# IAD: In-Context Learning Ability Decoupler of Large Language Models in Meta-Training

**Yuhan Liu**[1*], **Xiuying Chen**[2*], **Xing Gao**[3], **Ji Zhang**[3], **Rui Yan**[1†]
[1]Gaoling School of Artifical Intelligence, Renmin University of China, Beijing, China
[2]King Abdullah University of Science and Technology, Thuwal, Saudi Arabia
[3]Alibaba DAMO Academy, HangZhou, China
{yuhan.liu, ruiyan}@ruc.edu.cn, xiuying.chen@kaust.edu.sa,
{gaoxing.gx, zj122146}@alibaba-inc.com

**Abstract**

Large Language Models (LLMs) exhibit remarkable In-Context Learning (ICL) ability, where the model learns tasks from prompts consisting of input-output examples. However, the pre-training objectives of LLMs often misalign with ICL objectives. They're mainly pre-trained with methods like masked language modeling and next-sentence prediction. On the other hand, ICL leverages example pairs to guide the model in generating task-aware responses such as text classification and question-answering tasks. The basic pre-training task-related capabilities can sometimes overshadow or conflict with task-specific subtleties required in ICL. To address this, we propose an *In-context learning Ability Decoupler* (IAD). The model aims to separate the ICL ability from the general ability of LLMs in the meta-training phase, where the ICL-related parameters are separately tuned to adapt for ICL tasks. Concretely, we first identify the parameters that are suitable for ICL by transference-driven gradient importance. We then propose a new max-margin loss to emphasize the separation of the general and ICL abilities. The loss is defined as the difference between the output of ICL and the original LLM, aiming to prevent the overconfidence of the LLM. By meta-training these ICL-related parameters with max-margin loss, we enable the model to learn and adapt to new tasks with limited data effectively. Experimental results show that IAD's capability yields state-of-the-art performance on benchmark datasets by utilizing only 30% of the model's parameters. Ablation study and detailed analysis prove the separation of the two abilities.

**Keywords:** In-Context Learning, Large Language Models, Decoupler, Meta-training

## 1. Introduction

In the field of Natural Language Processing (NLP), LLMs exemplified by the notable GPT-3 (Wei et al., 2022a), have garnered substantial attention. These LLMs exhibit a remarkable proficiency in the ICL (Min et al., 2022b), a paradigm in which they acquire task-specific knowledge by processing input-output pairs provided as prompts (Brown et al., 2020). This paradigm has brought about significant advancements in various NLP tasks, ranging from text classification to text generation (Lu et al., 2022a). Recent research have been extensively directed towards augmenting the ICL capabilities of LLMs, often employing techniques such as supervised learning and meta-learning (Min et al., 2022b).

However, a pressing and fundamental concern persists—a substantial misalignment exists between the training objectives of the initial pre-training phase and those of the subsequent meta-training phase for ICL. This misalignment is graphically depicted in Figure 1. The pre-training phase exposes LLMs to vast amounts of textual data, enabling them to comprehensively grasp linguistic patterns and structures, which are mainly pre-trained with methods like masked language modeling and next-sentence prediction. Conversely, the ICL phase requires a targeted understanding of specific tasks, relying on a limited set of task-specific examples such as text classification or question answering. This misalignment brings

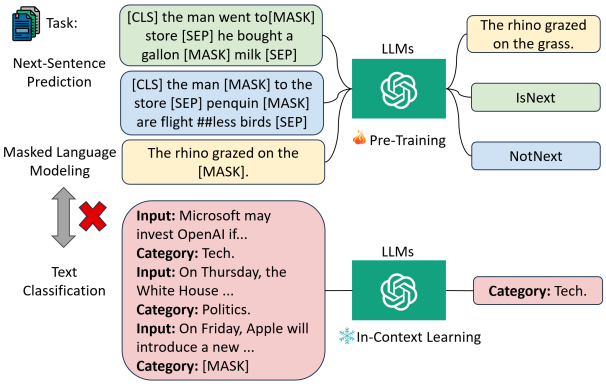



Figure 1: An instance sourced from the pre-training dataset of LLMs (Raffel et al., 2020), along with a clarifying ICL examples illustration focused on topic classification. The pre-training phase of LLMs and ICL objectives are misaligned.

[*] Equal contribution.
[†] Corresponding author: Rui Yan.



*LREC-COLING 2024*, pages 8535–8545
20-25 May, 2024. 

about two prominent challenges: firstly, the disparities in training data distribution and structure can impede LLMs' precise adaptation to novel ICL tasks; secondly, utilizing ICL data for meta-training might risk compromising the model's foundational linguistic comprehension. Such misalignment becomes a major hindrance when attempting to fully harness LLMs, which ideally should balance both general language proficiency and specific ICL capabilities.

To address this challenge, in this paper, we present a novel approach called the "*In-Context Learning Ability Decoupler*" (IAD). Our motivation is to distinguish the parameters optimized for ICL from those tailored for general linguistic knowledge, and then independently fine-tune the ICL-specific parameters. To accomplish this, we introduce a computational mechanism for assessing the significance of parameters associated with the ICL ability. This mechanism enables the seamless transfer of newly acquired ICL skills across a wide range of task datasets. Specifically, we treat the Multi-Head Attention (MHA) and Feed-Forward Network (FFN) components at each model layer as discrete units and evaluate the importance of each unit by analyzing its responsiveness to the loss function during ICL data training. Our approach addresses the challenge of task transfer by quantifying gradient importance and selectively stabilizing specific parameters during meta-training, thereby aiding LLMs in adapting dynamically to various ICL tasks. Moreover, to balance between the model's ICL ability and general ability during meta-training, we introduce a max-margin loss. This loss serves as a crucial bridge, aligning task-specific ICL objectives with the overall language modeling prowess of LLMs, thus preventing LLMs from becoming overconfident in ICL tasks. To mitigate the risk of producing responses with low prediction probabilities, we focus on minimizing this overconfidence indicator within the ICL framework. To empirically validate the effectiveness of IAD, we conduct a comprehensive series of experiments across a wide range of tasks sourced from diverse datasets such as Crossfit (Ye et al., 2021) and Numersense (Lin et al., 2020). The extensive evaluation on various tasks proves IAD's effectiveness and adaptability across diverse tasks.

Our contributions are summarized as follows:

- To the best of our knowledge, we are the first to employ the decoupling of ICL from the general abilities of LLMs as a strategic approach to mitigate the misalignment between pre-training and ICL objectives.
- We introduce an innovative method for calculating transference-driven gradient importance. During the meta-training of LLMs, we focus on training only the most crucial parameters for ICL.
- Our design of a max-margin loss mitigates the overconfidence of the LLM, thereby enhancing the model's ability to generalize.
- Extensive experimental results demonstrate that IAD achieves state-of-the-art performance compared to full-parameter meta-training baselines. Nevertheless, it accomplishes this by training only a subset of the model's parameters, thus significantly economizing computational resources.

## 2. Related Work

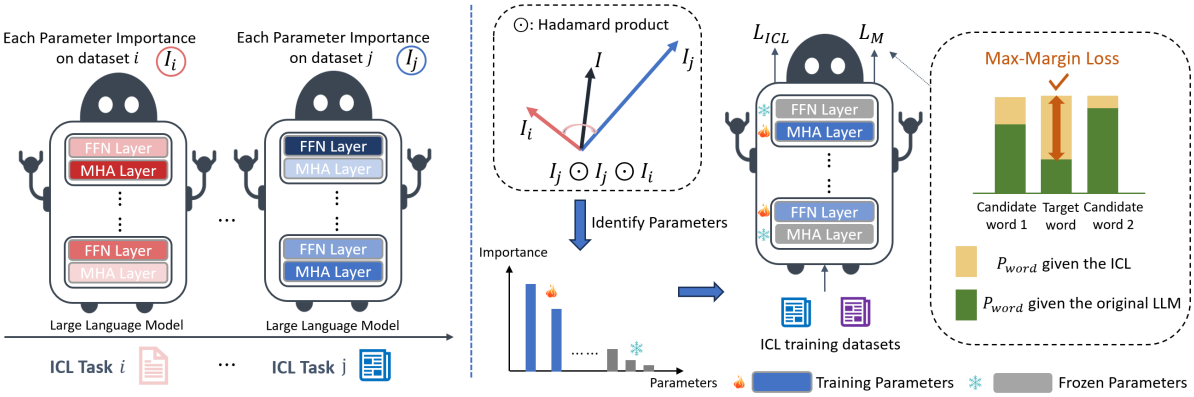
### 2.1. In-Context Learning

Initially proposed by Brown et al. (2020), ICL involves conditioning LLMs on a concatenated prompt of training examples, enabling the model to adapt to new tasks with no parameters update. ICL has undergone subsequent refinements through the works by Zhao et al. (2021) and Holtzman et al. (2021), yielding promising outcomes across diverse tasks. However, it is essential to note that the objectives of the ICL tasks driven by LLMs are misaligned with the training objectives of LLMs. Recent research has been directed toward comprehending its underlying mechanisms to enhance adaptive capabilities (Min et al., 2022b).

Prior research (Min et al., 2022b) has mainly focused on meta-training entire parameter sets within LLMs to enhance their ICL performance. However, it faces challenges due to the misalignment in the training objectives, as depicted in Fig 1. Such methods often result in compromising the model's general ability.

### 2.2. Continual Learning

Continual Learning (CL) has been applied in various tasks such as slot filling (Shen et al., 2019), sentiment assessment (Ke et al., 2021), topic discovery (Gupta et al., 2020), and knowledge enhancement (Lv et al., 2023). Recent studies in the CL domain have centered around LLMs. For instance, Madotto et al. (2021) introduces a system that learns distinct adapters for various domains, although it does not incorporate cross-domain fusion or knowledge transfer techniques. Another approach, DEMIX (Gururangan et al., 2022), initializes new adapters based on the nearest existing adapters. Unlike prior methods, which often involve adapting parameters in a continuous learning context, IAD selects foundational and task-specific parameters by leveraging insights from




(a) Transference-Driven Gradient Importance Computation (b) Meta-Training with Partial Parameters and Max-Margin Loss

Figure 2: Overview of the IAD Model framework consists of two parts: (1) Left: "Transference-driven Gradient Importance Calculation" assesses parameter significance via meta-training gradient analysis and inter-task transfer. (2) Right: "Max-Margin Loss" balances ICL and general model ability, promoting effective decoupling. Leveraging parameter importance, IAD selectively trains crucial ICL parameters while freezing others.

multiple ICL tasks. This unique strategy marks a departure from traditional continual learning approaches and plays a pivotal role in enhancing the model's ability for ICL. Additionally, to address the challenges associated with meta-learning across diverse tasks, we introduce a novel method for calculating gradient importance, streamlining the process of isolating and fine-tuning these critical parameters.

## 3. Methodology

We structured our methodology around three fundamental principles: (1) To identify parameters within the LLM that are most related to ICL, we harness gradients obtained during meta-training on ICL data. This procedure quantifies the significance of parameters by measuring inter-task transfer, which is defined as the reduction in loss for one task resulting from parameter updates driven by gradients from another task. (2) To enhance the LLMS' ICL ability while minimizing interference with its general ability, during training, IAD selectively prioritizes a subset of parameters identified as critical for ICL, while the remaining parameters are kept frozen. This decoupled approach ensures that the model concentrates on task-specific subtleties while preserving its broader general ability. (3) We introduce a max-margin loss that quantifies the difference between the output of ICL and the original LLM, aiming to prevent the overconfidence of the LLM. It encourages the model to maintain task-aware responses without overshadowing or conflicting with task-specific subtleties. The overall framework of IAD is visually depicted in Figure 2.

### 3.1. Transference-Driven Gradient Importance Computation

We regard the Multi-Head Attention (MHA) and Feed-Forward Network (FFN) at each model layer as individual units. Subsequently, we compute the significance of these units when training with ICL data. It has been noted that not all units within a layer are universally considered substantial (Michel et al., 2019). Specifically, within each layer of the transformer architecture, we independently evaluate the significance concerning the MHA and the FFN.

We compute the significance of each unit by evaluating its sensitivity to the loss function during training with ICL data. To elaborate, the importance score is rooted in the model output's reactivity to the parameters encapsulated within each unit, which delineates its ICL attributes. In each model layer, we utilize symbols $\theta_{MHA}$ and $\theta_{FFN}$ to respectively denote the parameters of the MHA and FFN units within the model. Consequently, we can derive the following expressions:

$$I_{MHA} = \mathop{\mathbb{E}}_{x\sim\mathcal{D}_x}\left(\frac{\partial\mathcal{L}_{icl}(x)}{\partial\theta_{MHA}}\right), I_{FFN} = \mathop{\mathbb{E}}_{x\sim\mathcal{D}_x}\left(\frac{\partial\mathcal{L}_{icl}(x)}{\partial\theta_{FFN}}\right). \quad (1)$$

Here, $\mathcal{L}_{icl}$ represents a loss function to the ICL training datasets, and $\mathcal{D}_x$ signifies the ICL data distribution. In practical terms, we calculate the mean value across the training dataset.

The computation of importance scores for all units in the LLMs is performed in a backward pass at the end of an epoch of fine-tuning. The importance score is intricately linked to the model's ICL ability. A lower importance score signifies that the corresponding unit makes only a minor contribu-

tion to the ICL ability. Conversely, a higher importance score suggests substantial significance in enhancing the ICL ability.

### 3.2. Transfer Ability Across Tasks

After training on the ICL training datasets, we determine the significance of units by employing Equation 1 on the datasets. However, within meta-training, the importance of the same unit computed across multiple datasets might vary in magnitude and sign. Directly aggregating them or, as seen by Ke et al. (2022), selecting the maximum result across different datasets as the importance would lead to the detrimental negative transfer of ICL ability during meta-training. Therefore, we introduce the quantification of transfer from dataset $i$ to $j$ (Where conducting meta-training across $T$ datasets, and $i, j \in T$) as the reduction in the loss for dataset $j$ caused by the gradient update from dataset $i$. Specifically, considering that the model parameters $\theta$ are updated by a dataset $i$ with a learning rate $\alpha_i > 0$. Furthermore, we use $g_i(\theta)$ to denote the gradient of the model's loss to its parameters on the dataset $i$. Through a first-order Taylor series expansion, the relationship is given by:

$$\begin{aligned} \Delta L_{i \to j} &= L_j(\theta) - L_j(\theta - \alpha_i g_i(\theta)) \\ &\approx \alpha_i g_i^T(\theta) g_j(\theta). \end{aligned} \tag{2}$$

It is worth noting that the expression $\alpha_i g_i(\theta) \approx \alpha_i g_i^T(\theta) g_j(\theta)$ indicates that a higher inner product value corresponds to more effective transference.

We optimize the gradient by maximizing the inter-datasets transfer $\Delta L_{i \to j}$ (as defined in Equation 2):

$$\begin{aligned} \frac{\partial \Delta L_{i \to j}}{\partial(\theta)} &= \frac{\partial}{\partial \theta}(\alpha_i g_i^T(\theta) g_j \theta) \\ &= \alpha_i H_j(\theta) g_i(\theta) \\ &\approx \alpha_i g_j(\theta) \odot g_j(\theta) \odot g_i(\theta), \end{aligned} \tag{3}$$

where $H_j(\theta)$ is the Hessian matrix of $L_j(\theta)$, and $\odot$ is Hadamard product (i.e., element-wise product)

Therefore, in conjunction with Equation 1, we can deduce that during the process of meta-training for ICL. Formally, let the importance score of unit $i$ be denoted as $I_i$, the importance score of the unit be denoted as $I_u, (u \in \{MHA, FFN\})$ and the ICL datasets are $D_{context}$ (denoted as $T$ in number), then the importance score is computed as:

$$I_u = \sum_{j \neq i, j \in T} \alpha_i I_j \odot I_j \odot I_i, \tag{4}$$

where $I_u$ represents the overall importance of a unit (MHA or FFN) across multiple ICL training datasets, during computation, we substitute the respective values for MHA and FFN, calculating their individual importance accordingly.

### 3.3. Meta-Training with Selected Parameters

Inspired by continual learning, this section will selectively freeze and meta-training parameters according to their importance scores. We employ tailored importance scores to evaluate the significance of each unit's contribution to the ICL task. This analysis quantitatively measures the unit's impact on the model's performance in ICL tasks. Importantly, we focus on the MHA and FFN components, known for their crucial roles in ICL tasks.

Subsequently, leveraging the importance scores, we implement a unit selection strategy. Specifically, we identify units with lower importance scores and freeze the associated parameters, ensuring they remain constant during subsequent training stages. This step effectively reduces the parameter space that requires updates during meta-training. By isolating the less critical parameters, we significantly diminish the computational resources required for training, leading to a more accurate and efficient process:

$$\theta_i = \begin{cases} \theta_i & \text{if } I_u > \text{top } n\text{-th percentile} \\ \text{frozen} & \text{otherwise} \end{cases} \tag{5}$$

### 3.4. Max-Margin Loss between the ICL and General Ability

When the ICL ability of the model is insufficient, the decoder tends to overlook certain input information about ICL, assuming a more prominent role as an open-ended LLM. Consequently, there is a rising risk of inference errors in the ICL task. Inspired by faithfulness-enhanced abstractive summarization task (Chen et al., 2022), we introduce a max-margin loss into the ICL task and employ it for the model's decoupled training process to enhance its ICL ability. This loss aims to maximize the discrepancy between the model's ICL ability and the LLM's predictive ability, effectively mitigating the tendency of LLMs to generate frequently seen collocations that do not align with the intended ICL inference.

To provide a more comprehensive explanation, we establish the margin between the ICL ability of the model and its general ability. This margin is characterized as the disparity in predictive probabilities:

$$M_t = P_t^{ICL}(y_t | y_{<t}, X) - P_t^{LM}(y_t | y_{<t}, X). \tag{6}$$

Here, $X$ denotes the input within the examples of ICL tasks. Furthermore, $P_t^{ICL}$ represents the predictive probability of the model for the t-th token during ICL, while $P_t^{LM}$ signifies the predictive probability of the original LLMs for the t-th token, reflecting the general ability of the language model. Intu-

itively, when $M_t$ is substantial, the model's ICL ability is evidently satisfactory. Conversely, when $M_t$ is relatively small, the possibility is that the model's general ability is inadequate yet overconfident, resulting in a diminished ICL ability.

Consequently, we introduce the max-margin loss, denoted as $L_M$, wherein a coefficient is incorporated into the margin:

$$L_M = \sum_t (1 - P_t^{ICL})(1 - M_t^5)/2. \quad (7)$$

Let $P_t^{ICL}(y_t|y_{<t}, X)$ be abbreviated as $P_t^{ICL}$. The component $(1-D_t^5)/2$ represents a non-linear, monotonically decreasing function concerning $D_t$, thus ensuring the optimization goal of maximizing $D_t$. For this purpose, we adopt the Quintic function (raised to the fifth power) as it has demonstrated greater stability (Miao et al., 2021). The initial factor $(1 - P_t^{ICL})$ serves to accommodate the two scenarios we previously discussed. A substantial value of $P_t^{ICL}$ indicates a proficient ICL ability of the model. This interpretation is encapsulated by $(1-P_t^{ICL})$, yielding a minor influence on $M_t$. Conversely, when $P_t^{ICL}$ is small, it signifies the necessity for the model's ICL ability to undergo refinement. This prompts the application of a significant coefficient $(1-P_t^{ICL})$, allowing the model to learn from the margin information effectively.

As per the reference (Min et al., 2022b), the loss employed during the meta-training phase of the model is a negative log-likelihood objective $L_{ICL}$. The losses, $L_M$ and $L_{ICL}$, are orthogonal and amenable to combination to enhance the model's ICL ability. The total loss is $L = L_M + L_{ICL}$.

# 4. Experiment

## 4.1. Experiment Setup

### 4.1.1. Datasets

We leveraged a comprehensive array of tasks curated from two prominent datasets: the CROSSFIT dataset by Ye et al. (2021) and the UNIFIEDQA dataset by Khashabi et al. (2020). The datasets comprise a wide spectrum of distinct tasks, spanning diverse problem domains such as text classification (Li et al., 2024) and question-answering (Chen et al., 2021).

In our experiments, we systematically explored various discrete configurations. In each experimental set, we notably employed three datasets for meta-training and two for testing, with strict segregation between the training and test datasets. Specifically, in our experiments, we utilized the SR (Dagan et al., 2005), TES (Barbieri et al., 2020), and TESF (Barbieri et al., 2020), ENO (Mollas et al., 2020) datasets as two distinct sets of testing data to evaluate the model's ICL ability. The data configuration is also presented in Table 1 for easy reference. Within each configuration, we carefully selected a subset of target tasks, ensuring that they do not share any domain congruence with the meta-training tasks. This distinction is exemplified across diverse domains, including finance, poetry, climate studies, and medical research. Our reporting encompasses results derived from all target tasks and results exclusively from target tasks without any domain overlap. We refer readers to the supplementary materials for a more detailed description of the training and test datasets.

| | |
|---|---|
| Input | Nevertheless over the last decade, daily record high temperatures occurred twice as often as record lows. options: {"Disputed", "Not enough info", "Refutes", "Supports"} |
| output | Refutes |

Table 1: Example input-output pairs for an ICL task

### 4.1.2. Baselines

We compare IAD with a range of baselines:

**0-shot**: We use a pre-trained LLM and run zero-shot inference, following Brown et al. (2020).

**In-context learning**: We use the pre-trained LLM and use ICL by conditioning on a concatenation of $k$ training examples, following Brown et al. (2020).

**Channel 0-shot, Channel In-context**: We use the noisy channel model by Min et al. (2022a) for 0-shot and ICL.

**Multi-task 0-shot**: We train the LLM on the same meta-training tasks without utilizing in-context learning objectives, essentially maximizing $P(y|x)$ without additional training examples (k = 0). This approach aligns with typical multi-task learning techniques found in prior work (Wei et al., 2021).

**MetaICL**: We employ a meta-learning approach to fine-tune the LLMs on the ICL dataset, following Min et al. (2022b).

### 4.1.3. Evaluation

For the test ICL tasks, we utilize Macro-F1 as an evaluation metric. Similar to the meta-training process, we adopt $k = 16$ training instances in a specific test task chosen through uniform random sampling. We relax the presumption of perfect label balance across the $k$ training instances, following the methodology of Min et al. (2022a). Acknowledging the inherent variance associated with ICL (Zhao et al., 2021; Perez et al., 2021; Lu et al., 2022b; Zhang et al., 2024), we engage distinct

| | Online Speech Detection | | | Tweet Classification | | | |
|---|---|---|---|---|---|---|---|
| Method | SR | ENO | F1 Score | TESF | TES | F1 Score | Average |
| 0-shot | 34.22 | 40.25 | 37.24 | 35.85 | 38.40 | 37.13 | 37.18 |
| In-context learning | 36.90 | 39.10 | 38.00 | 42.41 | 41.62 | 42.02 | 40.01 |
| Channel 0-shot | 37.03 | 38.10 | 37.57 | 41.33 | 40.23 | 40.78 | 39.18 |
| Multi-task 0-shot | 36.67 | 38.76 | 37.72 | 40.87 | 39.75 | 40.31 | 39.02 |
| Channel In-context | 43.69 | 40.27 | 41.98 | 44.03 | 45.84 | 44.94 | 43.46 |
| MetaICL | 43.55 | 41.37 | 42.46 | 45.28 | 45.27 | 45.28 | 43.87 |
| IAD | **46.74** | **42.66** | **44.70** | **47.27** | **46.80** | **47.04** | **45.87** |

Table 2: The experimental results were obtained by conducting tests on two distinct sets of datasets. Online Speech Detection datasets include SR (Dagan et al., 2005) and ENO (Mollas et al., 2020), while Tweet Classification datasets consist of TESF and TES (Barbieri et al., 2020). **Bold** indicates the model outperforms all baselines significantly in paired t-test at $p < 0.01$ level.

sets of k-training instances. Initially, we calculate performance across multiple random seeds for each test task. Subsequently, we present the Macro-F1 of these metrics across all test tasks, which we denote as "F1-SCORE" in the experiment.

#### 4.1.4. Experiment Details

The entire implementation is conducted within the PyTorch framework (Paszke et al., 2019) using the Transformers library (Wolf et al., 2020). During the process of meta-training, We used GPT2-Large as LLM for model training, and we considered a maximum of 16,384 training instances for each individual task. During our experiments, our model utilized the decoupling mechanism to freeze 70% of the parameters while training only 30% of the model's parameters. The training is performed using a batch size of 1, a learning rate of 1e-5, and a sequence length of 1024. The model undergoes training for a total of 3 epochs.

### 4.2. Experimental Results

#### 4.2.1. Main Results

We conducted comparative experiments between our model, IAD, and several classical ICL baselines. For Online Speech Detection datasets, we performed meta-training on three datasets and evaluated our model on SR and ENO datasets. In the case of Tweet Classification datasets, we utilized another three datasets for meta-training and evaluated on TESF and TES datasets. The evaluation results are presented in Table 2. Our model outperforms the baseline models across all metrics. This consistent improvement across both sets of datasets demonstrates our proposed IAD framework's superiority and general applicability.

Our approach not only surpassed all baseline models but also significantly reduced the computational resources required. Note that we only train 30% of the model's parameters, which proves the efficiency of our model. This finding highlights that, during the meta-training process, the decoupling mechanism should be considered, focusing training efforts on parameters more critical for ICL to achieve better results. We employed a two-tailed paired t-test with $\alpha = 0.01$ to assess the statistical significance of performance differences between two separate runs.

Overall, our results underscore the effectiveness and efficiency of our IAD model in improving ICL performance across various datasets and reinforce the importance of parameter decoupling during meta-training for enhanced model adaptability.

#### 4.2.2. Ablations Study

We have also presented the results of an ablation study in Table 3, to investigate the impacts of different modules within our proposed model. It can be observed that if we neither freeze a portion of parameters nor utilize the Max-Margin Loss, the performance of all metrics deteriorates, reducing the model to a state similar to the MetaICL model. This underscores the significance of decoupling abilities during training to enhance the model's ICL ability.

Furthermore, upon removing each of the two modules separately, we observed a decrease in scores by 0.9% and 2.1%, respectively. This indicates that the two modules we introduced contribute to the model's improved ICL ability. These findings emphasize the cooperative effect of the proposed modules, reinforcing the model's capacity in tasks requiring adaptation to context.

| | Online Speech Detection | | | Tweet Classification | | | |
|---|---|---|---|---|---|---|---|
| | SR | ENO | F1 Score | TESF | TES | F1 Score | Average |
| Full model - IAD | **46.74** | **42.66** | **44.70** | **47.57** | **46.79** | **47.18** | **45.95** |
| -w/o Max-Margin loss | 45.40 | 42.04 | 43.72 | 46.66 | 46.11 | 46.39 | 45.06 |
| -w/o Frozen part parameters | 43.55 | 41.37 | 42.46 | 45.28 | 45.07 | 45.18 | 43.82 |

Table 3: Ablations experiment results on two sets of test datasets. Online Speech Detection datasets include SR (Dagan et al., 2005) and ENO (Mollas et al., 2020), while Tweet Classification datasets consist of TESF and TES (Barbieri et al., 2020). The best score is in **bold**.

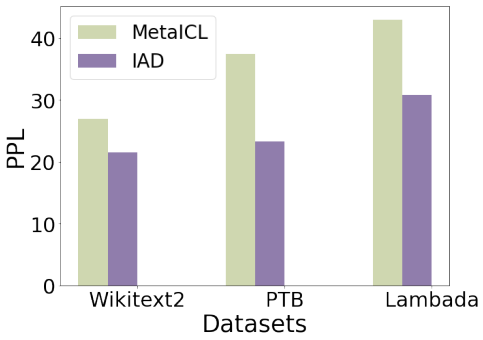



Figure 3: Experimental results on the Language Modeling task on three test datasets. Our model IAD has a significantly lower PPL than the strong baseline model MetaICL, indicating that our model possesses better language modeling capabilities.

## 4.3. Discussions

### 4.3.1. Decoupling of General and ICL Ability of LLMs

We evaluate the model's overall performance on language modeling tasks as a measure of its general ability. To assess this, we employ Perplexity (PPL) scores, which serve as an indicator of the language modeling proficiency of the models. Lower PPL scores correspond to stronger language modeling abilities.

As depicted in Figure 3, in contrast to the MetaICL, our IAD model consistently achieved significantly lower PPL scores across all three benchmark datasets: Wikitext2 (Gong et al., 2018), PTB (Gong et al., 2018), and Lambada (Paperno et al., 2016). This exceptional performance of the IAD model surpassing MetaICL underscores the effectiveness of the proposed decoupling methodology.

Our observations reveal that the IAD model, designed to disentangle ICL ability from the general LLM ability, demonstrates a discernible trade-off between these two attributes. It exhibited the capability to uphold competitive language modeling skills while enhancing its ICL ability. This observation supports the assertion that IAD effectively accomplishes its intended goal of isolating these two abilities. In summary, our experiments validate the efficacy of IAD as an approach to enhance the ICL ability of LLMs while preserving their language modeling prowess.

### 4.3.2. Influence of Frozen Parameters Percentage

We conducted experiments to investigate the relationship between the proportion of frozen parameters and the model's ICL ability. We froze different percentages of model parameters for experiment, and the results are presented in Figure 4. At a parameter freeze ratio of 70%, the Macro-F1 Score reached its highest value at 46.8. This indicates that, in this experiment, freezing a portion of parameters up to 70% can enhance the model's Macro-F1 Score. Beyond a parameter freeze ratio of 70%, the Macro-F1 Score starts to decline, with a significant drop observed at a parameter freeze ratio of 100%. This underscores once again that excessive parameter freezing can detrimentally affect the model's ICL ability.

The results also suggest that applying meta-training on all model parameters solely based on ICL data is not an optimal approach. Instead, a decoupling of the model's parameters is more suitable, focusing meta-training efforts on only a subset of parameters. This approach holds the potential for significantly improving the model's ICL ability while concurrently reducing the required computational resources. Furthermore, the findings emphasize the necessity to strike a balance between parameter freezing and adaptability in order to optimize ICL performance.

### 4.3.3. Visualization of Importance Across Different Layers

We present the visualized results of the method proposed by our model for calculating unit importance based on transference. This visualization aids in further analyzing the significance of each layer's structure in the model concerning ICL ability. This contributes to a deeper understanding of the decoupling of ICL ability and general abil-

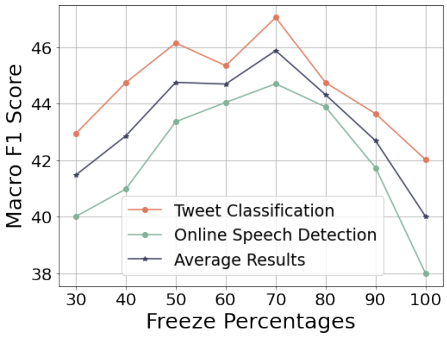



Figure 4: Performance of models with frozen different percentage parameters on ICL task.

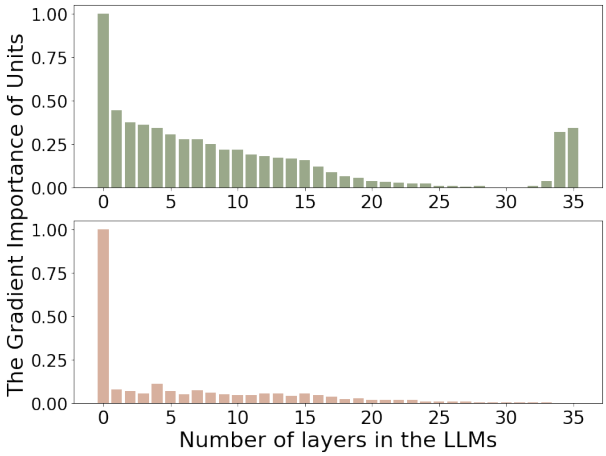



Figure 5: The gradient importance of parameters $I_u, (u \in \{MHA, FFN\})$ in the **FFN** units and **MHA** units of each layer in the model.

ity in our model, as illustrated in Figures 5. For both the FFN and the MHA layer, our experimental results demonstrate a consistent trend where the importance decreases as the number of layers increases. This implies that the higher layers of the model contribute less to ICL ability while the lower layers contribute more significantly.

However, in the analysis of the FFN, the middle layers (layers 1-10) and the top layers (layers 34-35) exhibit relatively higher importance. In contrast, in the analysis of the MHA layers, the importance of these layers gradually diminishes. This observation suggests that, in the context of ICL tasks, the middle and deep FFN layers might play a more critical role than the MHA layers.

Furthermore, as depicted in Figure 5, the contribution of the FFN is relatively distributed across both lower and middle layers. In contrast, Figure 5 shows that the contribution of the MHA layers is more prominent in the lower layers. This may imply that the self-attention mechanism is advantageous in capturing lower-level contextual features for ICL tasks. At the same time, the FFN layers might be more involved in localized feature processing at the lower and middle layers.

#### 4.3.4. Generalization of IAD Across Different Types of Tasks

As shown in Figure 6, our model has demonstrated noteworthy superiority in our experimental investigation when applied to a novel QA task. As in the previous experimental setup, we conducted meta-training on three datasets and then performed experiments on three non-overlapping datasets. For specific experimental details, please refer to the supplementary materials. Specifically, the IAD model achieved a mean score of 44.6, significantly surpassing the performance of the three baseline methods. This observation underscores the ability of the IAD model to excel across diverse tasks, achieving optimality consistently.

The success of the IAD model can be attributed to its decoupled methodology, which involves fine-tuning a critical subset of model parameters on ICL. This targeted refinement enhances the model's proficiency in ICL, conferring substantial advantages in task adaptation and generalization. Such flexibility and adaptability help the IAD model with the potential to manifest remarkable performance across distinct domains and problem domains. Consequently, the IAD model has achieved the highest mean score in our experimentation.

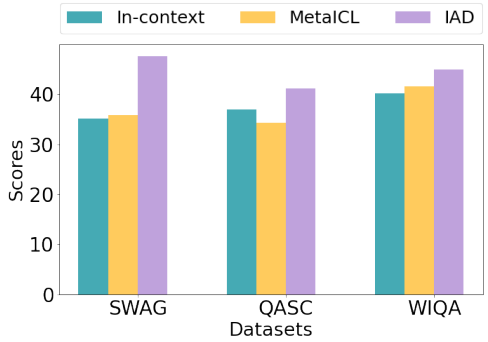



Figure 6: Experimental results on the QA task across three test datasets. The experiments have substantiated that our model IAD exhibits superior generalization capabilities across diverse task.

## 5. Conclusion and Broader Impacts

In this paper, we propose an innovative model called the"*In-Context Learning Ability Decoupler*" (IAD) to address the misalignment between the pre-training and ICL objectives of LLMs. The paper proposes a mechanism for identifying and fine-tuning ICL-specific parameters, as well as a max-margin loss to prevent overconfidence in ICL tasks.

The results of experiments demonstrate that IAD significantly improves LLMs' performance on a variety of tasks while using fewer parameters. This work offers a promising solution to the challenge of harmonizing general language modeling abilities with specific ICL requirements in LLMs.

We highlight prospective directions for ICL by delving into advanced fine-tuning techniques that complement the IAD approach, which can explore innovative loss functions and regularization methods to enhance the fine-tuning methods for ICL.

## 6. Ethical Considerations

**Data Privacy and Bias**: All datasets used in this research are published in previous studies and are publicly available. All of the datasets are widely used in the NLP domain. We also manually checked for offensive content in the data. There is no data bias against certain demographics with respect to these datasets.

## 7. Acknowledgement

This work was supported by the National Natural Science Foundation of China (NSFC Grant No. 62122089), Beijing Outstanding Young Scientist Program NO. BJJWZYJH012019100020098, and Intelligent Social Governance Platform, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, Renmin University of China, the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China.

## 8. References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context examples selection for machine translation. *arXiv preprint arXiv:2212.02437*.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Proc. of EMNLP Findings*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Proc. of NeurIPS*.

Xiuying Chen, Zhi Cui, Jiayi Zhang, Chen Wei, Jianwei Cui, Bin Wang, Dongyan Zhao, and Rui Yan. 2021. Reasoning in dialog: Improving response generation by context reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12683–12691.

Xiuying Chen, Mingzhe Li, Xin Gao, and Xiangliang Zhang. 2022. Towards improving faithfulness in abstractive summarization. *Proc. of NeurIPS*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2023. Why can GPT learn in-context? language models secretly perform gradient descent as meta-optimizers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4005–4019, Toronto, Canada. Association for Computational Linguistics.

Chengyue Gong, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2018. Frage: Frequency-agnostic word representation. *Proc. of NeurIPS*.

Pankaj Gupta, Yatin Chaudhary, Thomas Runkler, and Hinrich Schuetze. 2020. Neural topic modeling with continual lifelong learning. In *Proc. of ICML*.

Suchin Gururangan, Mike Lewis, Ari Holtzman, Noah A Smith, and Luke Zettlemoyer. 2022. Demix layers: Disentangling domains for modular language modeling. In *Proc. of NAACL*.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proc. of EMNLP*.

Zixuan Ke, Bing Liu, Nianzu Ma, Hu Xu, and Lei Shu. 2021. Achieving forgetting prevention and knowledge transfer in continual learning. *Proc. of NeurIPS*.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. 2022. Continual pre-training of language models. In *Proc. of ICLR*.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single qa system. In *Proc. of EMNLP Findings*.

Mingzhe Li, Xiuying Chen, Jing Xiang, Qishen Zhang, Changsheng Ma, Chenchen Dai, Jinxiong Chang, Zhongyi Liu, and Guannan Zhang. 2024. Multi-intent attribute-aware text matching in searching. *WSDM*.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! numersense: Probing numerical commonsense knowledge of pre-trained language models. In *Proc. of EMNLP*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022a. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proc. of ACL*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022b. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proc. of ACL*.

Ang Lv, Kaiyi Zhang, Shufang Xie, Quan Tu, Yuhan Chen, Ji-Rong Wen, and Rui Yan. 2023. Are we falling in a middle-intelligence trap? an analysis and mitigation of the reversal curse.

Andrea Madotto, Zhaojiang Lin, Zhenpeng Zhou, Seungwhan Moon, Paul A Crook, Bing Liu, Zhou Yu, Eunjoon Cho, Pascale Fung, and Zhiguang Wang. 2021. Continual learning in task-oriented dialogue systems. In *Proc. of EMNLP*.

Mengqi Miao, Fandong Meng, Yijin Liu, Xiao-Hua Zhou, and Jie Zhou. 2021. Prevent the language model from being overconfident in neural machine translation. In *Proc. of ACL*.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? *Proc. of NeurIPS*.

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022a. Noisy channel language model prompting for few-shot text classification. In *Proc. of ACL*.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022b. Metaicl: Learning to learn in context. In *Proc. of NAACL*.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2020. Ethos: an online hate speech detection dataset. *arXiv preprint arXiv:2006.08328*.

Denis Paperno, German David Kruszewski Martel, Angeliki Lazaridou, Ngoc Pham Quan, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda Torrent, Fernández Raquel, et al. 2016. The lambada dataset: Word prediction requiring a broad discourse context. In *Proc. of ACL*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Proc. of NeurIPS*.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. *Proc. of NeurIPS*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*.

Yilin Shen, Xiangyu Zeng, and Hongxia Jin. 2019. A progressive model to enable continual learning for semantic slot filling. In *Proc. of EMNLP*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *Proc. of ICLR*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Proc. of NeurIPS*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proc. of EMNLP*.

Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. Crossfit: A few-shot learning challenge for cross-task generalization in nlp. In *Proc. of EMNLP*.

Kaiyi Zhang, Ang Lv, Yuhan Chen, Hansen Ha, Tao Xu, and Rui Yan. 2024. Batch-icl: Effective, efficient, and order-agnostic in-context learning.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proc. of ICML*.