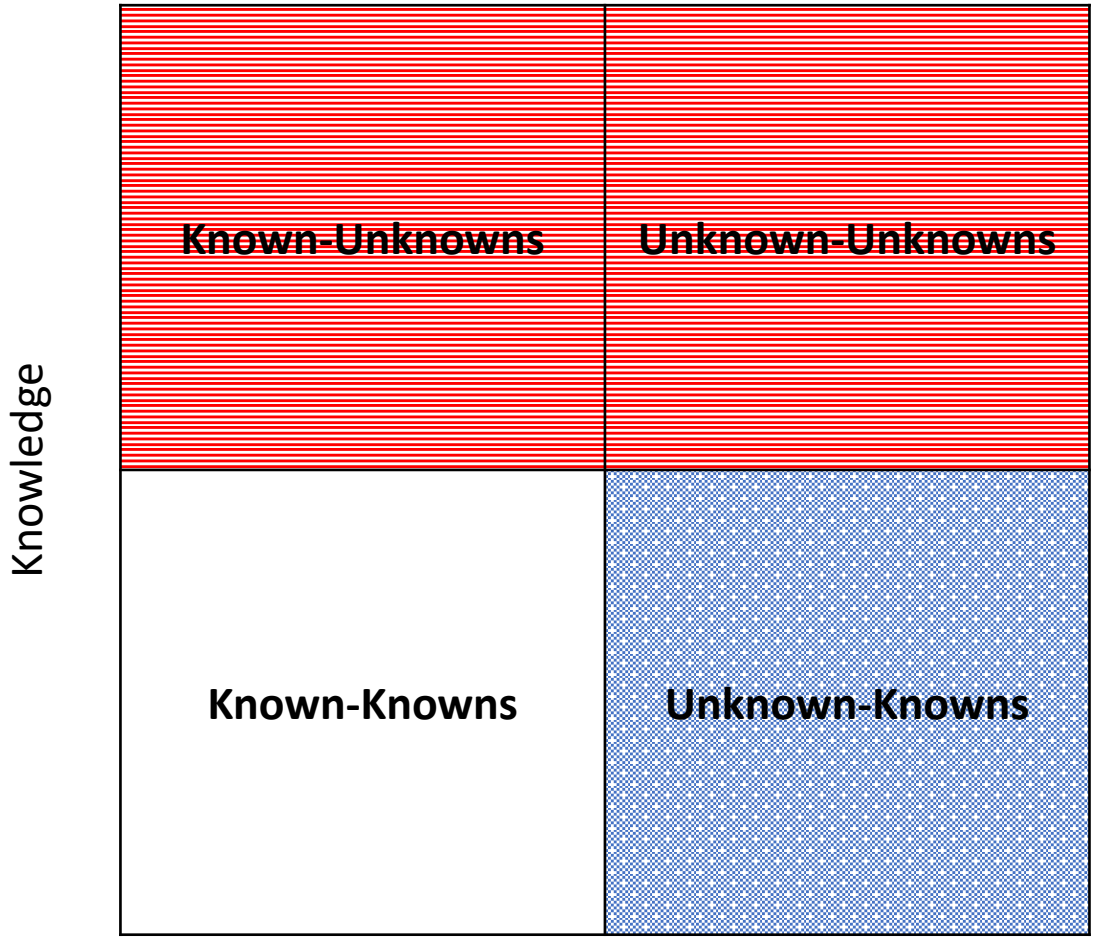
The "**Unknown**": either the absence of required knowledge pertaining directly to downstream tasks, or to the inadequate capacity to match knowledge to the task

Known-Unknown Quadrant

- **"Known-Knowns"** indicates knowledge and information that LLMs are aware of and can utilize in tasks.
- **"Known-Unknowns"** denotes the boundaries of LLMs' existing knowledge.
  - LLMs are aware that there are gaps in their knowledge.
- **"Unknown-Knowns"** represents that LLMs have untapped knowledge that is not properly applied to solve domain tasks.
- **"Unknown-Unknowns"** indicates knowledge that LLMs do not even know they should be considering, which remains beyond the purview of LLMs' training data.

- The aforementioned two challenges for LLMs, i.e., significant drops in model performance on domain tasks and lack of interpretability, also stem from "Known-Unknowns", "Unknown-Knowns" and "Unknown-Unknowns".

KePrompt Framework

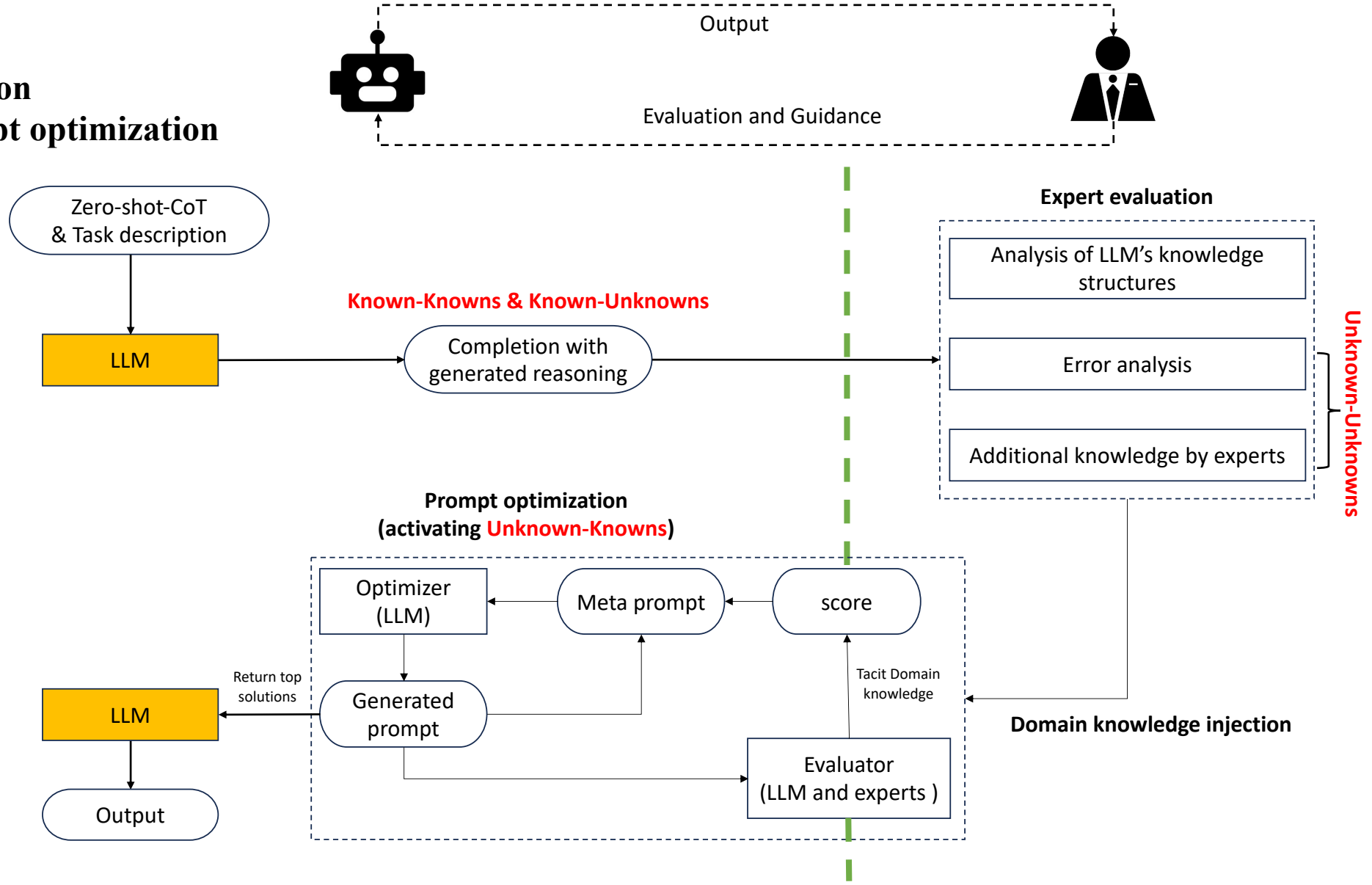


Expert knowledge injection

Prompt optimization

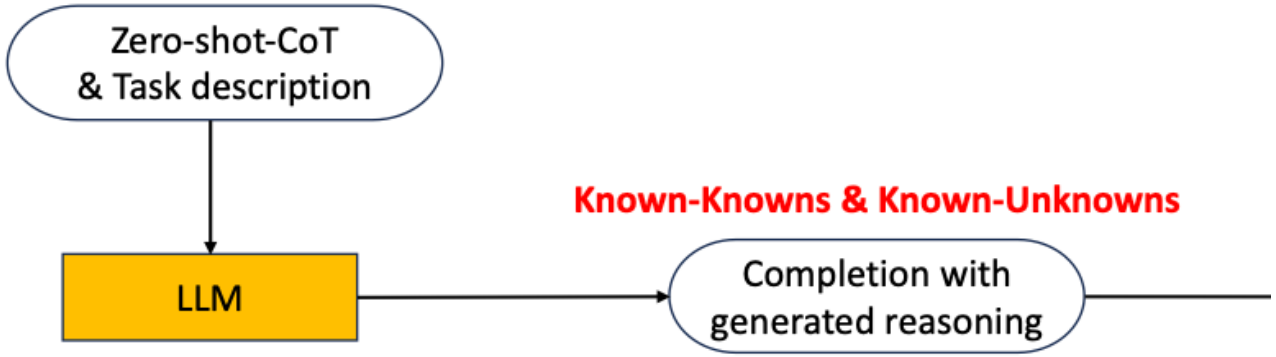
KePrompt Framework

Knowledge elicitation
Expert knowledge injection
Expert-in-the-loop prompt optimization



KePrompt Framework

- **Knowledge elicitation**



KePrompt Framework

復旦大學管理学院
SCHOOL OF MANAGEMENT
FUDAN UNIVERSITY

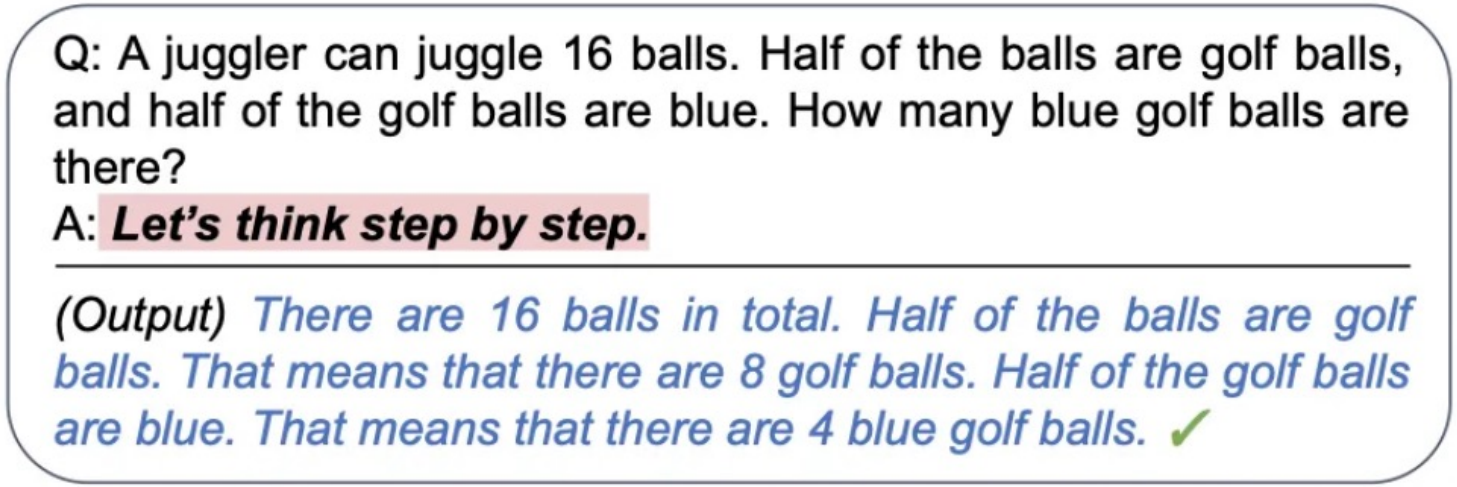
Chain of thought (CoT)

- **A series of intermediate natural language reasoning steps**
- Few-shot CoT (Wei et al. 2022)
- Zero-shot CoT (Kojima et al. 2022)
  - “LLMs are zero-shot reasoners”

Benefits:

- Decomposition -> easier intermediate problems
- **Interpretable**
- Leveraging prompting of LLMs

KePrompt Framework



Zero-shot-CoT prompting (Kojima et al. 2022)

KePrompt Framework

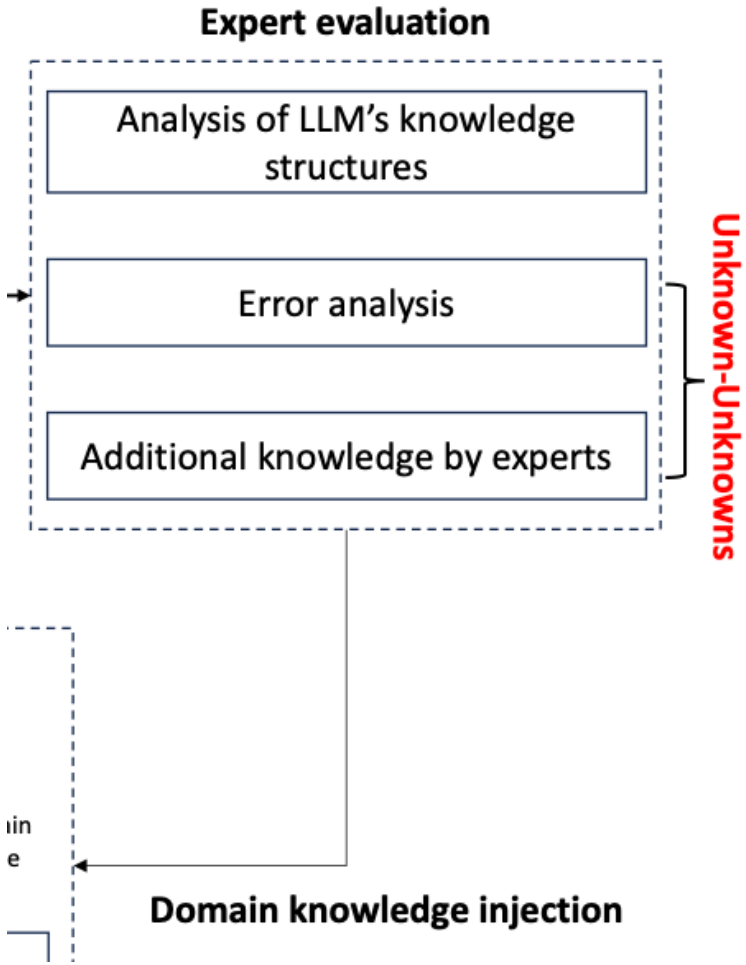
復旦大學管理学院
SCHOOL OF MANAGEMENT
FUDAN UNIVERSITY

- How about using zero-shot CoT for depression detection?

Input
{Context} *You are an expert in identifying depression. Next you need to help me diagnose depression based on user posts.* {CoT} *Let's think step by step.* {Instruction} *Fill in the <ans> part use 1 or 0, where 1 represents for depression and 0 represents for control. Just don't give me the output of None.* {Postings} …
Output by GPT-3.5
The user's posts reveal a mix of emotional resonance, thematic elements, and linguistic usage that fluctuates between lighthearted, casual interactions and moments of distress, including references to sadness, fear, and existential pondering. The posts also hint at past experiences of emotional turmoil and struggles, as well as seeking solace in hobbies and interactions. There are indications of seeking connection and support from others. 1

KePrompt Framework

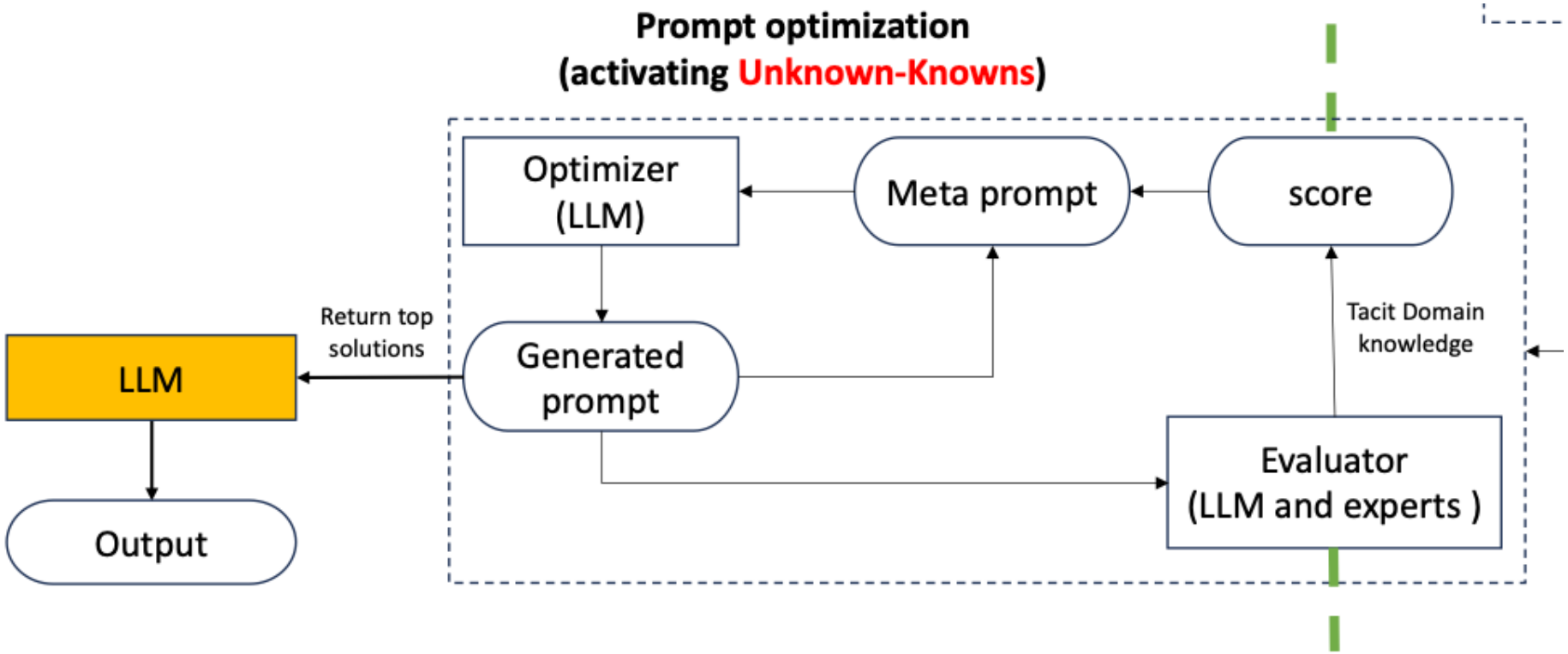
- Expert knowledge injection



KePrompt Framework

復旦大學管理学院
SCHOOL OF MANAGEMENT
FUDAN UNIVERSITY

- Expert-in-the-loop prompt optimization



KePrompt Framework

復旦大學管理学院
SCHOOL OF MANAGEMENT
FUDAN UNIVERSITY

Expert-in-the-loop prompt optimization

- **Initialization:** initial prompt templates along with their associated evaluation metrics (accuracy, precision, recall, and F1 score). This set of templates and their scores serve as our baseline {prompts_and_scores}.
- **Creation of Meta Prompt:** Using the task description, the current set of {prompts_and_scores}, and a few input-output examples, we construct a meta prompt.
- **Optimization Phase:** the meta prompt is sent to an optimizer, which then generates 3-5 prompt templates.
- **Evaluation:** The newly generated template is forwarded to expert and LLM evaluators to be scored based on its performance on the training set.
- **Rank and Update:** The new prompt templates along with their evaluation scores are merged with our existing {prompts_and_scores}, to update the {prompts_and_scores}. We then rank the prompt templates along with their scores. Only the top ten templates (based on their scores) are selected.
- **Iterative Refinement:** Steps 2-4 are repeated iteratively, refining the templates at each iteration.
- **Final Evaluation:** Once we've undergone sufficient optimization rounds, the top-performing templates from the training set are then evaluated on a separate test set to ascertain their real-world effectiveness.

KePrompt Framework

Meta prompt	Evaluator prompt
I have some prompt templates along with their corresponding evaluation index. The prompt templates are arranged in ascending order based on their F1 score. I need you to look at recall, precision, accuracy, and F1 score in an integrated way, here higher scores indicate better quality. prompt templates and their scores: {prompts_and_scores} The following exemplars show how to apply your prompt template: [you replace <prompt> with your prompt template, then read the question and give an output. We say your output is wrong if your output is different from the given output, and we say your output is correct if they are the same.] {exemplars} [Write your new prompt template that is different from the old ones and has a better performance in diagnosing depression which you think will get a f1_score as high as possible on the subsequent train set questions.] Because your answer will be directly used by me as a new template for the subsequent train set questions, remember that you only need to generate a new prompt template, no other analysis is required. Write the prompt template below in square brackets.	Q: {prompt} {question} Now, I need you to analyze based on the above user tweets and finally give me your predicted label of the user status. Your analysis should be specific and concise, including why you predict this user to be depressed or control. I want you to give me your answer in the following form: ["" ### <Analysis>: "Write your analysis of the user here", ### <Label>: "Write your predicted label of the user here" ""] I want you to give me the predicted label in the form of an integer between 1 and 0(1 for "depression" and 0 for "control"). So in your answer, you should only give me an integer between 1 and 0 after the <Label> tag. Even if you think it can't be classified, give directly the output integer between 1 and 0 which you think is closer.

Experiments

- Dataset

Table 2. Summary of the Reddit Data Set

Stage		No. of Users		Number of posts by each user		
		Diagnosed users	Control users	Average	Minimum	Maximum
Expert-in-the-loop Prompt optimization	Optimization	100	100	932.02	265	4839
	Evaluation	250	250	976.56	453	5323
Depression detection		1500	1500	946.06	230	10640

Experiments

Table 3. Examples of prompt content

Task description (TD)
{Context} *You are an expert in identifying depression. Next you need to help me diagnose depression based on user tweets.* {Instruction} *Fill in the <ans> part use 1 or 0, where 1 represents for depression and 0 represents for control. Just don't give me the output of None.* *Even if you think it can't be classified, give directly the output integer between 1 or 0 which you think is closer:*
Zero-shot CoT (ZS-COT)
Let's think step by step.
Type I domain knowledge (DM1)
Here are some key words of depression that may help you identify the user's mental state: *'depressed, sad, unhappy, miserable, sorrowful, dejected, downcast, downhearted, despondent ...etc'*
Type II domain knowledge (DM2)
***Additional Knowledge by Experts**:* *1. Depression is a unipolar affective disorder, typically without significant emotional fluctuations. Therefore, emotional conflicts such as happiness combined with anxiety are not suitable as criteria for judgment.* *2. A pervasive depressed mood and narrowed cognition are more reliable indicators for detection. Depression can lead to a constricted consciousness, such as excessive self-criticism, negating hope (not believing that the future will get better), and an excessive focus on negative information."* *3. ...*

Experiments

Table 4. Experiment Results

Prompt Optimization		Prompt Content	Depression Detection	Precision	Recall	F1 score
Evaluator	**Optimizer**					
Baseline						
-	-	TD	ChatGLM2	0.505	0.658	0.570
-	-	TD+COT	ChatGLM2	0.497	0.759	0.601
-	-		GPT-3.5	0.598	0.595	0.597
-	-	TD+COT+DM1	ChatGLM2	0.499	0.896	0.641
-	-		GPT-3.5	**0.865**	0.536	0.662
ChatGLM2	GPT-3.5	TD+COT	GPT-3.5	0.580	0.790	0.670
Ours						
ChatGLM2	GPT-3.5	TD+COT+DM2 (Knowledge Structure)	ChatGLM2	0.569	0.782	0.659
		TD+COT+DM2 (Error Analysis)		0.580	0.741	0.651
		TD+COT+DM2 (Additional Knowledge)		0.567	0.721	0.635
		TD+COT + DM1		0.639	0.738	0.685
		TD+COT+DM2 (3 Types)		0.636	**0.835**	**0.722**

Experiments

GPT-3.5 with the initial prompt (TD + zero-shot CoT)
The user's posts reveal a mix of emotional resonance, thematic elements, and linguistic usage that fluctuates between lighthearted, casual interactions and moments of distress, including references to sadness, fear, and existential pondering. The posts also hint at past experiences of emotional turmoil and struggles, as well as seeking solace in hobbies and interactions. There are indications of seeking connection and support from others.
GPT-3.5 with optimized prompt
The user's tweets display a range of emotions and experiences, including positive interactions, intellectual discussions, references to enjoying drinks, and discussions about gaming. There are also mentions of feeling upset, frustration, and sadness due to interpersonal conflicts and feelings of intellectual isolation. Additionally, there are references to mental health struggles and seeking help and support. There are indicators of self-awareness, a desire for connection, and a struggle with negative emotions.

More comprehensive

Provide reasons

Conclusions

Our study may have several contributions to healthcare management and large language models.

- Proposing a novel framework that not only leverages but also enhances the knowledge capabilities of LLMs for downstream tasks
- Introducing an expert-in-the-loop prompt optimization strategy to improve the model performance
- Our study sheds light on the promising intersection of LLMs and healthcare management.

Thanks!