TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214
DOI: 10.26599/TST.2025.9010116

Prompt-MAML: Model-Agnostic Meta-in-Context Learning for Major Depressive Disorder Classification

Zita Lifelo, Jianguo Ding, Zongjie Wang*, Feifei Shi, Huansheng Ning, and Sahraoui Dhelim

Abstract: The classification of major depressive disorders (MDDs) is a challenging task in clinical practice, especially in low-resource scenarios where generalization is essential for effective adaptation. Recent progress in meta-training large language models (LLMs) via in-context learning (ICL) offers promise for robust adaptation to unseen tasks without parameter updates. However, existing methods rely on multitask fine-tuning and do not fully exploit the optimization advantages of model-agnostic meta learning (MAML) techniques, limiting their generalization. This study proposes prompt-MAML, a novel method for meta-training LLMs that enhances multimodal ICL for classifying MDD tasks. The method integrates audio-textual features through a transformer-based cross-modal alignment module and incorporates bi-level optimization to learn generalizable model parameters that adapt well to unseen tasks. Extensive experiments demonstrate that prompt-MAML outperforms strong baseline models by an average improvement in macro-F1 of +4% on seen domains, +3% on unseen domains, and +3% in few-shot settings, demonstrating robustness and effectiveness in data-scarce and cross-domain clinical scenarios. Additionally, exploration depth is shown to play a key role in task performance, and further analysis of task complexity, modality, and optimiser configurations highlights critical design considerations for meta-training LLMs.

Key words: in-context learning; large language model; major depressive disorder detection; model-agnostic meta learning; multimodality

1 Introduction

Deep representation learning has transformed machine learning by enabling models to extract unique features from data[1, 2]. This paradigm shift has greatly improved the performance of the model in various tasks, including natural language processing (NLP) and automatic speech recognition (ASR). However, the success of these models frequently requires significant amounts of data to address a specific task, rendering them impractical in real-world scenarios where data availability is limited, costly, or sensitive. This constraint becomes particularly pronounced in domains like nueral machine translation[3] and medical diagnostics[4], such as the classification of major depressive disorders (MDDs), where patient privacy, demographic diversity, and data availability are limited. Most existing approaches rely on supervised learning, which is adapted to specific tasks, or unsupervised learning, which yields general representations that may not transfer well to new tasks[1]. Moreover, training models from scratch for

• Zita Lifelo, Zongjie Wang, Feifei Shi, and Huansheng Ning are with School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China. E-mail: b20200691@xs.ustb.edu.cn; wangzj@ustb.edu.cn; shifeifei@ustb.edu.cn; ninghuansheng@ustb.edu.cn.
• Jianguo Ding is with Department of Computer Science, Blekinge Institute of Technology, Karlskrona 37179, Sweden. E-mail: jianguo.ding@bth.se.
• Sahraoui Dhelim is with School of Computing, Dublin City University, Dublin, D09 V209, Ireland. E-mail: sahraoui.dhelim@dcu.ie.
* To whom correspondence should be addressed.
Manuscript received: 2025-05-12; accepted: 2025-07-04

each new task is not only computationally inefficient but also infeasible under low-resource (LR) or rapidly evolving clinical conditions. As a result, robust generalization and adaptation for MDD classification tasks, under low-resource and across linguistic or demographic shifts remains an open challenge.

Recent advancements in NLP, particularly through large language models (LLMs), have shown significant in-context learning (ICL) capabilities, enabling models to make predictions based on a limited number of illustrative examples presented during inference[5, 6]. ICL removes the necessity for extensive fine-tuning typically associated with deep learning models, facilitating swift generalization to new tasks with reduced computational demands. While pretrained LLMs demonstrate notable ICL abilities, research indicates that meta-training during pretraining can substantially improve this capability[7–9]. These methods provide models with diverse prompt exemplars during training, thereby enhancing the model's performance on new tasks. After the meta-training phase, models undergo evaluation on completely unseen tasks to assess their generalization abilities. While these methods enhance adaptability, they primarily depend on continuous multitask fine-tuning, which fails to fully utilize the optimization benefits of bi-level meta learning frameworks. Furthermore, although there have been promising advancements in multimodal ICL across various tasks[6], the integration of multimodal signals, including audio and text, during meta-training is still largely unexamined. This is especially relevant for clinical applications such as MDD, where multimodality is essential for precise classification.

To overcome these limitations, we turn to model-agnostic meta learning (MAML)[10], a classical meta learning framework designed to train models to quickly adapt to new tasks with minimal data. Unlike standard transfer learning or multitask learning, MAML employs a bi-level optimization scheme: In the inner loop, the model rapidly adapts its parameters to diverse tasks via task-specific gradient updates, while in the outer loop, it aggregates these task-specific adaptations through second-order gradient computations[1, 11]. This setup enables deeper exploration of the parameter space, resulting in more transferable and robust model representations. MAML is particularly well-suited for dynamic, low-resource scenarios, where task diversity and distributional shifts present substantial generalization challenges.

To this end, we introduce prompt-MAML, a novel adaptation of MAML for meta-training LLMs, aimed at enhancing multimodal in-context learning for MDD classification tasks. Our approach embeds a bi-level optimization strategy within an LLM meta-training framework, effectively integrating text and audio data. Audio inputs are encoded via a pretrained speech encoder and projected into the LLM's embedding space using a transformer-based cross-modal alignment module[12]. This modality-aware integration during meta-training improves the model's adaptability to diverse and previously unseen MDD classification tasks. Our key contributions are summarized as follows:

(1) We propose prompt-MAML, a novel meta learning framework that incorporates bi-level optimization into meta-training LLMs for improved multimodal in-context learning performance across MDD classification tasks. Unlike other approaches, our method enables effective parameter initialization and faster adaptation, making it more robust in low-resource and cross-domain scenarios.

(2) For effective multimodal in-context learning, we propose a cross-modal alignment approach using Q-Former, effectively integrating audio features into the LLM embedding space to enrich the contextual understanding necessary for improved MDD classification performance, where both modalities are informative.

(3) We evaluate our method against existing meta-trained ICL approaches under both high-data and limited-data settings, with experiments that systematically examine the impact of task diversity, optimiser choice, and modality on model generalization. Prompt-MAML consistently achieves superior performance across a range of MDD task settings, notably demonstrating strong few-shot adaptation to unseen tasks, which underscores its robustness and effectiveness in data-scarce clinical scenarios.

The remainder of the paper is structured as follows. Section 2 reviews related works, critically analyzing existing approaches and identifying research gaps. Section 3 outlines our proposed method. Section 4 discusses the experimental setup. Datasets, evaluation metrics, and results are discussed in Section 5. In Section 6, we conclude with key insights, limitations, and potential directions for future research.

2 Related Work

2.1 In-context learning

In-context learning, as introduced by Brown et al.[13], allows language models to perform new tasks by utilizing a prompt that contains several task-specific examples, all without necessitating parameter updates. This method has demonstrated efficiency and reduced resource requirements compared to conventional fine-tuning, exhibiting robust performance across various tasks, including complex mathematical problems and reasoning[14, 15]. The behaviour of LLMs in ICL settings has been examined through both theoretical and empirical approaches, revealing that prompt design, exemplar selection, and model scale substantially affect performance[6]. ICL encounters challenges, notably performance degradation when tasks diverge significantly from the pretraining objective or when utilising smaller models, with worst-case performance that can be highly variable[8]. The recent success of ICL in NLP has led to investigations in additional modalities, such as visual[16, 17]. These efforts indicate that findings in textual ICL do not consistently transfer across modalities, highlighting the necessity for domain-specific strategies[6]. The reasons behind the proficiency of LLMs in ICL continue to be an active area of research, garnering interest.

2.2 Meta learning

Meta learning aims to train models that can quickly adapt to new tasks with minimal data[1], offering a promising direction for improving generalization in low-resource settings. Model-agnostic meta learning[10] is a widely adopted meta learning framework that enables models to quickly adapt to new tasks using bi-level optimization: an inner loop that adapts task-specific parameters and an outer loop that updates the model's initialization using second-order gradients. This approach has been shown to improve few-shot learning and task generalization across domains and task shifts[18], including language modeling[19]. However, MAML's benefits often come with computational challenges, such as instability and sensitivity to hyperparameters[11], and memorization effects during training[1]. To improve ICL, meta training strategies such as those introduced by Chen et al.[9] (MetaICT) and Min et al.[8] (MetaICL) have emerged, leveraging diverse prompt exemplars to warm up pre-trained LLMs. These approaches, however, primarily perform multitask fine-tuning rather than bi-level optimization. Recent related work attempts to apply MAML to improve prompt tuning through learned soft embedding of tokens[20]. In contrast, our work performs generalization on model parameters to improve multimodal ICL.

2.3 Major depression disorder detection

Current techniques for the detection of MDD can be classified into two primary categories: unimodal and multimodal approaches. Unimodal models depend on a single data source, such as text, which reflects linguistic indicators of depression[21], or audio, which utilizes prosodic and acoustic signals from speech[22], to generate predictions. Although these models have shown promising results, they often do not encompass the complete range of depressive symptoms, which may present through various behavioural and communicative modalities. In contrast, multimodal approaches seek to integrate various signals, such as audio and text, to enhance the understanding of depression. For instance, recent methods have used multimodal attention and spatio-temporal feature fusion to combine linguistic, facial, and acoustic features for improved prediction[23]. Despite their success, these systems still face significant challenges in generalization, especially across domains, languages, or institutions, due to limited annotated data and variability in real-world clinical settings.

Previous approaches aim to improve ICL by applying multitask fine-tuning to meta-train LLMs on a range of prompt-based tasks. However, these methods update model parameters continuously without using the two-step optimization process that helps models learn how to adapt, limiting their generalization ability. In contrast, our method adopts a bi-level optimization process to meta-train models. Unlike past work, which concentrates on NLP text classification tasks, we combine text and audio inputs using a cross-modal alignment module, and we test our model in low-resource and cross-lingual conditions. These settings closely reflect real-world clinical challenges, making our approach more practical and effective in improving generalization and adaptation for MDD classification tasks.

3 Methodology

In this section, we present our meta training method for

enhancing in-context learning of LLMs to improve MDD classification. We begin by formalizing the bi-level meta training formulation and outlining our problem setup. We then introduce our proposed method and detail its main components, including feature encoding, task adaptation, aggregated meta-update phases, and optimization strategies.

3.1 Problem statement

In-context learning treats the LLM as an inference-only model by prepending j labeled exemplars to each test prompt, where j denotes the number of exemplar samples sampled from the same task. We mirror this set-up during meta training and extend this to the multimodal data distribution setting. Given training sample $\mathcal{X}$ and task label $\mathcal{Y}$, for each pair of training examples $\{x_i = (x_i^{\mathrm{t}}, x_i^{\mathrm{a}}), y_i\} \in (\mathcal{X}, \mathcal{Y})$ in a task C, where x_i^{t} and x_i^{a} represent the text and audio embeddings, we uniformly sample a set of supports $\mathcal{S} = \{(x_1^{\mathrm{t}}, x_1^{\mathrm{a}}, y_1), (x_2^{\mathrm{t}}, x_2^{\mathrm{a}}, y_2), \ldots, (x_j^{\mathrm{t}}, x_j^{\mathrm{a}}, y_j)\} \subset C$ and construct the joint prompt $P = [x_1^{\mathrm{t}}, x_1^{\mathrm{a}}, y_1, x_2^{\mathrm{t}}, x_2^{\mathrm{a}}, y_2, \ldots, x_j^{\mathrm{t}}, x_j^{\mathrm{a}}, y_j, x_i^{\mathrm{t}}, x_i^{\mathrm{a}}]$. The j support examples are pre-appended in P along with the final train pair $(x_i^{\mathrm{t}}, x_i^{\mathrm{a}})$. We then compute the standard classification loss $\ell(f_\theta(P), y_i)$ using the target label y_i, and meta-train the pre-trained LLM f over all tasks C. We update the model parameter θ via

$$\theta \leftarrow \theta - \nabla_\theta \ell(f_\theta(P), y_i), \forall (x_i^{\mathrm{t}}, x_i^{\mathrm{a}}, y_i) \in C \tag{1}$$

In Section 3.3, we extend this bi-level formulation to our prompt-MAML method.

3.2 Feature encoding

Text and audio inputs are processed through separate encoding pathways that generate embeddings aligned with the representation space of the LLM. Text inputs are tokenized using the LLM's pretrained tokenizer, resulting in embedding $x_i^{\mathrm{t}} \in \mathbb{R}^{d_{\mathrm{LLM}}}$, where d_{LLM} represents the hidden dimension of the LLM. Audio inputs, denoted as $\tilde{x}_i^{\mathrm{a}}$, are initially encoded into intermediate feature $M \in \mathbb{R}^{T \times d_{\mathrm{a}}}$ using a frozen pretrained speech encoder, such as Whisper[24]. $\tilde{x}_i^{\mathrm{a}}$ denotes the audio signal's intermediate representation prior to projection, T is the number of temporal frames (time steps) in the encoded speech sequence, and d_{a} denotes the dimensionality of the speech-encoder feature vector. These features are then projected into the LLM's embedding space using a trainable Q-Former[12, 25], a lightweight transformer module that facilitates cross-modal alignment through windowed self-attention. The Q-Former converts M into audio embedding represented as $x_i^{\mathrm{a}} \in \mathbb{R}^{d_{\mathrm{LLM}}}$. After encoding, both modalities are concatenated to form a unified prompt sequence. Each sample is represented as $x_i = (x_i^{\mathrm{t}}, x_i^{\mathrm{a}})$, which is flattened and appended in the format $P = [x_1^{\mathrm{t}}, x_1^{\mathrm{a}}, y_1, x_2^{\mathrm{t}}, x_2^{\mathrm{a}}, y_2, \ldots, x_j^{\mathrm{t}}, x_j^{\mathrm{a}}, y_j, x_i^{\mathrm{t}}, x_i^{\mathrm{a}}]$. This alignment enables the LLM to interpret audio embeddings as contextual prompts, similar to additional text tokens.

3.3 Meta-in-context learning

Standard in-context meta training methods such as MetaICL[8] and MetaICT[9] perform optimization in line with multi-task fine-tuning. Mathematically, they solve the single-stage objective

$$\min_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_\theta) \tag{2}$$

where $\mathcal{T}_i$ denotes the i-th task sampled from the probability distribution $p(\mathcal{T})$, with p denoting the probability and $\mathcal{T}$ denoting the set of all tasks considered in meta-training. $\mathcal{L}_{\mathcal{T}i}(\cdot)$ is the loss function evaluated on task $\mathcal{T}_i$.

In contrast, prompt-MAML follows the bi-level optimization meta training of MAML-en-LLM[11], as illustrated in Fig. 1, learns a single initialization θ that can rapidly adapt across a diverse collection of task distributions $\mathcal{T}_i$, drawing from the paired $(x^{\mathrm{t}}, x^{\mathrm{a}})$ examples, by structuring meta training into two nested phases. In Fig. 1, Prompt_n denotes the n-th input sample (prompt) consisting of the input pair $(x_n^{\mathrm{t}}, x_n^{\mathrm{a}})$ and its task label y_n. y_n is the groundtruth label for the n-th prompt, and $\hat{y}_n$ is the corresponding predicted task

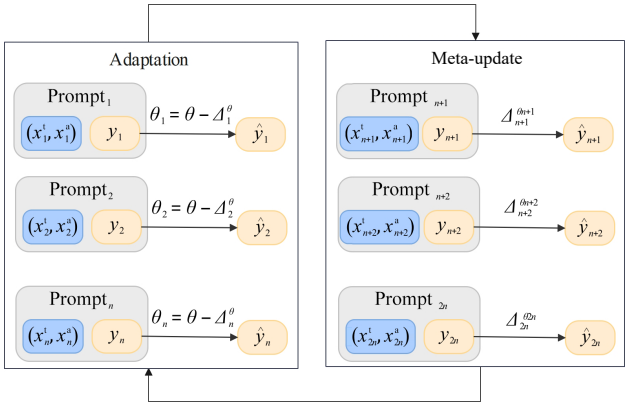

Fig. 1 Overview of the prompt-MAML bi-level training loop, showing the inner loop task adaptation on support prompts and the outer loop meta-updates across query prompts.

label. Δ_n^θ denotes the gradient-based inner-loop update computed from Prompt_n with respect to the parameter θ. $\theta_n = \theta - \Delta_n^\theta$ is the parameter adapted after the update, and n is the number of tasks used to compute the updated parameters. The inner loop performs task-specific adaptation, taking a few gradient steps on each task's support set to explore the local parameter space, while the outer loop aggregates the resulting gradients from each task's query set to meta-update θ, thereby controlling the overall update magnitude. Formally, this yields the bi-level optimization objective

$$\min_{\theta} \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}\left(f_{\theta - \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)}\right) \tag{3}$$

3.3.1 Task adaptation

In the inner loop, we adapt the shared initialization θ to each multimodal task by sampling n tasks to a distribution of tasks $\mathcal{T}_i \sim p(\mathcal{T})$, where $p(\mathcal{T})$ represents a probability distribution over diverse tasks. It is important to recognize that a task denotes a randomly sampled batch of training prompts, as outlined in Ref. [8]. For each task $\mathcal{T}_i$, we perform k gradient descent steps on its support set $\mathcal{S}_i$. Here, the support-set size k exactly determines the number of adaptation steps. Intuitively, increasing the number of tasks n widens this exploration of the parameter space. We compute the adapted parameters via

$$\theta_i \leftarrow \theta - \alpha \nabla_\theta \ell_{\mathcal{T}_i}(f_\theta) \tag{4}$$

where $\mathcal{T}_i \sim p(\mathcal{T})$ is sampled from the full task distribution, θ represents the model parameter, α is the adaptation learning rate, and ℓ is the cross-entropy loss.

3.3.2 Meta-update

In the outer loop, prompt-MAML refines the original initialization θ by leveraging a distinct query set $\mathcal{Q}_i$ for each task $\mathcal{T}_i$, where $\mathcal{Q}_i$ is disjoint from its support set $\mathcal{S}_i$ and $|\mathcal{Q}_i| = |\mathcal{S}_i| = k$. For every task, we evaluate its adapted parameter θ_i (from Formula (4)) on $\mathcal{Q}_i$. We then aggregate these per-task losses across all n tasks and take a second-order gradient step on θ to update the unadapted parameters. The meta-update is thus applied to the initial parameters based on the second-order gradients derived from the task-specific adapted parameters θ_i'. Mathematically, the second-order update can be represented as

$$\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \ell_{\mathcal{T}_i}(f_{\theta_i}) \tag{5}$$

where β is the learning rate for the outer-loop optimization.

3.4 Shared adaptive optimizer

Prompt-MAML's bi-level training creates a dual optimization challenge: We must run an inner adaptation and an outer meta-update for the same parameter θ. The dual optimization problem presents distinct challenges in LLMs, as the selection of optimizers significantly influences generalization. Using adaptive optimizers like AdamW is critical for rapid convergence, yet naively resetting their internal state between inner and outer loops can destabilize training[11]. Concretely, standard adaptive optimizers maintain two bias-corrected moment estimates at step t

$$m_t^B \leftarrow \frac{\beta_1 \cdot m_{t-1} + (1-\beta_1)\cdot g_t}{1-\beta_1^t},$$

$$v_t^B \leftarrow \frac{\beta_2 \cdot v_{t-1} + (1-\beta_2)\cdot g_t^2}{1-\beta_2^t},$$

where t denotes the optimization step, g_t is the gradient at step t, computed from observations sampled in a mini-batch B, m_t and v_t are the first-moment and second-moment moving averages of the gradients, and β_1 and β_2 are the exponential decay hyperparameters controlling the moving-average rates for the first and second moments, respectively. Crucially, when we reset the optimizer after each meta-update, both m_t and v_t are re-initialized. Over long inner-loop runs, especially as the model approaches a minimum, this repeated resetting causes later gradient steps to carry disproportionately more weight, destabilizing fine-tuning.

To address this, we share optimizer parameters between the inner and outer loops. Rather than maintaining two independent sets of (m, v), we keep a single, continuously updated pair of moving averages. This shared state ensures that inner updates benefit from the full history of past gradients, smoothing optimization and improving convergence.

3.5 Meta training with prompt-MAML

Figure 2 illustrates the detailed inner-loop adaptation and outer-loop meta-update steps, while Algorithm 1 lays out the prompt-MAML training steps. Similar to MAML-en-LLM[11], we denote our configuration as prompt-MAML-2k-n, where n is the number of tasks used to compute the adapted parameters and k is the number of task batches used for the inner-loop adaptation stage and for computing the outer-loop meta-updates. For example, when $|\mathcal{S}| = |\mathcal{Q}| = k = 1$ and $n = 1$, where $\mathcal{S}$ is the support set and $\mathcal{Q}$ is the disjoint

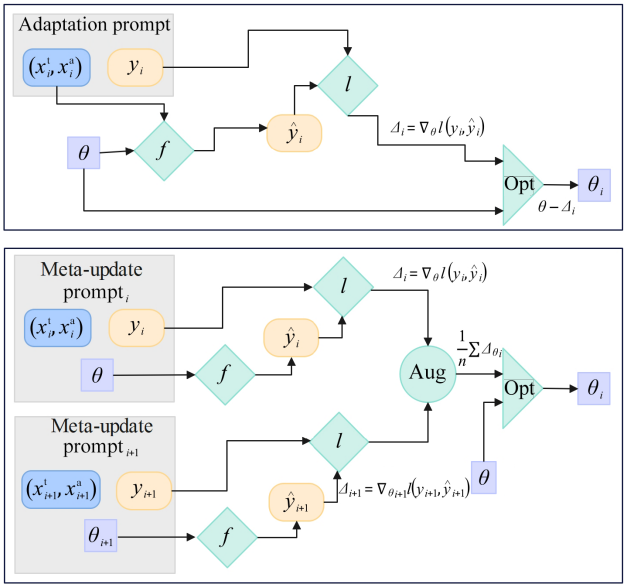

Fig. 2 Detailing one full inner-loop adaptation step and two example outer-loop meta-update steps. Aug denotes the aggregated meta-update operation, and Opt denotes an optimization step. The figure is adapted from the schematic in MAML-en-LLM[11].

query set used for the outer-loop meta-update, we write prompt-MAML-2-1, which yields meta-updates every $2kn = 2\times1\times1 = 2$ steps. Similarly, in prompt-MAML-2-4 (i.e., $k = 1$, and $n = 4$), meta-updates occur every $2\times1\times4 = 8$ steps.

4 Experiment

In this section, we introduce the dataset details, experiments settings, and comparison methods.

4.1 Dataset description

We used four multimodal benchmark datasets, distress analysis interview corpus-Wizard-of-Oz (DAIC-WOZ)[26], emotional audio-textual depression (EATD) corpus[27], Chinese multimodal depression corpus (CMDC)[28], and multimodal open dataset for mental-disorder analysis (MODMA)[29], each providing synchronized audio-transcript pairs with expert annotations for depression. From these, we derive meta-tasks for binary MDD detection and severity classification, formulated across all datasets. All subjects are split at the dataset level so that no individual appears in both meta-train and meta-test. The testing tasks consist of tasks with similar domains (tasks drawn from the same distribution) in the training set and tasks sampled from unseen domains (tasks drawn from the different distributions, e.g., a different

Algorithm 1 Prompt-MAML

Input:

Training dataset $\mathcal{X}$, label $\mathcal{Y}$, training sample $\{x_i = (x_i^{\mathrm{t}}, x_i^{\mathrm{a}}) \in \mathcal{X};\ y_i \in \mathcal{Y}\}$, total step T, model f pre-trained parameter θ_{pt}, learning rates α and β, hyperparameters β_1 and β_2, weight decay coefficient λ, stability constant ϵ, support/query set size n, meta-gradient $g'^{(t)}$, and loss function ℓ.

Output: Meta-trained parameter θ

$\theta \leftarrow \theta_{\mathrm{pt}}$

$t \leftarrow 0,\ m \leftarrow 0,\ v \leftarrow 0$

for $t = 1$ to T **do**

  Sample support set of size n: $\{x_i\}^n,\ \{y_i\}^n \in \{\mathcal{X},\ \mathcal{Y}\}$

  for $i = 1$ to n **do**

$$g_i^{(t)} \leftarrow \nabla_\theta \ell(f_\theta(x_i),\ y_i)$$

$$m \leftarrow \frac{\beta_1 \cdot m + (1-\beta_1)\cdot g_i^{(t)}}{1-\beta_1^t}$$

$$v \leftarrow \frac{\beta_2 \cdot v + (1-\beta_2)\cdot (g_i^{(t)})^2}{1-\beta_2^t}$$

$$\theta_i \leftarrow \theta - \left[\alpha \cdot \frac{m}{\sqrt{v}+\epsilon} + \alpha \cdot \lambda \cdot \theta\right]$$

  end for

  Sample query set of size n: $\{x_j\}^n,\ \{y_j\}^n \in \{\mathcal{X},\ \mathcal{Y}\}$

  for $i = 1$ to n **do**

$$g'^{(t)} \leftarrow \frac{1}{n}\sum_{j=1}^{n} \nabla_\theta \ell(f_{\theta_i}(x_j),\ y_j)$$

$$m \leftarrow \frac{\beta_1 \cdot m + (1-\beta_1)\cdot g'^{(t)}}{1-\beta_1^t}$$

$$v \leftarrow \frac{\beta_2 \cdot v + (1-\beta_2)\cdot (g'^{(t)})^2}{1-\beta_2^t}$$

$$\theta \leftarrow \theta - \left[\beta \cdot \frac{m}{\sqrt{v}+\epsilon} + \beta \cdot \lambda \cdot \theta\right]$$

  end for

end for

Return θ

dataset) in the training set. Table 1 summarizes each dataset's statistics, splits, and train/test tasks. Specifically, we consider five train/test task pairs under complete data setting and limited data setting: (1) binary → binary (B→B), (2) binary → severity (B→S), (3) severity → binary (S→B), (4) severity → severity (S→S), and (5) high-resource (HR) → low-resource (LR) (H→L) transitions. In Table 1, the numbers in parentheses indicate sample counts, "train/val" gives the sizes of the training and validation splits, and single numbers in the "Test setting (test)" and "Unseen" columns give the sizes of the

Table 1 Summary of train/test task configuration used in meta-training experiment. Each task specifies the train setting (train/val), test setting (test), and unseen evaluation datasets, with splits stratified by MDD:NC ratios across MDD datasets.

Train/test task	Train setting (train/val)	Test setting (test)	Unseen
Binary → Binary	DAIC-WOZ (76/16), CMDC (32/7), and EATD (40/10)	EATD (52)	MODMA (27)
Binary → Severity	DAIC-WOZ (76/16), CMDC (32/7), and EATD (40/10)	DIAC-WOZ (48)※	MODMA (27)
Severity → Binary	DAIC-WOZ (76/16), CMDC (32/7), and EATD (40/10)	CMDC (39)	MODMA (27)
Severity → Severity	DAIC-WOZ (76/16), CMDC (32/7), and EATD (40/10)	DIAC-WOZ (48)	MODMA (27)
HR → LR	DAIC-WOZ (76/16)	CMDC (39)	MODMA (27)

Note: DIAC-WOZ (48)※ indicates the binary→severity test split, distinct from the unstarred DIAC-WOZ (48) used in severity→severity.

corresponding evaluation splits.

4.2 Experiment setting

For all our experiments, we utilize a pre-trained GPT-2 Medium[14] model comprising 355 million parameters. Training is conducted on a NVIDIA T4 GPU using Google Colab Pro, leveraging PyTorch and HuggingFace pre-trained checkpoints. Meta-training is performed for 20 000 steps. The input training sequence length is fixed at 1024 tokens, and the batch size is fixed at 1 during meta-training. For both the inner-loop and outer-loop learning rates, we use a fixed value of 1×10^{-5}. The size of support and query sets per task is set to 1, and the number of tasks per batch is either 1 (prompt-MAML-2-1) or 4 (prompt-MAML-2-4), implying that meta-updates occur every 2 or 8 steps, respectively. For both the inner and outer loops, we use AdamW for optimization with identical hyperparameters. To ensure consistent optimization dynamics, the final moment states of the inner optimizer are copied to the outer optimizer prior to each meta-update step. The training seed is set at 100 across all experiments. During inference, the test sequence length is set at 256 tokens, with the batch size of 16 exemplars, or whichever is lower. The overall training time for prompt-MAML is approximately 22 h.

4.3 Comparative method

We evaluate our proposed method in comparison to various models and prompt setups. In line with the evaluation frameworks proposed by Min et al.[8] and Sinha et al.[11], we incorporate both standard models and channel models[8] for the evaluation of all experiments. Although the related works focus on NLP text datasets, this evaluation paradigm is modality-agnostic and remains applicable to our multimodal setting. We align this framework to our experimental design, adapting both standard and channel training objectives accordingly. In standard models, provided with exemplars $(x_1, y_1, x_2, y_2, \ldots, x_j, y_j)$ and a target input instance $x_i = (x_i^{\mathrm{t}}, x_i^{\mathrm{a}})$, the training and inference procedures optimize the model parameter θ by utilizing cross-entropy loss ℓ as follows:

$$\theta = \theta - \nabla_\theta \ell\Big(f_\theta((x_1^{\mathrm{t}}, x_1^{\mathrm{a}}), y_1, (x_2^{\mathrm{t}}, x_2^{\mathrm{a}}), y_2, \ldots, (x_j^{\mathrm{t}}, x_j^{\mathrm{a}}), y_j, (x_i^{\mathrm{t}}, x_i^{\mathrm{a}})), y_i\Big) \tag{6}$$

At inference, the model predicts the label as

$$y_i = \arg\max_{y\in\mathcal{C}} P\Big(y\,\Big|\,(x_1^{\mathrm{t}}, x_1^{\mathrm{a}}), y_1, (x_2^{\mathrm{t}}, x_2^{\mathrm{a}}), y_2, \ldots, (x_j^{\mathrm{t}}, x_j^{\mathrm{a}}), y_j, (x_i^{\mathrm{t}}, x_i^{\mathrm{a}})\Big) \tag{7}$$

where $\mathcal{C}$ denotes the set of all possible classes. Unlike standard models, channel models treat the problem as a generative task by reversing the prompts and labels. During training, labels precede their corresponding inputs, and the target label is included within the prompt, whereas the target instance becomes the output to be predicted. Formally, training optimizes the parameters using

$$\theta = \theta - \nabla_\theta \ell\Big(f_\theta(y_1, (x_1^{\mathrm{t}}, x_1^{\mathrm{a}}), y_2, (x_2^{\mathrm{t}}, x_2^{\mathrm{a}}), \ldots, y_j, (x_j^{\mathrm{t}}, x_j^{\mathrm{a}}), y_i), (x_i^{\mathrm{t}}, x_i^{\mathrm{a}})\Big) \tag{8}$$

At inference, prompts with all possible labels $c \in \mathcal{C}$ are constructed, and the label yielding the highest conditional probability of generating the target instance $(x_i^{\mathrm{t}}, x_i^{\mathrm{a}})$ is selected

$$y_i = \arg\max_{c\in\mathcal{C}} P\Big((x_i^{\mathrm{t}}, x_i^{\mathrm{a}})\,\Big|\,y_1, (x_1^{\mathrm{t}}, x_1^{\mathrm{a}}), y_2, (x_2^{\mathrm{t}}, x_2^{\mathrm{a}}), \ldots, y_j, (x_j^{\mathrm{t}}, x_j^{\mathrm{a}}), c\Big) \tag{9}$$

For both standard and channel models, we evaluate three settings based on the models used and the prompt structures.

(1) No ICL: The prompt consists solely of the target instance x_i, with no exemplars. The model directly estimates the label y_i.

(2) RawLM: The prompt includes exemplars

followed by the target sample to estimate the target label. The model employed is a pre-trained, standard large LLM. For channel models, the prompt structure is reversed as illustrated in Eq. (9).

(3) MetaICL: We replicate the MetaICL training and inference procedures described in Ref. [8], adapted explicitly to match our experimental settings and prompt structures.

4.3.1 Evaluation criterion and metric

To evaluate model performance across all tasks, we report macro-F1 score. Macro-F1 is particularly suitable for binary depression and severity classification tasks, where class imbalance is significant. We compute results over five random seeds, and include both the average and worst-case performance. We also report win-rates where a model is considered to win a task setting if it outperforms alternatives in both average and worst-case metrics, and the results are statistically significant. The values highlighted in bold in Table 2 and Table 3 indicate the best-performing methods for each data setting.

5 Result and Discussion

5.1 Performance evaluation

To evaluate the robustness of our framework in reflecting real-world clinical scenarios, we perform experiments under two data environments: complete data setting and limited data setting. These settings reflect common challenges in MDD applications, where labeled datasets often vary in size and quality. In the complete data setting, we use the full training split of each dataset. All training and test splits are defined at the dataset level to prevent subject overlap. Table 1 summarizes the dataset statistics used for training, validation, and testing across tasks.

In the limited data setting, we subsample 10% of the training set from each dataset using fixed random seeds for consistency. Sampling is equally stratified to preserve the original class distribution, thereby minimizing bias due to class imbalance. To assess generalization, we evaluate model performance first on all test tasks and then on tasks drawn from unseen domains. Since in-context learning is sensitive to prompt construction, we run all experiments across five random seeds. For each test instance, exemplars are drawn from the training set of the same task.

5.1.1 Performance on all tasks

We present the results of our method across five tasks under both complete and limited data settings in Table 2. The prompt-MAML models, both standard and channel-based, are compared against MetaICL, Raw LLMs, and No ICL baselines. In standard models, prompt-MAML outperforms MetaICL in 3 out of 5 tasks, achieving a win rate of 0.60 in both settings. Channel models perform significantly better in standard models, outperforming MetaICL in all tasks with a win rate of 1.00. In the limited data setting, channel variants of prompt-MAML demonstrate robust performance surpassing channel MetaICL in 4 of 5 tasks, achieving a win rate of 0.80.

Across both data settings, our prompt-MAML models, especially in the channel configuration, demonstrate superior generalization capabilities over MetaICL. This is especially important in scenarios with limited data, underscoring the effectiveness of meta-training in enabling adaptation to various MDD related tasks. The results observed in both complete and limited task settings demonstrate that, although prompt-MAML performs meta-updates every $2kn$ batches, it is not affected by the amount of training data available.

5.1.2 Performance on unseen task

We evaluate the generalization capability of our approach across unseen test domains under both complete and limited data settings. Unseen domains are disjoint from the meta-training phase in both distribution and task type. Table 3 presents results across the five evaluation tasks. In the complete data setting, prompt-MAML standard models show better performance than MetaICL in 3 out of 5 tasks, achieving a win rate of 0.60, with particularly better performance in binary →binary and severity →binary. However, they underperform in binary →severity task, consistent with patterns observed on seen tasks. Channel models exhibit a notable advantage, as prompt-MAML outperforms MetaICL in 4 out of 5 tasks, yielding a win rate of 0.80. Improvements are notably observed in severity →binary, demonstrating enhanced cross-domain generalization.

Similarly, with limited data, the standard models of prompt-MAML outperform MetaICL in 4 out of 5 tasks, achieving a win rate of 0.80, and achieving the best performance on severity →binary. In channel models, we observe that prompt-MAML outperforms MetaICL in 3 out of 5 tasks, achieving a win rate of 0.60. Notably, they improve performance on binary → severity and severity →severity, suggesting that

Table 2 Overall performance evaluation on all tasks using complete data and limited data settings. Best values are bolded.

(a) Performance on all tasks using complete data setting.

Model	Average/worst-case macro-F1				
	Binary → Binary	Binary → Severity	Severity → Binary	Severity → Severity	HR → LR
No ICL	33.33	33.43	25.72	37.42	31.40
Raw LLM	34.36/34.16	33.51/32.72	32.86/31.62	34.36/34.16	35.18/34.36
MetaICL	42.85/42.80	33.84/32.02	49.59/46.06	**42.56/40.00**	**42.85/41.80**
Prompt-MAML-2-1	41.47/41.07	18.81/18.67	**50.01/46.63**	40.82/39.95	41.33/40.52
Prompt-MAML-2-4	**42.98/42.08**	**38.20/37.26**	44.66/42.52	40.03/39.68	41.72/40.21
Channel no ICL	33.64	46.59	33.32	33.94	36.64
Channel raw LLM	45.68/45.01	45.27/44.89	38.66/37.51	45.68/45.01	41.45/40.35
Channel MetaICL	50.40/49.76	51.78/49.57	48.40/47.76	49.76/48.33	46.65/45.44
Channel prompt-MAML-2-1	51.08/50.04	**54.11/51.21**	**52.91/51.45**	**50.55/49.54**	**48.81/47.57**
Channel prompt-MAML-2-4	**51.52/50.71**	53.86/51.03	52.02/51.46	50.50/49.05	48.43/47.06

(b) Performance on all task using limited data setting.

Model	Average/worst-case macro-F1				
	Binary → Binary	Binary → Severity	Severity → Binary	Severity → Severity	HR → LR
No ICL	33.33	33.43	25.72	37.42	31.40
Raw LLM	34.36/34.16	33.51/32.72	32.86/31.62	34.36/34.16	35.18/34.36
MetaICL	38.85/37.80	**37.53/35.72**	31.59/29.06	**41.56/40.00**	39.18/38.47
Prompt-MAML-2-1	**42.47/42.37**	33.81/33.67	**44.71/42.63**	38.82/37.95	39.33/38.52
Prompt-MAML-2-4	42.15/40.98	34.29/34.26	37.66/32.52	41.03/40.68	**39.72/38.21**
Channel no ICL	33.64	46.59	33.32	33.94	36.64
Channel raw LLM	45.68/45.01	45.27/44.89	38.66/37.51	45.68/45.01	41.45/40.35
Channel MetaICL	**48.53/47.16**	48.78/47.57	44.40/42.76	46.76/46.33	45.65/43.44
Channel prompt-MAML-2-1	48.08/47.04	46.11/45.21	**46.91/45.45**	**47.55/46.54**	46.81/45.57
Channel prompt-MAML-2-4	47.52/47.71	**49.86/48.03**	46.02/45.46	47.50/46.05	**49.43/48.06**

channel models benefit more from meta-learned initializations under low-resource conditions. Across both settings, prompt-MAML demonstrates strong generalization on unseen tasks. These gains support the hypothesis that meta learning enables the model to explore a broader parameter space, leading to better initializations and stronger adaptation, even across heterogeneous datasets.

5.2 Effect of modality and task complexity

5.2.1 Modality-specific analysis

As shown in Fig. 3, we analyze the impact of the modality on performance in unseen domains by isolating audio and text modalities. Across all tasks, the text modality consistently outperforms audio, with MetaICL outperforming prompt-MAML with a win rate of 0.60 in standard models. In channel models, prompt-MAML outperforms MetaICL in 4 out of 5 tasks, particularly excelling in the HR → LR task when using text. These observations highlight the effectiveness of pre-trained LLMs in capturing linguistic signals and suggest that depressive indicators may be more distinctly expressed through text rather than speech patterns, consistent with prior research. The gap between modalities probably reflects the inherent complexity of audio data, which is more susceptible to noise, language differences, and recording inconsistencies, whereas text provides more structured and reliable emotional markers. Within our prompt-MAML, these results underscore the critical role of text in guiding exemplar selection and generalization, especially in low-resource or cross-domain scenarios where precise language-driven indicators are essential for adaptation.

5.2.2 Task complexity and exploration state

We begin our discussion on exploration states by outlining the specific task settings. Our study comprises five task configurations, all classification-based, with the final one, HR → LR, representing a composite setting distinguished by data volume and domain shift. While all tasks share the classification format, they vary in complexity. For instance, binary

Table 3 Performance evaluation on unseen task using complete data and limited data settings. Best values are bolded.

(a) Performance on unseen task utilizing complete data setting.

Model	Average/worst-case macro-F1				
	Binary → Binary	Binary → Severity	Severity → Binary	Severity → Severity	HR → LR
No ICL	28.96	35.42	33.80	28.96	28.96
Raw LLM	29.61/25.17	41.84/33.57	42.59/33.67	29.61/25.17	29.14/25.10
MetaICL	38.94/36.24	**59.24/54.41**	78.74/73.33	**36.75/32.49**	42.25/38.76
Prompt-MAML-2-1	**42.53/41.60**	33.39/32.29	**80.28/77.00**	31.54/26.64	34.81/33.64
Prompt-MAML-2-4	40.41/39.66	34.96/34.18	60.94/52.86	32.68/27.84	**43.49/42.13**
Channel no ICL	30.96	41.69	33.72	30.96	30.96
Channel raw LLM	42.89/39.34	52.01/47.39	38.66/34.41	42.89/39.34	42.79/39.19
Channel MetaICL	**49.91/47.51**	48.55/46.24	55.90/45.13	44.88/43.21	48.81/47.79
Channel prompt-MAML-2-1	48.24/47.14	**59.36/56.80**	64.11/58.72	46.03/44.41	**51.18/48.81**
Channel prompt-MAML-2-4	49.58/47.73	57.29/54.45	**64.26/59.42**	**46.52/44.76**	50.92/47.22

(b) Performance on unseen task utilizing limited data setting.

Model	Average/worst-case macro-F1				
	Binary → Binary	Binary → Severity	Severity → Binary	Severity → Severity	HR → LR
No ICL	28.96	35.42	33.80	28.96	28.96
Raw LLM	29.61/25.17	41.84/33.57	42.59/33.67	29.61/25.17	29.14/25.10
MetaICL	38.83/35.45	38.49/32.39	33.85/33.47	35.86/32.30	**43.27/41.57**
Prompt-MAML-2-1	**38.99/36.59**	34.97/32.87	53.72/42.86	34.55/31.58	41.00/39.03
Prompt-MAML-2-4	38.46/36.35	**42.08/38.91**	**57.82/55.97**	**39.50/38.31**	42.63/40.61
Channel no ICL	30.96	41.69	33.72	30.96	30.96
Channel raw LLM	42.89/39.34	52.01/47.39	38.66/34.41	42.89/39.34	42.79/39.19
Channel MetaICL	**46.59/44.33**	43.34/42.09	53.23/34.40	**53.30/49.26**	45.76/42.74
Channel prompt-MAML-2-1	45.44/44.35	52.17/48.43	**54.54/37.44**	45.00/43.39	**48.78/46.44**
Channel prompt-MAML-2-4	45.98/44.80	**53.38/48.84**	54.52/37.39	45.47/44.03	46.31/43.21

classification is relatively simple because of its distinct categories and narrow label space. In contrast, severity classification necessitates that the model identifies more subtle distinctions in depressive symptom levels, which adds to the complexity. The HR → LR setting presents significant challenges, merging data scarcity with the need for cross-lingual adaptation between English and Chinese datasets.

To understand how this complexity affects learning, we compare two versions of our method: prompt-MAML-2-1, which updates using a single task, and prompt-MAML-2-4, which updates across multiple tasks. We find that prompt-MAML-2-1 performs well on simpler tasks, while prompt-MAML-2-4 yields better results on complex tasks by exploring a broader parameter space, though at the cost of slower convergence.

Although our task setup ensures no subject overlap between training and testing, we observe that not all tasks respond equally to meta-training, particularly in low-resource unseen domains. In tasks like binary → severity, standard MetaICL and prompt-MAML models sometimes underperform relative to Raw LLMs. This indicates that meta-training, although beneficial, may sometimes erase valuable pre-trained knowledge, particularly when there is a misalignment in complexity between the training and test tasks. The findings highlight the necessity of aligning the difficulty of meta-training tasks with the requirements of the target setting, especially in resource-constrained or cross-domain settings.

5.3 Few-shot and zero-shot adaptation

To further evaluate adaptability, we test our models on few-shot adaptation in unseen domains using 0-shot, 4-shot, and 16-shot settings as illustrated in Fig. 4. For each condition, prompts and adaptation samples are drawn from the training split of the test datasets, ensuring that the target instance is never repeated in the exemplars. The model is fine-tuned using 0, 4, and 16 adaptation examples, respectively, with a single training pass (one epoch) over the available samples,

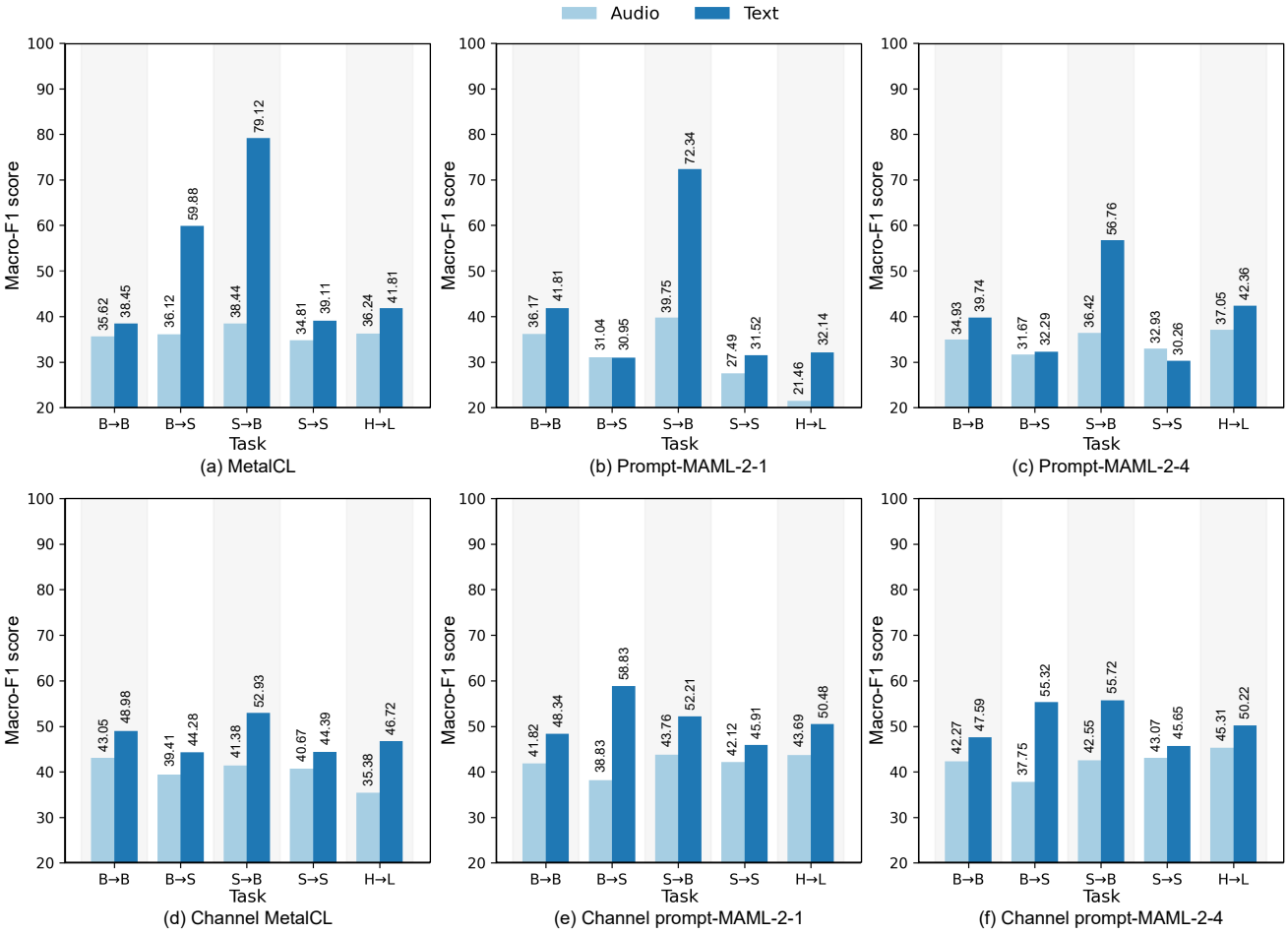

Fig. 3 Modality-specific performance across tasks in unseen domain. We compare macro-F1 scores of models trained with either audio-only or text-only exemplars across five task configurations: (B→B), (B→S), (S→B), (S→S), and (H→L).

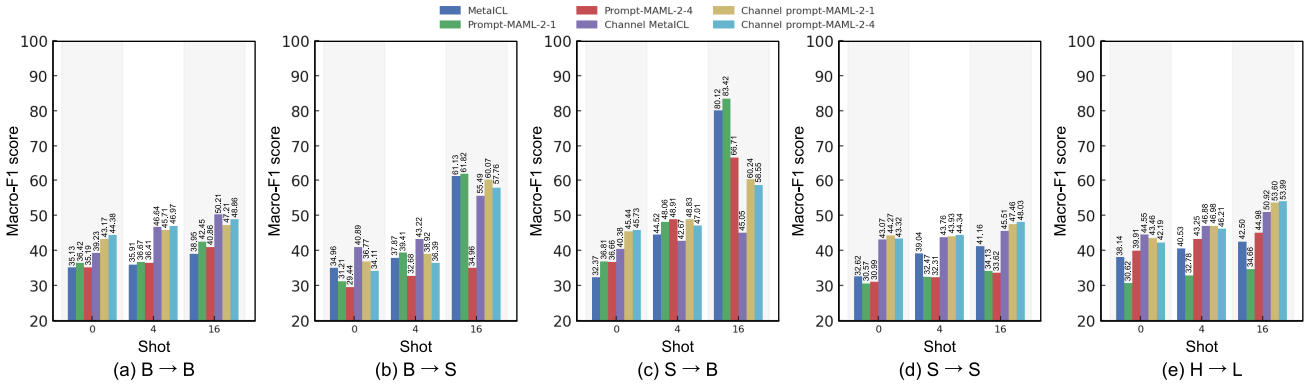

Fig. 4 Few-shot adaptation performance across tasks under 0-shot, 4-shot, and 16-shot settings. We evaluate prompt-MAML and baseline models in unseen domains by varying the number of adaptation examples per task. Performance is measured using macro-F1 across five classification settings.

and a low learning rate to simulate minimal supervision. It should be noted that before performing any adaptation, the standard and channel models used are identical to those trained in the complete data setting.

This setup reflects the real-world challenge of adapting to new clinical or linguistic domains with very limited labelled data. Across tasks, prompt-MAML standard models consistently outperform MetaICL in few-shot adaptation, with a clear

advantage emerging by 16-shot, particularly on severity → binary and binary → binary. Similarly, channel-based prompt-MAML models outperform channel MetaICL in the majority of tasks and shot settings, with prompt-MAML-2-4 attaining the highest performance on complex tasks such as HR → LR and severity → severity. The results indicate that meta-training with broader exploration enables the model to obtain more transferable initializations, thereby improving its generalization capabilities with minimal updates. This represents a notable advantage of prompt-MAML in low-resource MDD classification.

5.4 Optimizer choice on meta-training stability

To assess the role of optimizer choice in meta-training stability and performance, we conduct ablation studies using different optimizer configurations for our prompt-MAML-2-1 model, shown in Table 4. Similar to prior work, we experiment with combinations of stateless optimizer like stochastic gradient descent (SGD) and adaptive AdamW optimizers for the inner-loop (task adaptation) and outer-loop (meta-update). We evaluate performance on the HR → LR task using a 10% data subset under two random seeds, where seed 10 and seed 20 indicate runs with random seeds of 10 and 20, respectively. We compare five configurations: SGD+SGD, SGD+AdamW, AdamW+SGD, AdamW+AdamW†, where † indicates setting utilizing AdamW without moment sharing, and AdamW+AdamW indicates the setting with moment parameter sharing. We also report results for MetaICL (AdamW), Raw LLM, and No ICL as baselines.

We observe that using stateless optimizers results in lower performance. Switching to adaptive optimizers, particularly AdamW in both loops, yields the strongest performance, confirming prior findings that adaptive optimizers enhance stability in LLM training. Interestingly, hybrid configurations like SGD (inner) plus AdamW (outer) also lead to improved results over pure SGD, suggesting that retaining gradient state in at least one loop contributes positively. On the other hand, AdamW (inner) plus SGD (outer) performs similarly to pure SGD, highlighting the critical role of the outer loop in meta-updates. These results reinforce the importance of optimizer choice in MAML-style training for MDD classification tasks. They also validate our use of AdamW for both adaptation and meta-updates in all main experiments, including with moment state sharing.

Table 4 Optimizer ablation study on prompt-MAML-2-1 for HR → LR task. Macro-F1 under 10% data using two random seeds. AdamW† indicates no moment sharing.

Method	Optimizer	Macro-F1 score	
		Seed 10	Seed 20
No ICL	—	21.4	24.4
Raw LLM	—	34.3	31.6
MetaICL	SGD	34.8	35.3
	AdamW	34.2	35.5
Prompt-MAML-2-1	SGD+SGD	35.1	34.7
	SGD+AdamW	41.5	39.8
	AdamW+SGD	35.0	34.5
	AdamW+AdamW†	41.5	39.8
	AdamW+AdamW	**42.4**	**43.1**

6 Conclusion

This paper proposed prompt-MAML, a method for meta-training LLMs designed to improve multimodal ICL for MDD classification tasks. Empirical results demonstrate that prompt-MAML consistently surpasses strong baselines in both generalization and adaptation performance under high and limited data settings. Further, by examining a wider parameter space before meta-updates, prompt-MAML enhances the effectiveness of few-shot adaptation to new tasks. In addition to performance improvements, the study gives an in-depth analysis of how task complexity, input modality, and optimizer configurations affect meta-training dynamics. Nevertheless, challenges remain. Bi-level optimization increases training complexity and memory overhead, and stability can be sensitive to exploration strategies and data distribution. Future work will explore more scalable and efficient meta-training methods, extend to broader clinical tasks, and incorporate personalized adaptation strategies to better reflect the diversity of real-world mental health scenarios.

Acknowledgment

The authors would like to acknowledge the support of the University of Science and Technology, Beijing.

References

[1] A. Vettoruzzo, M. R. Bouguelia, J. Vanschoren, T. Rögnvaldsson, and K. C. Santosh, Advances and challenges in meta-learning: A technical review, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 7, pp. 4763–4779, 2024.

[2] W. Li, J. Zhao, L. Zhu, L. Zhang, and H. Wang, TopoPharmDTI: Improving interactions prediction by enhanced deep learning representation for both drug and target molecules, *Tsinghua Science and Technology*, vol. 31, no. 1, pp. 399–417, 2025.

[3] S. Zhu, D. Jian, and D. Xiong, A survey of multilingual neural machine translation based on sparse models, *Tsinghua Science and Technology*, vol. 30, no. 6, pp. 2399–2418, 2025.

[4] B. Wang, Y. Liu, H. Tian, R. Hua, K. Chang, J. Xia, X. Dai, Z. Gao, S. Liu, R. Wang, et al., LLM4DEU: Fine tuning large language model for medical diagnosis in outpatient and emergency department visits of neurosurgery, *Tsinghua Science and Technology*, vol. 30, no. 6, pp. 2487–2504, 2025.

[5] J. Wei, J. Wei, Y. Tay, D. Tran, A. Webson, Y. Lu, X. Chen, H. Liu, D. Huang, D. Zhou, et al., Larger language models do in-context learning differently, arXiv preprint arXiv: 2303.03846, 2023.

[6] Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, B. Chang, et al., A survey on in-context learning, in *Proc. 2024 Conf. Empirical Methods in Natural Language Processing*, Miami, FL, USA, 2024, pp. 1107–1128.

[7] Y. Li, X. Ma, S. Lu, K. Lee, X. Liu, and C. Guo, MEND: Meta demonstration distillation for efficient and effective in-context learning, arXiv preprint arXiv: 2403.06914, 2024.

[8] S. Min, M. Lewis, L. Zettlemoyer, and H. Hajishirzi, MetaICL: Learning to learn in context, in *Proc. 2022 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, WA, USA, 2022, pp. 2791–2809.

[9] Y. Chen, R. Zhong, S. Zha, G. Karypis, and H. He, Meta-learning via language model in-context tuning, in *Proc. 60th Annu. Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*), Dublin, Ireland, 2022, pp. 719–730.

[10] C. Finn, P. Abbeel, and S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in *Proc. 34th Int. Conf. Machine Learning*, Sydney, Australia, 2017, pp. 1126–1135.

[11] S. Sinha, Y. Yue, V. Soto, M. Kulkarni, J. Lu, and A. Zhang, MAML-en-LLM: Model agnostic meta-training of LLMs for improved in-context learning, in *Proc. 30th ACM SIGKDD Conf. Knowledge Discovery and Data Mining*, Barcelona, Spain, 2024, pp. 2711–2720.

[12] J. Pan, J. Wu, Y. Gaur, S. Sivasankaran, Z. Chen, S. Liu, and J. Li, COSMIC: Data efficient instruction-tuning for speech in-context learning, in *Proc. Interspeech 2024*, Kos, Greece, 2024, pp. 4164–4168.

[13] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, in *Proc. 34th Int. Conf. Neural Information Processing Systems*, Vancouver, Canada, 2020, pp. 1877–1901.

[14] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, Language models are unsupervised multitask learners, *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[15] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in *Proc. 36th Int. Conf. Neural Information Processing Systems*, New Orleans, LA, USA, 2022, pp. 24824–24837.

[16] Z. Wang, Y. Jiang, Y. Lu, Y. Shen, P. He, W. Chen, Z. Wang, and M. Zhou, In-context learning unlocked for diffusion models, in *Proc. 37th Int. Conf. Neural Information Processing Systems*, New Orleans, LA, USA, 2023, pp. 8542–8562.

[17] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, InstructBLIP: Towards general-purpose vision-language models with instruction tuning, in *Proc. 37th Int. Conf. Neural Information Processing Systems*, New Orleans, LA, USA, 2023, pp. 49250–49267.

[18] W. Liu, X. Xu, J. Wu, and J. Jiang, Federated meta reinforcement learning for personalized tasks, *Tsinghua Science and Technology*, vol. 29, no. 3, pp. 911–926, 2024.

[19] F. Lux and N. T. Vu, Language-agnostic meta-learning for low-resource text-to-speech with articulatory features, in *Proc. 60th Annu. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Dublin, Ireland, 2022, pp. 6858–6868.

[20] C. Qin, S. Joty, Q. Li, and R. Zhao, Learning to initialize: Can meta learning improve cross-task generalization in prompt tuning?, in *Proc. 61st Annu. Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*), Toronto, Canada, 2023, pp. 11802–11832.

[21] M. E. Aragón, A. P. López-Monroy, L. C. González-Gurrola, and M. Montes-Y-Gómez, Detecting mental disorders in social media through emotional patterns—The case of anorexia and depression, *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 211–222, 2023.

[22] S. S. Leal, S. Ntalampiras, and R. Sassi, Speech-based depression assessment: A comprehensive survey, *IEEE Trans. Affect. Comput.*, vol. 16, no. 3, pp. 1318–1333, 2025.

[23] Y. Wang, Z. Lin, Y. Teng, Y. Cheng, H. Jiang, and Y. Yang, SIMMA: Multimodal automatic depression detection via spatiotemporal ensemble and cross-modal alignment, *IEEE Trans. Computat. Soc. Syst.*, vol. 12, no. 5, pp. 3548–3564, 2025.

[24] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, Robust speech recognition via large-scale weak supervision, in *Proc. 40th International Conference on Machine Learning*, Honolulu, HI, USA, 2023, pp. 28492–28518.

[25] J. Li, D. Li, S. Savarese, and S. Hoi, BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, in *Proc. 40th Int. Conf. Machine Learning*, Honolulu, HI, USA, 2023, pp. 19730–19742.

[26] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, et al., The distress analysis interview corpus of human and

computer interviews, in *Proc. International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, 2014, pp. 3123–3128.

[27] Y. Shen, H. Yang, and L. Lin, Automatic depression detection: An emotional audio-textual corpus and a Gru/Bilstm-based model, in *Proc. 2022 IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Singapore, Singapore, 2022, pp. 6247–6251.

[28] B. Zou, J. Han, Y. Wang, R. Liu, S. Zhao, L. Feng, X. Lyu, and H. Ma, Semi-structural interview-based Chinese multimodal depression corpus towards automatic preliminary screening of depressive disorders, *IEEE Trans. Affect. Comput.*, vol. 14, no. 4, pp. 2823–2838, 2023.

[29] H. Cai, Z. Yuan, Y. Gao, S. Sun, N. Li, F. Tian, H. Xiao, J. Li, Z. Yang, X. Li, et al., A multi-modal open dataset for mental-disorder analysis, *Sci. Data*, vol. 9, no. 1, p. 178, 2022.

Zita Lifelo received the BS and MS degrees in computer science from Copperbelt University, Kitwe, Zambia, in 2012 and in 2016, respectively. She is currently pursuing the PhD degree at University of Science and Technology Beijing, Beijing, China. Her research interests include affective computing, smart health, smart agriculture, meta learning, and continual learning.

Jianguo Ding received the PhD degree in engineering from Faculty of Mathematics and Computer Science, University of Hagen, Hagen, Germany, in 2008. He is currently an associate professor at Department of Computer Science, Blekinge Institute of Technology, Karlskrona, Sweden. He is a senior member of the Association for Computing Machinery. His research interests include cybersecurity, critical infrastructure protection, intelligent technologies, blockchain, distributed systems management and control, and serious game.

Zongjie Wang received the BS and PhD degrees in computer science and technology from University of Science and Technology Beijing, China, in 1992 and 2007, respectively. He is currently an associate professor at School of Computer and Communication Engineering, University of Science and Technology Beijing, China. He has published dozens of papers in journals such as *Food Chemistry*, *Applied Intelligence*, etc. His research interests include artificial intelligence, big data mining, and intelligent manufacturing.

Feifei Shi received the PhD degree from University of Science and Technology Beijing, Beijing, China, in 2024. She is currently a lecturer at School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China. Her research interests include Internet of Things, artificial intelligence, and smart health.

Huansheng Ning received the BS degree from Anhui University, China, in 1996 and the PhD degree from Beihang University, China, in 2001. He is currently a professor and vice dean at School of Computer and Communication Engineering, University of Science and Technology Beijing, China. He has presided many research projects including Natural Science Foundation of China and National High Technology Research and Development Program of China (863 Project). He has published more than 200 journal/conference papers and authored 5 books. He serves as an associate editor of *IEEE Systems Journal* (2013 to now) and *IEEE Internet of Things Journal* (2014 to 2018), and as a steering committee member of *IEEE Internet of Things Journal* (2016 to Now). His research interests include Internet of Things, general cyberspace and metaverse, smart education, cyber syndrome, and cyber-health.

Sahraoui Dhelim received the BS degree in computer science from University of Djelfa, Algeria, in 2012, the MS degree in networking and distributed systems from University of Laghouat, Algeria, in 2014, and the PhD degree in computer science and technology from University of Science and Technology Beijing, China, in 2020. From 2020 to 2021 he was a visiting researcher at Ulster University, UK. He is currently a senior postdoctoral researcher at Dublin City University, Ireland. He serves as a guest editor in several reputable journals, including *Electronics* and *Applied Science*. His research interests include social computing, smart agriculture, deep-learning, recommendation systems, and intelligent transportation systems.