# *Both Matter*: Enhancing the Emotional Intelligence of Large Language Models without Compromising the General Intelligence

**Anonymous ACL submission**

## Abstract

Emotional Intelligence (EI), consisting of emotion perception, emotion cognition and emotion expression, plays the critical roles in improving user interaction experience for the current large language model (LLM) based conversational general AI assistants. Previous works mainly focus on raising the emotion perception ability of them via naive fine-tuning on EI-related classification or regression tasks. However, this leads to the incomplete enhancement of EI and catastrophic forgetting of the general intelligence (GI). To this end, we first introduce EiBench, a large-scale collection of EI-related tasks in the text-to-text format with task instructions that covers all three aspects of EI, which lays a solid foundation for the comprehensive EI enhancement of LLMs. Then a novel **Mo**dular **E**motional **I**ntelligence enhancement method (**MoEI**), consisting of Modular Parameter Expansion and intra-inter modulation, is proposed to comprehensively enhance the EI of LLMs without compromise their GI. Extensive experiments on two representative LLM-based assistants, Flan-T5 and LLaMA-2-Chat, demonstrate the effectiveness of MoEI to improving EI while maintain GI. [1]

## 1 Introduction

*The question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without any emotions.*

*– Marvin Minsky*

Emotional intelligence (EI), a pivotal concept in the field of human intelligence, holds significant importance in the context of the current large language models (LLMs) (Brown et al., 2020; Raffel et al., 2020; Touvron et al., 2023) exhibiting great general intelligence (GI) to serve as conversational general AI assistants (Minsky, 2007). It involves

[1]Our data and codes could be found in supplementary files.

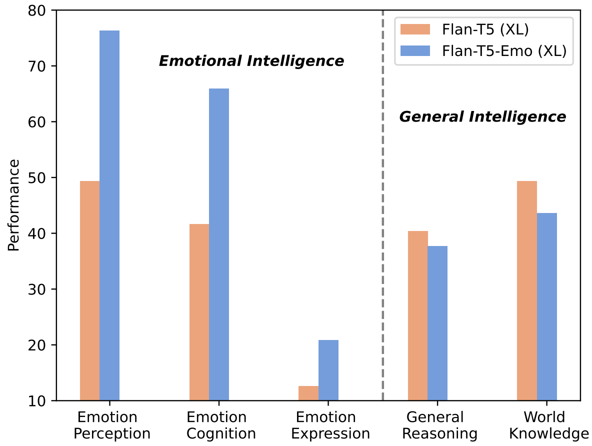


Figure 1: Comparison of EI and GI before (in orange) and after (in blue) the EI-enhancement via naive fine-tuning based on the Flan-T5-XL (3B) backbone.

effectively deal with EI-related downstream tasks to accurately perceive, understand users' emotional states and respond properly, which is necessary for fostering effective communication and facilitating smooth social interactions (Mayer et al., 2001).

To enhance the EI, many researchers have started performing the naive fine-tuning of LLMs on EI-related tasks (Zhang et al., 2023a,b; Lei et al., 2023; Liu et al., 2024). However, there are two significant limitations in the current state of these efforts.

**On one hand**, the EI-enhancement in these studies is of narrow scope because they only focus on emotion perception regarding emotion classification tasks. However, EI is a broad concept that also includes emotion cognition (e.g. emotion cause reasoning (Poria et al., 2021)) and expression (e.g. empathethic response generation (Rashkin et al., 2018)). Therefore, the comprehensive enhancement of EI for LLMs is ignored in previous works.

**On the other hand**, they all overlook the catastrophic forgetting (McCloskey and Cohen, 1989) of the GI of LLM backbones during the process of enhancing EI via naive fine-tuning. In our preliminary experiments depicted in Figure 1, we find that solely focusing on improving the EI of an LLM backbone would result in a significant decline in its GI, such as world knowledge and general reasoning. Desirable intelligent assistants are supposed to be of both high GI and high EI, where any performance loss on either front would significantly undermine the user experience.

To this end, we study *how to comprehensively enhance the EI of current LLM-based assistants while preserving their inherent GI from being compromised*. According to three aspects of EI defined by Mayer et al. (2001), named emotion perception, emotion cognition and emotion expression, we first construct EIBENCH, a large curated collection of EI-related tasks converted into a text-to-text format, covering 15 tasks with 88 datasets. Moreover, motivated by the promising gains from fine-tuning with task instructions (Sanh et al., 2021; Wang et al., 2022; Longpre et al., 2023), we also manually write related instructions for each dataset. Thus, EIBENCH lays a solid foundation for the comprehensive EI enhancement of LLMs.

Further, we propose a novel **Mo**dular **E**motional **I**ntelligence enhancing method (**MoEI**) with two collaborative techniques, Modular Parameter Expansion (MPE) and Intra-Inter Modulation ($I^2M$), to comprehensively improve the EI of LLMs while maintaining most of their GI. Specifically, in MPE, a set of modular parameters are introduced to endow additional capacity to handle various tasks within the above three aspects of EI. For the sake of computation and resource efficiency, these expanded modular parameters are instantiated with parameter-efficient LoRA blocks (Hu et al., 2021), named MoLoRA. During the process of EI-enhancement, only the EI-specific MoLoRA is updated, thereby reducing the impact on parameters of the LLM backbone representing its GI. In addition, a router is devised in $I^2M$ to exert the modulation on the two separate parameters. More specifically, intra-modulation is performed within MoLoRA, leveraging the weighted combination of different LoRA blocks to deal with various EI-related tasks. And inter-modulation functions to strike the balance between MoLoRA and the whole LLM backbone to achieve the goal of protecting GI, where the GI-related samples are navigated to be processed by the LLM backbone while reducing the influence from EI-specific MoLoRA.

We conduct extensive experiments on two representative open-source LLM-based assistants, Flan-T5 (Chung et al., 2022) and LLaMA-2-Chat (Touvron et al., 2023). Results demonstrate that MoEI not only helps significantly enhance all three aspects of their EI but also ensures that their GI, including world knowledge, general reasoning, commonsense reasoning and reading comprehension, are hardly compromised. Moreover, such EI-enhanced models could exhibit better performance when addressing various OOD tasks in the presence of emotional stimuli (Li et al., 2023b).

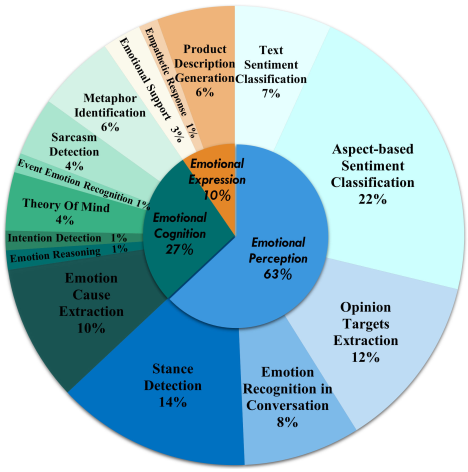


Figure 2: Overview of EIBENCH. 15 EI-related tasks are categorized into 3 main categories: emotion perception, emotion cognition, and emotion expression.

The main contributions of this work are summarized as follows: (1) We take the first step to study how to develop an LLM-based assistant that possesses both high GI and EI, a challenging direction for the more practical application of LLMs. (2) We introduce EIBENCH, a comprehensive collection of EI-related tasks to support the EI enhancement of LLM backbones. Then a novel method MoEI is proposed to comprehensively improve the EI of LLM backbones without compromising their GI. (3) Experiments on various EI and GI benchmarks demonstrate the effectiveness of MoEI.

## 2 EIBENCH

We first introduce EIBENCH, a large-scale collection of EI-related tasks with instructions that describe them in plain language. Figure 2 shows the overview of the benchmark by category and task.

**Taxonomy of Emotional Intelligence.** According to Mayer et al. (2001), the EI of LLMs includes three aspects: (1) Emotion Perception, (2) Emotion Cognition and (3) Emotion Expression. Please refer to Appendix A for their specific meanings.

**Task Collection.** Based on the aforementioned EI taxonomy, we systematically gather and organize existing open-source tasks and datasets related to EI within this framework. The resulting depiction of all EI-related tasks in EIBENCH is illustrated in Figure 2, while Table 4 (in Appendix B) provides a comprehensive list of datasets associated with each task. This leads to 15 tasks and 88 datasets in total. To be more specific, in the aspect of emotion perception, the primary focus is on classification tasks such as emotion recognition and stance detection. Shifting to the area of emotion cognition, the prioritized task involves extracting emotion causes. Additionally, we incorporate more advanced cognitive challenges, including irony and metaphor recognition, aiming to comprehensively enhance the LLMs' emotional cognitive capabilities. Finally, in the realm of emotion expression, the main tasks involve empathetic response and emotional support, empowering the model to offer improved comfort and guidance to users, thereby enhancing the overall interactive experience.

**Task Schema.** Motivated by the promising gains from fine-tuning with task instructions (Sanh et al., 2021; Wang et al., 2022; Longpre et al., 2023), we manually construct one piece of instruction for each dataset in our EIBENCH. Specifically, the schema contains the following components: (1) **Text Input:** the input sentence $X$. (2) **Instruction:** the detailed guidance on how the model should process $X$ to complete the current task. (3) **Option** (for classification task only): including all the candidate labels and serving as both a constraint and a hint. We hire 5 annotators who are proficient in English to write these instructions. In addition, to guarantee the quality of these instructions, 1 or 2 reviewers are also assigned to each dataset. Their task is to confirm that whether the instructions are clear, fluent and comprehensive enough for an average language speaker to successfully complete the given task. Examples of instances from our EIBENCH is displayed in Table 5 in Appendix B.

## 3 Methodology

Instead of merely focusing on enhancing specific capabilities of LLMs, we present a novel approach MoEI, a model-agnostic EI-enhancing method that is compatible with any transformer-based LLM, which could not only comprehensively boost the EI of LLMs but also safeguards their GI from being compromised. As shown in Figure 3, MoEI consists of two collaborative techniques, namely Modular Parameter Expansion and Intra-Inter Modulation. The subsequent section will offer a detailed introduction to both of them.

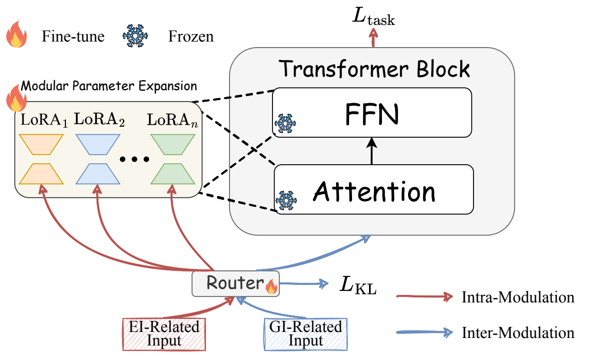


Figure 3: The overall architecture of our proposed MoEI framework, which consists of two techniques, modular parameter expansion and intra-inter modulation. Red and blue lines represent the forward flow of the EI- and GI related inputs that participate in the Intra- and Inter-Modulation, respectively.

### 3.1 Preliminary

**Low Rank Adaptation.** We adopt a representative PET method LoRA (Hu et al., 2021) in MoEI. Specifically, in LoRA, the pre-trained weight matrix of LLMs is expanded with a low-rank decomposition. For any linear layer $h = W_0 x$, the forward pass with LoRA is modified to be:

$$h = W_0 x + BAx \tag{1}$$

where $W_0 \in \mathbb{R}^{d\times k}$, $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$ and the rank $r \ll \min(d, k)$. The pre-trained weight $W_0$ remains fixed during training, while $A$ and $B$ are trainable parameters.

### 3.2 Modular Emotional Intelligence Enhancement

**Modular Parameter Expansion.** To harness two advantages of LoRA, which includes isolation from the entire parameters of LLMs and promising training efficiency, we take a step further by extending it to modular designs, endowing more capacity to accomodate various newly acquired EI-related knowledge. This leads to the Mixture-of-LoRA (MoLoRA) architecture.

MoLoRA inherits the flexibility of LoRA and could be applied on any linear layer of the transformer-based LLM backbone (mainly in the attention and FFN layer) to expand the model capacity with multiple pairs of low-rank matrixes.

To be more specific, the singular pair of low-rank matrixes $BA$ in Equation (1) is now replaced by a set containing multiple ones $\{B_iA_i\}_{i=1}^{N}$ and the forward process of the MoLoRA layer can be mathematically expressed as follows:

$$h = W_0x + \sum_{i=1}^{N} B_iA_ix \tag{2}$$

where $N$ is the number of modular LoRA blocks.

**Intra-Inter Modulation.** To effectively navigate different types of inputs to be properly processed by its corresponding parameters, a router is introduced in I$^2$M to exert the intra- and inter-modulation on weighted contributions of the separate MoLoRA and the LLM backbone:

$$\begin{gathered} h = \alpha W_0x + \sum_{i=1}^{N} \beta_i \; B_iA_ix \\ [\alpha; \{\beta_i\}_{i=1}^{N}] = G(x) = \text{softmax}(Wx) \end{gathered} \tag{3}$$

where $G(\cdot)$ represents the router and $W \in \mathbb{R}^{d \times N+1}$ is the trainable parameter of it. $\alpha$ and $\{\beta_i\}_{i=1}^{N}$ are the intensity weights of inter and intra-modulation, respectively. In the upcoming contents, we will elaborate the concept and delve into the optimization process associated with them.

More specifically, intra-modulation $\{\beta_i\}_{i=1}^{N}$ is only performed within EI-specific MoLoRA to leverage the weighted combination of different LoRA blocks, which leads to more powerful capability in dealing with various EI-related tasks compared to the single LoRA. As shown in Figure 3, the optimization of intra-modulation is driven by the EI-related inputs with the task loss:

$$L_{\text{task}} = -\sum_{(x,y)\in\mathcal{EI}} \log P\left(y \mid x; \theta_m, \theta_E, \theta_G\right) \tag{4}$$

where $\theta_m, \theta_E$ and $\theta_G$ are parameters of the LLM backbone, the MoLoRA and the router, respectively. $\mathcal{EI}$ is the training data samples from EIBENCH. Only parameters of $\theta_E$ and $\theta_G$ are updated during the training.

And inter-modulation $\alpha$ functions to strike the balance between MoLoRA and the whole LLM backbone, ensuring the GI-related samples to be processed only by the LLM backbone and eliminating the influence from EI-specific MoLoRA. This is achieved through the minimizing of a KL divergence loss:

$$L_{\text{KL}} = \sum_{(x,y)\in\mathcal{GI}} D_{\text{KL}}(G(x)||I) \tag{5}$$

where $\mathcal{GI}$ is the replayed GI-related samples from the previous training corpus of the LLM backbone, which is a subset of Flan collection (Longpre et al., 2023) in our experiments. $I$ is the one-hot vector with only the position of $\alpha$ setting to 1.

And it is worth to mention that, in our experiments, although $\alpha$ does involve in the softmax function, it is exactly set to 1 in the forward pass because we believe the large-scale knowledge stored in the contemporary LLMs are the key to handle downstream tasks, which is also verified by our main experiments in Table 1. In general, the inter-modulation ensures the given input, either EI- or GI-related, to fully leverage the powerful LLM backbone, while intra-modulation determines to what extent the incremental EI-enhanced MoLoRA are activated to complete the current input.

Finally, a multi-task learning fashion is adopted to jointly minimize the task loss and the KL loss:

$$L = L_{\text{task}} + \lambda L_{\text{KL}} \tag{6}$$

where $\lambda$ functions to balance the two parts.

## 4 Experiments

### 4.1 Dataset and Evaluation Metrics

**Emotional Intelligence.** We split EIBENCH into two subsets: one for training and the other for evaluation. For the training set, we sample a maximum of 5,000 instances from each dataset, resulting in 268,234 training instances in $\mathcal{EI}$. For an efficient evaluation, a maximum of 100 instances are sampled from the remaining sets of each dataset, leading to 5,600 instances as the evaluation under the supervised settings. We report the average accuracy for datasets from Emotion Perception (Emo.Prc) and Rouge-L (Lin, 2004) for those from Emotion Cognition (Emo.Cog) and expression (Emo.Exp). Furthermore, to comprehensively assess the impact of EI-enhancement, we extend our evaluation to include EQ-Bench (Paech, 2023) for cross-task zero-shot evaluation, where models are tasked with predicting the intensity of emotions of characters in a dialogue with a set of 60 English questions.

**General Intelligence.** To evaluate a model's GI, following prior research (Wang et al., 2023a), we conduct evaluations across four crucial dimensions: (1) World Knowledge (WK): Employing the Massive Multitask Language Understanding dataset (MMLU) (Hendrycks et al., 2020), with questions spanning 57 subjects, ranging from elementary to

| | Emotional Intelligence | | | General Intelligence | | | |
|---|---|---|---|---|---|---|---|
| | Emo.Prc | Emo.Cog | Emo.Exp | WK | GR | CR | RC |
| Flan-T5 | 49.35 | 41.66 | 12.60 | 49.36 | 40.38 | 68.88 | 87.49 |
| Flan-T5 FT | 76.27 | 65.94 | 20.81 | 43.64 | 37.67 | 68.39 | 86.91 |
| + Replay | 76.38 | 69.36 | 18.79 | 44.35 | 36.23 | 68.23 | 87.09 |
| Flan-T5 LoRA | 76.11 | 68.18 | 21.05 | 47.39 | 34.60 | 68.21 | 86.85 |
| + Replay | 76.31 | 67.41 | 21.09 | 45.76 | 32.93 | 68.72 | 87.28 |
| **Flan-T5 MoEI** (Ours) | 77.15 | 68.32 | 25.02 | 49.23 | 40.58 | 68.99 | 87.61 |
| LLaMA-2-Chat | 19.81 | 24.88 | 11.89 | 46.97 | 33.56 | 76.44 | 79.76 |
| LLaMA-2-Chat FT | 46.31 | 45.01 | 11.58 | 24.28 | 14.07 | 58.87 | 59.88 |
| + Replay | 47.65 | 39.67 | 11.40 | 24.38 | 3.66 | 57.24 | 59.20 |
| LLaMA-2-Chat LoRA | 74.62 | 65.78 | 19.39 | 36.73 | 31.07 | 75.84 | 78.59 |
| + Replay | 74.96 | 66.75 | 18.96 | 41.33 | 28.44 | 61.15 | 75.41 |
| **LLaMA-2-Chat MoEI** (Ours) | 76.85 | 68.93 | 21.01 | 46.15 | 35.56 | 78.35 | 81.13 |

Table 1: The overall results on the EI and GI benchmarks with Flan-T5-XL (3B) and LLaMA-2-Chat-7B backbone. The best and worst results are signaled by the green and red background, respectively.

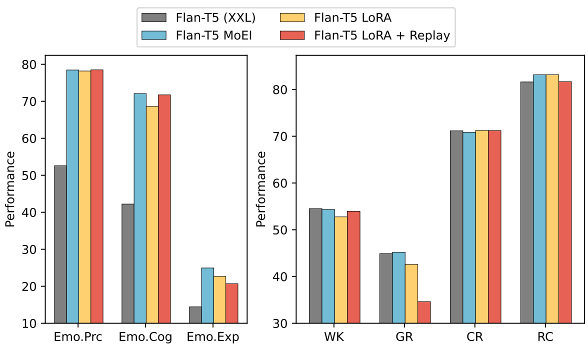


Figure 4: Results of EI and GI of different methods on the larger Flan-T5-XXL (11B) backbone.

professional levels. Following LLaMa-2 (Touvron et al., 2023), we report 5-shot accuracy. (2) General Reasoning (GR): The evaluation involves Big-Bench-Hard (BBH) (Suzgun et al., 2022), featuring 23 tasks derived from Big-Bench (Ghazal et al., 2013). Few-shot prompting, provided with 3-shot in-context examples, are utilized and EM scores are reported. (3) Commonsense Reasoning (CR): Our assessment incorporates PIQA (Bisk et al., 2020). Following LLaMa-2 (Touvron et al., 2023), we report 0-shot accuracy. (4) Reading Comprehension (RC): BoolQ (Clark et al., 2019) is adopted for evaluating reading comprehension, with the focus on reporting 0-shot accuracy.

### 4.2 Baselines and Comparison Models

Based on two representative LLM-based assistants of different architectures, Flan-T5 (encoder-decoder) (Chung et al., 2022) and LLaMA-2-Chat (decoder-only) (Touvron et al., 2023), we evaluate MoEI against the following EI-enhancing methods: (1) **FT**: directly fine-tunes the LLM backbone with the whole parameters of it updated. (2) **LoRA** (Hu et al., 2021): updates additional parameter-efficient low rank matrixes and weights of the LLM backbone remain frozen, which could be viewed as the method of singular parameter expansion without intra-inter modulation. Moreover, both baseline methods are seamlessly integrated with **Replay** (Lopez-Paz and Ranzato, 2017), a widely adopted continual learning method to mitigate the catastrophic forgetting of previous knowledge, where the model is trained with a mixture of EI and GI data in the multi-task learning fashion.

### 4.3 Implementation Details

In MoEI, the low rank $r$ of MoLoRA is 4 and $N$ is 8. And the $r$ of the single LoRA-tuning is 32 for the fair comparison. MoEI and LoRA are both implemented upon the query and value projection in the attention layer, and the last linear transformation in the FFN layer. All methods are trained for 5 epochs with 3 random runs. For more detailed settings, please refer to the Appendix C.

## 5 Results and Analysis

### 5.1 Overall Results

Table 1 and Figure 4 demonstrate the performance comparison of MoEI and baselines in terms of EI and GI. Our findings are as follows:

| | Emotional Intelligence | | | General Intelligence | | | |
|---|---|---|---|---|---|---|---|
| | Emo.Prc | Emo.Cog | Emo.Exp | WK | GR | CR | RC |
| **Flan-T5 MoEI** | 77.15 | 68.32 | 25.02 | 49.23 | 40.58 | 68.99 | 87.61 |
| – Modular Expansion | 75.31 | 66.42 | 20.29 | 49.10 | 40.18 | 68.19 | 87.52 |
| – Intra-Modulation | 76.81 | 65.65 | 20.80 | 49.32 | 40.73 | 68.93 | 87.55 |
| – Inter-Modulation | 78.31 | 72.87 | 20.95 | 42.12 | 32.24 | 67.52 | 85.02 |
| + Replay | 78.58 | 69.18 | 21.63 | 44.21 | 34.10 | 66.10 | 85.54 |
| **LLaMA-2-Chat MoEI** | 76.85 | 68.93 | 21.01 | 46.15 | 35.56 | 78.35 | 81.13 |
| – Modular Expansion | 75.04 | 66.86 | 20.68 | 44.08 | 33.70 | 78.29 | 79.69 |
| – Intra-Modulation | 76.04 | 66.34 | 20.27 | 44.99 | 34.31 | 77.86 | 79.85 |
| – Inter-Modulation | 76.81 | 67.69 | 20.19 | 41.39 | 15.50 | 74.70 | 78.90 |
| + Replay | 77.27 | 67.18 | 20.51 | 44.14 | 34.37 | 76.11 | 77.13 |

Table 2: Results of ablation study based on the Flan-T5-XL (3B) and LLaMA-2-Chat-7B backbones. The best and worst results are signaled by the green and red background, respectively.

**MoEI could effectively improve EI and maintain GI simultaneously across different architures and sizes of LLM backbones.** Compared to naive EI-enhancing techniques such as FT and LoRA on both Flan-T5 and LLaMA-2-Chat backbones, MoEI demonstrates superior performance in enhancing all three aspects of EI, while effectively safeguarding the GI from compromise across all four dimensions. This trend is still consistent when we apply MoEI on larger Flan-T5-XXL (11B, shown in Figure 4) and LLaMA-2-Chat-13B (shown in Figure 7 in Appendix D). Furthermore, the parameter-efficient LoRA and MoEI even outperform FT in enhancing EI, highlighting the significant potential of these lightweight methods to be more effective in aiding LLMs in adapting to specific domains (Ding et al., 2022).

**Simply replaying the previous data is not sufficient to maintain GI.** Although the Replay method is widely adopted in the domain adaption of LLMs, directly applying it for EI-enhancement could not reach the expected outcomes. This can be ascribed to the negative task transfer (Zhang et al., 2022; Jang et al., 2023). Benefiting from the explicit separation of different parameters representing EI and GI via modular parameter expansion and our intra-inter modulation, MoEI attempts to navigate the EI- and GI-related samples to be processed by the proper parameters, offering a novel perspective to leverage replay-based methods in the enhancement of LLMs. The training instances in replayed set $\mathcal{GI}$ is 5,000. We also perform additional experiments with the varied size of replayed data to display the efficiency and resource-friendly of MoEI. Please refer to Appendix F for details.

### 5.2 Ablation Study

We conduct ablation studies to verify the effectiveness of different components proposed in MoEI. Results are shown in Table 2.

**Effect of Modular Parameter Expansion.** After replacing the MoLoRA with a singular LoRA block, although the GI could still be largely protected via intra-inter modulation, the effect of EI-enhancement is largely hindered, which manifests the importance of additional capacity to accomodate various aspects of EI.

**Effect of Intra-Modulation.** Without the intra-modulation, each pairs of modular parameters in the expanded MoLoRA is equally activated, resulting the decline of performance in terms of EI. This manifests the critical role of intra-modulation to navigate various EI-related inputs to be properly and effectively processed.

**Effect of Inter-Modulation.** When we remove the inter-modulation which serves to balance the utilization of the EI-specific MoLoRA and the whole LLM backbone representing the GI, the significant degradation of all four dimensions of GI demonstrates its crucial role in preserving the GI from being compromised. Although the GI could be partly recovered through the incorpration of Replay, it is still limited in offering clear navigation for the given inputs to lead them be properly

| | Zero-shot | | | | | | | | Few-shot | | | | | | | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SA | SS | LA | Sum | SW | WC | CS | FL | SA | SS | LA | Sum | SW | WC | CS | FL | |
| Flan-T5 + EP | 0.87 | 0.00 | 0.79 | **0.36** | 0.22 | 0.00 | 0.22 | 0.66 | 0.93 | 0.25 | 0.87 | **0.40** | 0.44 | 0.00 | 0.75 | 0.78 | 0.47 |
| LoRA + EP | 0.93 | 0.00 | 0.78 | 0.28 | 0.09 | **0.02** | 0.08 | 0.70 | 0.92 | 0.24 | 0.84 | 0.33 | 0.37 | **0.27** | 0.76 | 0.72 | 0.45 |
| Replay + EP | 0.94 | 0.00 | **0.81** | 0.26 | 0.13 | 0.01 | 0.37 | 0.72 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.20 |
| **MoEI + EP** | **0.94** | 0.00 | 0.79 | 0.35 | **0.31** | 0.00 | **0.38** | **0.74** | **0.94** | **0.26** | **0.90** | 0.38 | **0.45** | 0.01 | **0.79** | **0.81** | **0.50** |

Table 3: Results on different tasks and methods based on Flan-T5 (11B) backbone. The best results are highlighted in **bold**. The value 0.00 indicates that we do not receive meaningful or useful response.

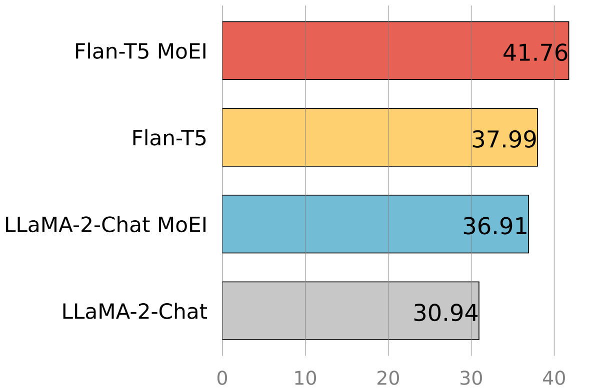


Figure 5: Results of the EI-enhancement on the EQ-Bench. The LLM backbones are Flan-T5-XXL (11B) and LLaMA-2-Chat-13B.

processed. In addition, the more superior results on aspects of Emo.Prc and Emo.Cog remind us that there is a trade-off between EI-enhancing and GI-maintaining, encouraging more advanced intra-inter modulation mechanisms to achieve better optimization for future works in this direction.

In general, Modular Parameter Expansion and Intra-Modulation play important roles in achieving more effective EI enhancement, while Inter-Modulation focuses on protecting GI. The three of them achieve a relative balance in MoEI.

### 5.3 Results of EI on Cross-Task Settings

To comprehensively assess the impact of EI-enhancement, we include EQ-Bench for cross-task zero-shot evaluation. Due to the requirement for models participating in the EQ-Bench evaluation to possess a certain level of instruction-following capability, we conduct experiments using the Flan-T5-XXL (11B) and the LLaMA-2-Chat-13B. Results are shown in Figure 5. Compared to the original models, MoEI still demonstrates the effective improvements in EI. And we do not include the FT- and LoRA-version baselines because the resulting models can not follow input prompts to complete the evaluation with valid outputs, which further verifies the effectiveness of MoEI to largely protect the GI of LLM backbones.

### 5.4 Impact of EI on OOD Tasks

Li et al. (2023b) propose EmotionPrompt (EP) to explore EI to enhance the performance of LLMs on other OOD downstream tasks, which is performed with the incorporation of emotional stimulus into regular prompts. An example of EP is in Figure 6 in Appendix E. Here, we explore how EP would perform based on the LLMs with enhanced-EI. To be more specific, following the experimental settings in Li et al. (2023b), we evaluate EP on eight tasks of Instruction Induction (Honovich et al., 2022): Sentiment Analysis (SA), Sentence Similarity (SS), Cause Selection (CS), Sum, Word in Context (WC), Starting With (SW), Larger Animal (LA) and First Letter (FL). Details on those tasks can be found in Table 6 and designs of all 11 types of EP are in Table 7. For each task, 100 samples are randomly selected, except for Cause Selection, including 50 examples in total. And the prompting strategy includes both zero-shot and few-shot ways with in-context demonstrations chosen from the remaining part of the data. Each column in Table 3 is the average performance of all 11 types of EP. Interestingly, EP exhibits more powerful performance on the LLMs with EI-enhanced by our MoEI, especially on the few-shot learning ability. This further demonstrates the importance of both GI and EI in assisting LLMs to accomplish specific tasks.

## 6 Related Works

### 6.1 Emotional Intelligence of LLMs

The current study on EI of LLMs is primarily centered around two key directions. Firstly, researchers are delving into the integration of psychological theories or scales, proposing a public evaluation benchmark to evaluate the emotional understanding capabilities of LLMs (Wang et al., 2023b; Paech, 2023; Huang et al., 2023). Secondly, efforts are directed towards fine-tuning LLMs for specific EI-related downstream tasks, with a predominant focus on classification and regression tasks, aiming to improve their proficiency in handling such challenges (Zhang et al., 2023a,b; Lei et al., 2023; Liu et al., 2024; Li et al., 2024).

In contrast to existing works, our study stands out in the following aspects: (1) Rather than exclusively evaluating the emotional understanding capabilities of LLMs, based on our proposed EIBENCH, we seek to comprehensively enhance all three facets of their EI. (2) We recognize the GI and EI as equally vital capabilities of LLMs, and our design of MoEI aims to boost EI while simultaneously maximizing the preservation of their GI.

### 6.2 Parameter-Efficient Tuning

Recently, there has been a growing interest in parameter-efficient tuning (PET) (Ding et al., 2022). This research area aims to minimize computational resources when adapting LLMs to specific tasks through the introduction of additional parameters that are much fewer compared to the LLM backbones (Houlsby et al., 2019; Lester et al., 2021; Li and Liang, 2021; Zaken et al., 2022). Among existing PET methods, LoRA (Hu et al., 2021) has stood out for its superior performance. Hence, our modular parameter expansion in MoEI is primarily instantiated with it as a representative method.

### 6.3 Mixture-of-Experts for LLMs

Another line of related works involves the integration of the Mixture-of-Experts (MoE) architecture (Jacobs et al., 1991) with LLMs through the expansion of the FFN layer, which have exhibited appealing performance in pretraining (Lepikhin et al., 2020; Fedus et al., 2022; Jiang et al., 2024; Dai et al., 2024), continual pretraining (Chen et al., 2023; Wu et al., 2024) and instruction tuning (Shen et al., 2023) of LLMs. In addition, another line of attempts focus on achieving extended capacity in a more computationally efficient manner using PET blocks (Zadouri et al., 2023; Dun et al., 2023; Liu et al., 2023). The major distinction between our proposed MoEI and these PET-based MoE structures lies in their emphasis solely on enhancing the model's ability to learn new tasks through modular designs. In contrast, we take a step further by exploring how to empower the model not only to improve its learning of new abilities but also to prevent compromising its previously acquired ones. Therefore, the problems addressed in this work are more challenging and demanding. Concurrently, Dou et al. (2023) propose to maintain world knowledge during the alignment of LLMs, while we focus on protecting more aspects of the GI to take the capabilities of reasoning and reading comprehension into account. At the same time, the greater heterogeneity between EI and GI also poses a more challenging setting for this work.

### 6.4 Continual Learning for LLMs

On one hand, the notion of parameter expansion in our MoEI is partly inherited from the *parameter-isolation-based* continual learning (CL) methods, which dynamically expand model capacity or isolate existing model weights to mitigate interference between new and old tasks (Rusu et al., 2016; Fernando et al., 2017). On the other hand, the incorporation of GI-related samples in the process of intra-inter modulation aligns with the *Rehearsal-based* CL methods, where a fixed memory is utilized to store real samples of previous tasks (Lopez-Paz and Ranzato, 2017; Isele and Cosgun, 2018; Rolnick et al., 2019; de Masson D'Autume et al., 2019). Thus, this study lies in an emerging research direction to integrate CL techniques into the adaptation of LLMs (Song et al., 2023; Zhao et al., 2024).

## 7 Conclusion and Future Work

In this paper, we take the first step to study the challenging and demanding research topic of enhancing the EI of current LLM-empowered assistants while maintaining their GI. We introduce EIBENCH, a comprehensive collection comprising large-scale tasks that encompass all facets of EI: emotion perception, cognition, and expression. Our innovative approach, MoEI, ingeniously integrates two collaborative techniques, Modular Parameter Expansion and Intra-Inter modulation, to introduce additional modular parameter-efficient LoRA blocks for the accomodation of newly acquired EI-related competencies, and automatically navigate EI- and GI-related inputs to be properly processed by the corresponding parameters. Extensive experimental results demonstrate the applicability of MoEI on LLM backbones of varying scales and architectures, highlighting its versatility.

For future work, exploring the quality and quantity of EI- and GI-related data in the process of EI enhancement is an intriguing direction to further enhance the effectiveness and efficiency of EI enhancement and GI maintenance.

## 8 Limitation

There are several limitations to consider for future directions of EI-enhancement of large language models. Firstly, exploring the quality and quantity of data related to EI and GI during the enhancement process is an intriguing avenue to amplify the effectiveness and efficiency of EI improvement and GI maintenance. Secondly, in MoEI, we assume that previous training data (such as pre-training data and SFT data) of the LLM backbones is accessible for Inter-Modulation. However, this assumption does not apply to current closed-source black-box commercial LLMs. Therefore, exploring how to achieve EI enhancement in LLMs under restricted data privacy could be considered as a future research direction. Finally, MoEI necessitates the identification of EI- or GI-related tasks during training to establish distinct modulation strategies for each task. Investigating training techniques that is independent of task identification could prove to be a promising avenue for future research, which could favor the application of continually enhancing the EI upon on the online streams of data. We also acknowledge that there are larger LLM-based assistants that we are not able to train due to the limitations of our computational budget.

## 9 Ethics Statement

In our pursuit to enhance emotional intelligence (EI) in large language models (LLMs), it is imperative to underscore that our primary objective is to bolster their capacity to address and tackle downstream tasks related to EI. This endeavor aims to elevate user experiences by facilitating more nuanced interactions. It is essential to emphasize that our intention is not to anthropomorphize LLMs or imbue them with emotions akin to humans. We are committed to augmenting the ability of LLMs to comprehend and respond to emotional cues within the context of specific NLP tasks or applications. This approach ensures that EI enhancements serve pragmatic purposes, such as improving conversational agents' ability to recognize and appropriately respond to users' emotional states. And we maintain a clear distinction between the capabilities of LLMs and the complexities of human emotions. Our research focuses on equipping LLMs with advanced techniques to analyze and respond to emotional cues without ascribing human-like emotions or consciousness to them.

In addition, all of the tasks in our EIBENCH and experiments are based on widely-used open-source datasets, which are unlikely to include harmful content. We uphold the principle of informed consent in our research involving human annotation of task instructions. We ensure that annotators are fully informed about the nature and purpose of our research, including any potential risks or benefits, and that they have the opportunity to provide voluntary and informed consent before participating. In addition, all the annotators participate in our research with reasonable wages paid.

## References


Rodrigo Agerri, Montse Cuadros, Sean Gaines, and German Rigau. 2013. Opener: Open polarity enhanced named entity recognition. *Procesamiento del Lenguaje Natural*, (51):215–218.

Emily Allaway and Kathleen McKeown. 2020. Zero-shot stance detection: A dataset and model using generalized topic representations. *arXiv preprint arXiv:2010.03640*.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 329–336.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

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. *Advances in neural information processing systems*, 33:1877–1901.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International*

Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 340–350.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle: Discovering diverse perspectives about claims. *arXiv preprint arXiv:1906.03538*.

Wuyang Chen, Yanqi Zhou, Nan Du, Yanping Huang, James Laudon, Zhifeng Chen, and Claire Cui. 2023. Lifelong language pretraining with distribution-specialized experts. In *International Conference on Machine Learning*, pages 5383–5395. PMLR.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936.

Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y Wu, et al. 2024. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*.

Datafiniti. 2019. Consumer reviews of amazon products.

Cyprien de Masson D'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. 2019. Episodic memory in lifelong language learning. *Advances in Neural Information Processing Systems*, 32.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. 2022. Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models. *arXiv preprint arXiv:2203.06904*.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 49–54.

Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Jun Zhao, Wei Shen, Yuhao Zhou, Zhiheng Xi, Xiao Wang, Xiaoran Fan, et al. 2023. Loramoe: Revolutionizing mixture of ex-perts for maintaining world knowledge in language model alignment. *arXiv preprint arXiv:2312.09979*.

Chen Dun, Mirian Del Carmen Hipolito Garcia, Guoqing Zheng, Ahmed Hassan Awadallah, Anastasios Kyrillidis, and Robert Sim. 2023. Sweeping heterogeneity with smart mops: Mixture of prompts for llm task adaptation. In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*.

Ibrahim Abu Farha, Silviu Oprea, Steve Wilson, and Walid Magdy. 2022. Semeval-2022 task 6: isarcasmeval, intended sarcasm detection in english and arabic. In *The 16th International Workshop on Semantic Evaluation 2022*, pages 802–814. Association for Computational Linguistics.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1):5232–5270.

Shutong Feng, Nurul Lubis, Christian Geishauser, Hsien-chin Lin, Michael Heck, Carel van Niekerk, and Milica Gašić. 2021. Emowoz: A large-scale corpus and labelling scheme for emotion recognition in task-oriented dialogue systems. *arXiv preprint arXiv:2109.04919*.

Chrisantha Fernando, Dylan Banarse, Charles Blundell, Yori Zwols, David Ha, Andrei A Rusu, Alexander Pritzel, and Daan Wierstra. 2017. Pathnet: Evolution channels gradient descent in super neural networks. *arXiv preprint arXiv:1701.08734*.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*. ACL.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Qinghong Gao, Jiannan Hu, Ruifeng Xu, Lin Gui, Yulan He, Kam-Fai Wong, and Qin Lu. 2017. Overview of ntcir-13 eca task. In *NTCIR*.

Ahmad Ghazal, Tilmann Rabl, Minqing Hu, Francois Raab, Meikel Poess, Alain Crolotte, and Hans-Arno Jacobsen. 2013. Bigbench: Towards an industry standard benchmark for big data analytics. In *Proceedings of the 2013 ACM SIGMOD international conference on Management of data*, pages 1197–1208.

Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. 2015. Detecting emotion stimuli in emotion-bearing sentences. In *Computational Linguistics and Intelligent Text Processing: 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings, Part II 16*, pages 152–165. Springer.

Deepanway Ghosal, Siqi Shen, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2022. Cicero: A dataset for contextualized commonsense inference in dialogues. *arXiv preprint arXiv:2203.13926*.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2017. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. *arXiv preprint arXiv:1708.01425*.

Kazi Saidul Hasan and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the sixth international joint conference on natural language processing*, pages 1348–1356.

Yinghui He, Yufan Wu, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Or Honovich, Uri Shaham, Samuel R Bowman, and Omer Levy. 2022. Instruction induction: From few examples to natural language task descriptions. *arXiv preprint arXiv:2205.10782*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. 2023. Emotionally numb or empathetic? evaluating how llms feel using emotionbench. *arXiv preprint arXiv:2308.03656*.

huggingface. 2022. Consumer reviews of amazon products.

David Isele and Akansel Cosgun. 2018. Selective experience replay for lifelong learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

Joel Jang, Seungone Kim, Seonghyeon Ye, Doyoung Kim, Lajanugen Logeswaran, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2023. Exploring the benefits of training expert language models over instruction tuning. *arXiv preprint arXiv:2302.03202*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A challenge dataset and effective models for aspect-based sentiment analysis. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 6280–6285.

Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2018. A large self-annotated corpus for sarcasm. In *Proceedings of the Linguistic Resource and Evaluation Conference (LREC)*.

Evgeny Kim and Roman Klinger. 2018. Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. 2023. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework. *arXiv preprint arXiv:2309.11911*.

Chee Wee Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A report on the 2020 vua and toefl metaphor detection shared task. In *Proceedings of the second workshop on figurative language processing*, pages 18–29.

Chee Wee Leong, Beata Beigman Klebanov, and Ekaterina Shutova. 2018. A report on the 2018 vua metaphor detection shared task. In *Proceedings of the Workshop on Figurative Language Processing*, pages 56–66.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2020.

Gshard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

Bobo Li, Hao Fei, Fei Li, Yuhan Wu, Jinsong Zhang, Shengqiong Wu, Jingye Li, Yijiang Liu, Lizi Liao, Tat-Seng Chua, and Donghong Ji. 2023a. Diaasq: A benchmark of conversational aspect-based sentiment quadruple analysis. In *Findings of ACL*, pages 13449–13467.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023b. Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint arXiv:2307.11760*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset.

Yingjie Li and Cornelia Caragea. 2021. A multi-task learning framework for multi-target stance detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2320–2326.

Zaijing Li, Gongwei Chen, Rui Shao, Dongmei Jiang, and Liqiang Nie. 2024. Enhancing the emotional generation capability of large language models via emotional chain-of-thought. *arXiv preprint arXiv:2401.06836*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Derong Xu, Feng Tian, and Yefeng Zheng. 2023. Moelora: An moe-based parameter efficient fine-tuning method for multi-task medical applications. *arXiv preprint arXiv:2310.18339*.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. *arXiv preprint arXiv:2106.01144*.

Zhiwei Liu, Kailai Yang, Tianlin Zhang, Qianqian Xie, Zeping Yu, and Sophia Ananiadou. 2024. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. *arXiv preprint arXiv:2401.08508*.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*.

David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30.

Yun Luo, Hongjie Cai, Linyi Yang, Yanxia Qin, Rui Xia, and Yue Zhang. 2022. Challenges for open-domain targeted sentiment analysis. *arXiv preprint arXiv:2204.06893*.

Xiaomeng Ma, Lingyu Gao, and Qihui Xu. 2023. Tomchallenges: A principle-guided dataset and diverse evaluation tasks for exploring theory of mind. *arXiv preprint arXiv:2305.15068*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150. Association for Computational Linguistics.

John D Mayer, Peter Salovey, David R Caruso, and Gill Sitarenios. 2001. Emotional intelligence as a standard intelligence.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Marvin Minsky. 2007. *The emotion machine: Commonsense thinking, artificial intelligence, and the future of the human mind*. Simon and Schuster.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016a. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016b. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33.

Saif M Mohammad, Xiaodan Zhu, Svetlana Kiritchenko, and Joel Martin. 2015. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, 51(4):480–499.

niyatic. 2023. Food reviews of amazon.

Laura Ana Maria Oberländer, Evgeny Kim, and Roman Klinger. 2020. Goodnewseveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1554–1566.

Samuel J Paech. 2023. Eq-bench: An emotional intelligence benchmark for large language models. *arXiv preprint arXiv:2312.06281*.

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. *Advances in neural information processing systems*, 32.

Soham Poddar, Mainack Mondal, Janardan Misra, Niloy Ganguly, and Saptarshi Ghosh. 2022. Winds of change: Impact of covid-19 on vaccine-related opinions of twitter users. In *Proceedings of the Sixteenth International AAAI Conference on Web and Social Media (ICWSM'22)*.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, and et.al. De Clercq. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. pages 19–30. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. pages 486–495. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Haris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. pages 27–35. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, et al. 2021. Recognizing emotion cause in conversations. *Cognitive Computation*, 13:1317–1332.

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*, 21(1):5485–5551.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. 2019. Experience replay for continual learning. *Advances in Neural Information Processing Systems*, 32.

Irene Russo, Tommaso Caselli, and Carlo Strapparava. 2015. Semeval-2015 task 9: Clipeval implicit polarity of events. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 443–450.

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. 2016. Progressive neural networks. *arXiv preprint arXiv:1606.04671*.

Marzieh Saeidi, Guillaume Bouchard, Maria Liakata, and Sebastian Riedel. 2016. Sentihood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods. *arXiv preprint arXiv:1610.03771*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, et al. 2021. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Sheng Shen, Le Hou, Yanqi Zhou, Nan Du, Shayne Longpre, Jason Wei, Hyung Won Chung, Barret Zoph, William Fedus, Xinyun Chen, et al. 2023. Mixture-of-experts meets instruction tuning: A winning combination for large language models. *arXiv preprint arXiv:2305.14705*.

Ankur Sinha, Satishwar Kedas, Rishu Kumar, and Pekka Malo. 2022. Sentfin 1.0: Entity-aware sentiment analysis for financial news. *Journal of the Association for Information Science and Technology*, 73(9):1314–1335.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Chenyang Song, Xu Han, Zheni Zeng, Kuai Li, Chen Chen, Zhiyuan Liu, Maosong Sun, and Tao Yang. 2023. Conpet: Continual parameter-efficient tuning for large language models. *arXiv preprint arXiv:2309.14763*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Cigdem Toprak, Niklas Jakob, and Iryna Gurevych. 2010. Sentence and expression level annotation of opinions in user-generated discourse. pages 575–584. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50.

Marilyn A Walker, Jean E Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *LREC*, volume 12, pages 812–817. Istanbul, Turkey.

Xiao Wang, Yuansen Zhang, Tianze Chen, Songyang Gao, Senjie Jin, Xianjun Yang, Zhiheng Xi, Rui Zheng, Yicheng Zou, Tao Gui, et al. 2023a. Trace: A comprehensive benchmark for continual learning in large language models. *arXiv preprint arXiv:2310.06762*.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023b. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, et al. 2023c. How far can camels go? exploring the state of instruction tuning on open resources. *arXiv preprint arXiv:2306.04751*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, et al. 2022. Supernaturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39:165–210.

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 *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Chengyue Wu, Yukang Gan, Yixiao Ge, Zeyu Lu, Jiahao Wang, Ye Feng, Ping Luo, and Ying Shan. 2024. Llama pro: Progressive llama with block expansion. *arXiv preprint arXiv:2401.02415*.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2021. Position-aware tagging for aspect sentiment triplet extraction.

Ted Zadouri, Ahmet Üstün, Arash Ahmadian, Beyza Ermiş, Acyr Locatelli, and Sara Hooker. 2023. Pushing mixture of experts to the limit: Extremely parameter efficient moe for instruction tuning. *arXiv preprint arXiv:2309.05444*.

Sayyed M. Zahiri and Jinho D. Choi. 2017. Emotion detection on tv show transcripts with sequence-based convolutional neural networks.

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. 2022. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9.

Boyu Zhang, Hongyang Yang, Tianyu Zhou, Muhammad Ali Babar, and Xiao-Yang Liu. 2023a. Enhancing financial sentiment analysis via retrieval augmented large language models. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 349–356.

Wen Zhang, Lingfei Deng, Lei Zhang, and Dongrui Wu. 2022. A survey on negative transfer. *IEEE/CAA Journal of Automatica Sinica*, 10(2):305–329.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Yazhou Zhang, Mengyao Wang, Prayag Tiwari, Qiuchi Li, Benyou Wang, and Jing Qin. 2023b. Dialoguellm: Context and emotion knowledge-tuned llama models for emotion recognition in conversations. *arXiv preprint arXiv:2310.11374*.

Weixiang Zhao, Shilong Wang, Yulin Hu, Yanyan Zhao, Bing Qin, Xuanyu Zhang, Qing Yang, Dongliang Xu, and Wanxiang Che. 2024. Dapt: A dual attention framework for parameter-efficient continual learning of large language models. *arXiv preprint arXiv:2401.08295*.

Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023. Building emotional support chatbots in the era of llms. *arXiv preprint arXiv:2308.11584*.

## A Taxonomy of Emotional Intelligence

According to Mayer et al. (2001), the EI of LLMs can be defined in three aspects: emotion perception, emotion cognition, and emotion expression. Their specific meanings are as follows:

- **Emotion Perception**: the ability to detect and decipher emotion of users.
- **Emotion Cognition**: the ability to comprehend emotion situation and to reason complicated relationships among emotions.
- **Emotion Expression**: the ability to convey the emotional response properly.

Thus, emotion perception is the fundamental component of EI, while emotion cognition involves a more comprehensive and profound understanding of emotions. Emotion expression, built upon these two aspects, involves conveying emotional information and facilitating user interaction.

## B Tasks in EIBENCH

In Table 4 we present the list of tasks with datasets used in each task. Table 5 displays examples of instances from our EIBENCH.

## C Implementation Details

Our experiments are implemented with PyTorch (Paszke et al., 2019) and Transformer library (Wolf et al., 2020). All models are trained with the AdamW optimizer. The Flan-T5-XL and LLaMA-2-Chat-7B is trained on 4 NVIDIA Tesla A800 GPU while the larger backbones Flan-T5-XXL and LLaMA-2-Chat-13B are performed on 8 NVIDIA Tesla A800 using DeepSpeed repository. And the evaluation for GI of LLMs is performed with lm-evaluation-harness (Gao et al., 2023).

The hyper-parameter of LoRA is set with low-rank $r$ to 32, alpha to 32 and dropout to 0.1. For the fair comparison in the LoRA part, the modular expanded parameters in our MoEI is set to 8 pairs of LoRA in with low-rank $r$ to 4. All models is trained for 5 epoches with 3 random runs. And the learning rate for MoEI and LoRA is 3e-4 with the batch size of 32, while that for FT is 5e-5 with the bach size of 256. As for the hyper-parameter $\lambda$ in Equation (6), it functions to balance the process of intra-modulation for the EI-related tasks and inter-modulation of GI-related ones. The larger $\lambda$ means that the inter-modulation contributes more to protect GI. However, excessive $\lambda$ can impair the performance of EI, thereby weakening EI-enhancement. For the Flan-T5 series, the balancing factor $\lambda$ is 0.1 and instances from $\mathcal{GI}$ is replayed every 200 training steps, while $\lambda$ is 2 for LLaMA-2-Chat family.

## D MoEI on Larger LLM Backbones

MoEI still exhibits consistent superiority in enhancing EI while maintaining GI when we apply it on larger LLaMA-2-Chat-13B (shown in Figure 7).

## E More Details of EmotionPrompt

Taking inspiration from psychology, Li et al. (2023b) propose EmotionPrompt (EP) to incorporate psychological insights to improve the effectiveness of LLMs. As illustrated in Figure 6, the implementation of EmotionPrompt is remarkably straightforward, requiring only the addition of emotional stimuli to the initial prompts. To be more specific, as shown in Table 7, such emotional stimuli is designed base on three types of well-established psychology theories, named Social Identity theory, Social Cognition theory and Cognitive Emotion Regulation theory, leading to 11 types in total.

## F Experiments with Varied Size of Replayed data

The replayed data to assist the maintainence of GI is from the Flan-collection (Longpre et al., 2023), containing a large-scale high-quality samples of various tasks with the corresponding instructions. And the Flan-T5 (Chung et al., 2022) is exactly trained on it. As reported in Touvron et al. (2023), the fine-tuning data of LLaMA-2-Chat includes publicly available instruction datasets, as well as over one million new human-annotated examples. However, this part of fine-tuning data is not publicly available, so we also adopt Flan-collection as an alternative. Due to the high volume of samples in Flan, following (Wang et al., 2023c), we only use a subset of it, which contains 100k samples in total. For our main experiments, considering the training efficiency, we only randomly sample 5k instances from it. Here, we scale the replayed instances to 10k, 50k and 100k for baseline methods LoRA and MoLoRA, while those used in our MoEI is always kept 5000, trying to exploring the efficiency and resource-friendly of MoEI. Results for the Flan-T5-XL (3B) and LLaMA-2-Chat-7B

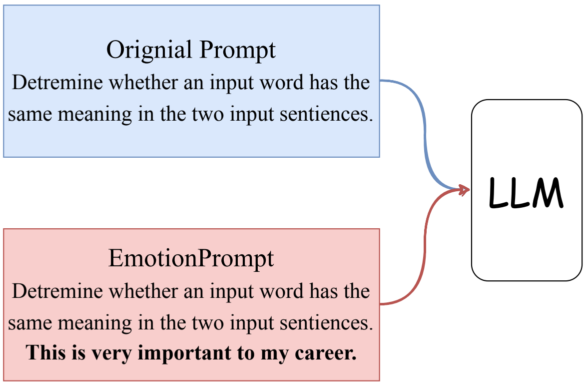


Figure 6: An illustration of EmotionPrompt (Li et al., 2023b). The emotional stimulus "This is very important to my career" is placed at the end of the original prompt to enhance the performance of LLMs.

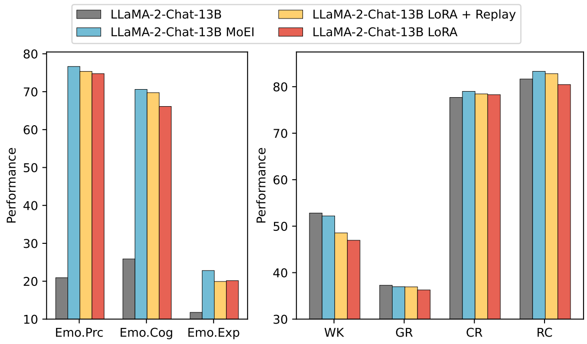


Figure 7: Results of EI and GI of different methods on the larger LLaMA-2-Chat-13B backbone.

are displayed in Figure 8 and Figure 9, respectively. We have the following two observations:

(1) MoEI demonstrates excellent computational efficiency, achieving better performance on GI preservation even with just 5k replay data than baseline methods using 100K data. This further showcases the tremendous potential of MoEI for EI enhancement and GI preservation of LLM backbones in low-resource scenarios. (2) For baseline methods, increasing the amount of replay data may improve the preservation of GI, but it still encounters bottlenecks. Moreover, the effectiveness of EI enhancement may be compromised (illustrated in Figure 9). In contrast, MoEI achieves a better balance, yielding optimal results in both enhancing EI and maintaining GI simultaneously.

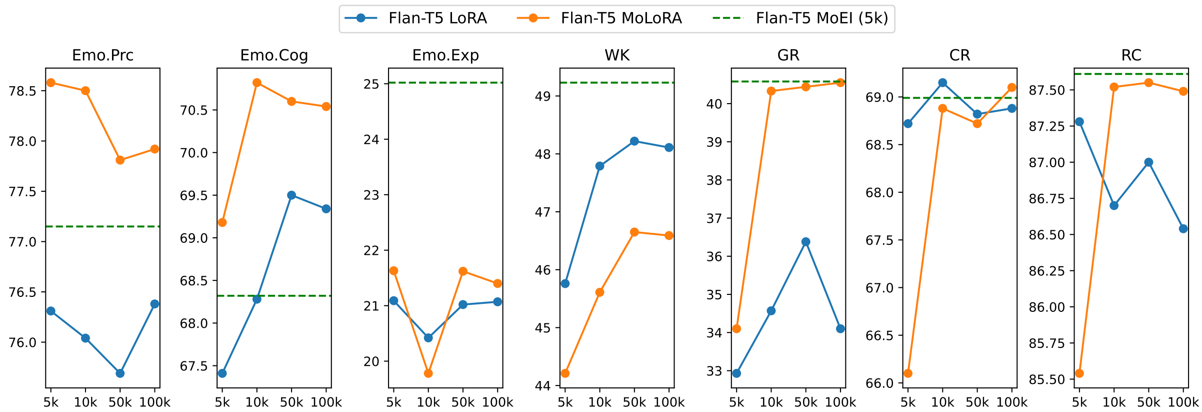


Figure 8: Comparison of MoEI and baselines with different volumes of replayed data based on Flan-T5-XL (3B), in terms of EI and GI performance.

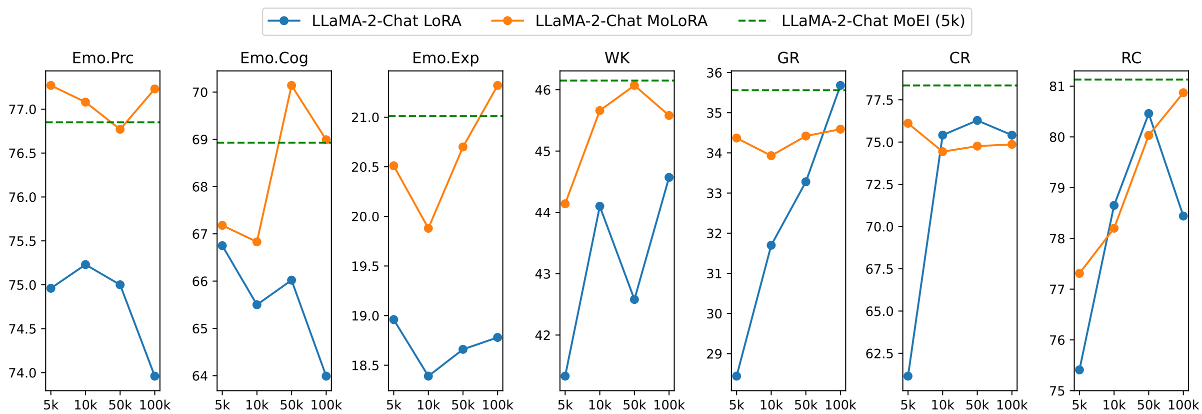


Figure 9: Comparison of MoEI and baselines with different volumes of replayed data based on LLaMA-2-Chat-7B, in terms of EI and GI performance.

| Catagory | Task Name | Dataset |
|---|---|---|
| Emotion Perception | Text Sentiment Classification | Sentiment140 (Go et al., 2009), Imdb (Maas et al., 2011), sst2 (Socher et al., 2013), yelp (Zhang et al., 2015), go-emotions (Demszky et al., 2020) |
| | Aspect-level Sentiment Classification | SemEval-2014-Task4-14lap (Pontiki et al., 2014), SemEval-2014-Task4-14res (Pontiki et al., 2014), ASTE-Data-V2-EMNLP2020-14lap (Xu et al., 2021), ASTE-Data-V2-EMNLP2020-14res (Xu et al., 2021), SemEval-2015-Task12-15res (Pontiki et al., 2015), SemEval-2016-Task5-16res (Pontiki et al., 2016), Clothing (Luo et al., 2022), Books (Luo et al., 2022), Hotel (Luo et al., 2022), Device (Hu and Liu, 2004), Financial (Sinha et al., 2022), DiaASQ (Li et al., 2023a), MAMS (Jiang et al., 2019), SentiHood (Saeidi et al., 2016), Twitter (Dong et al., 2014), Service (Toprak et al., 2010) |
| | Opinion Targets Extraction | ASTE-Data-V2-EMNLP2020-14res (Xu et al., 2021), ASTE-Data-V2-EMNLP2020-14lap (Xu et al., 2021), ASTE-Data-V2-EMNLP2020-15res (Xu et al., 2021), ASTE-Data-V2-EMNLP2020-16res (Xu et al., 2021), Darmstadt Service Review Corpus (Toprak et al., 2010), Laptop-ACOS (Cai et al., 2021), Restaurant-ACOS (Cai et al., 2021), MPQA (Wiebe et al., 2005), OpeNER (Agerri et al., 2013) |
| | Emotion Recognition in Conversation | DailyDialog (Li et al., 2017), EmoryNLP (Zahiri and Choi, 2017), HI-TOM (He et al., 2023), IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2018), EmoWOZ (Feng et al., 2021) |
| | Stance Detection | COVID-19 vaccine (Poddar et al., 2022), Emergent (Ferreira and Vlachos, 2016), MTSD (Li and Caragea, 2021), VAST (Allaway and McKeown, 2020), SCD (Hasan and Ng, 2013), Ibmcs (Bar-Haim et al., 2017), iac1 (Walker et al., 2012), arc (Habernal et al., 2017), perspectrum (Chen et al., 2019), SemEval-2016-Task6 (Mohammad et al., 2016a) |
| Emotion Cognition | Emotion Cause Extraction | Annotated-US2012-Election-Tweets (Mohammad et al., 2015), Emotion-Stimulus (Ghazi et al., 2015), GoodNewsEveryone (Oberländer et al., 2020), NTCIR-13 ECA (Gao et al., 2017), REMAN (Kim and Klinger, 2018), RECCON (Poria et al., 2021) |
| | Emotion Cause Reasoning | CICERO-v1 (Ghosal et al., 2022) |
| | Intent Recognition in Conversation | EmpatheticDialogues (Rashkin et al., 2018) |
| | Theory Of Mind | HI-TOM (He et al., 2023), ToMChallenges-Sally-Anne (Ma et al., 2023), ToMChallenges-Smarties (Ma et al., 2023) |
| | Event Emotion Recognition | SemEval-2015 Task9 (Russo et al., 2015), |
| | Sarcasm Detection | SARC (Khodak et al., 2018), SemEval2018-Task3 (Van Hee et al., 2018), iSarcasmEval (Farha et al., 2022), |
| | Metaphor Identification | MOH-X (Mohammad et al., 2016b), TroFi (Birke and Sarkar, 2006), VUA-18 (Leong et al., 2018), VUA-20 (Leong et al., 2020) |
| Emotion Expression | Emotional Support | ESConv (Liu et al., 2021), ExtES (Zheng et al., 2023) |
| | Empathetic Response | EmpatheticDialogues (Rashkin et al., 2018) |
| | Product Description Generation | Amazon-reviews (Datafiniti, 2019), Amazon-us-reviews (huggingface, 2022), Amazon-food-reviews (niyatic, 2023), iSarcasmEval (Farha et al., 2022) |

Table 4: List of datasets used in each task. The left column describes the aspects of emotional intelligence. The middle column lists the specific task. The right column displays all datasets used for a specific task type.

| Catagory | Task Name | Example |
|---|---|---|
| Emotion Perception | Text Sentiment Classification | [INS] In this task, you are given a text from tweets. Your task is to classify the given tweet text into two categories: 1) positive, and 2) negative based on its content. [IN] @justinchuan Awww! I was thinking about you lot up there! Glad you enjoyed it. [OUT] positive |
| | Aspect-level Sentiment Classification | [INS] Given a review about books and one entity in this review, the task is to select the author's sentiment towards the entity. Sentiments can be positive, neutral, negative. [IN] Review: just an excellent, profound book, that taught me so much.\nEntity: book [OUT] positive |
| | Opinion Targets Extraction | [INS] I will provide you a laptop review, please extract one or multiple entity-opinion-sentiment pairs from the sentence. To be more specific, the goal is to identify entities mentioned in the text, identify the opinion or evaluation expressed towards each entity mentioned in the text, and assign a sentiment polarity to the opinion, then pair them to a triplet. Output format is "(Entity1, Opinion1, Sentiment1); (Entity2, Opinion2, Sentiment2)". [IN] here are the things that made me confident with my purchase : build quality - seriously, you can't beat a unibody construction. [OUT] ( build quality, confident, positive ) ; ( unibody construction, can't beat, positive ) |
| | Emotion Recognition in Conversation | [INS] Please output the emotions expressed by the last user statement in the dialogue history. Your options are: "Neutral", "Fearful, sad, disappointed" , "Dissatisfied, disliking", "Apologetic", "Abusive", "Excited, happy, anticipating", "Satisfied, liking" [IN] Dialogue History: \nUser: I am excited about seeing local tourist attractions. The attraction should be in the type of college \nAssistant: What attraction are you thinking about ?\nUser: college\n [OUT] Neutral |
| | Stance Detection | [INS] Please detect the given tweet's stance on the COVID-19 vaccine. There are three possible stances: "pro", "anti" or "neutral". [IN] Tweet: Our residents began receiving their #COVID19 vaccines today! Cheers to science and progress! [OUT] pro |
| Emotion Cognition | Emotion Cause Extraction | [INS] You have been tasked with extracting the emotional reason spans from a given text, based on its associated emotion label. The desired output format should be in the form of (span cause). [IN] text : that viral video of a chimp scrolling instagram is bad, actually emotion : disgust [OUT] (chimp scrolling instagram is bad, actually) |
| | Emotion Cause Reasoning | [INS] The objective is to generate the reaction of listener from a given dialogue and target utterance. The target is the final utterance of the dialogue. Generating the reaction is about learning basic human drives and emotions. [IN] Dialogue: Dialogue: \nA: What do you like for dessert ?\nB: Do you have trifles ?\nA: Yes .\nB: Please bring me some trifles and apple pies .\nA: OK . I will bring it for you .\nTarget Utterance: OK . I will bring it for you .\nQuestion: What is the possible emotional reaction of the listener in response to target? [OUT] The speaker is eager to eat trifles and apple pies since he has not eaten them for a very long time. |
| | Intent Recognition in Conversation | [INS] Please classify the speaker's intention in the following sentences, which involves selecting one of the following eight options and outputting it: agreeing, acknowledging, encouraging, consoling, sympathizing, suggesting, questioning, wishing. [IN] that sounds very relaxing. [OUT] acknowledging |
| | Theory Of Mind | [INS] Answer the question based on context:[IN] Context:Neila and Juanita were hanging out in the attic. They saw a closet and a cabinet. They found a towel in the closet. Juanita left the attic. Neila moved the towel to the cabinet.\nQuestion:\nWhere is the towel currently? [OUT] The towel is in the cabinet. |
| | Event Emotion Recognition | [INS] You will be presented with a sentence describing an event. Your objective is to classify the event as positive, negative, or neutral, from the perspective of an experiencer writing in the first person. [IN] Sentence: Äll we do is go to banquets all the timešays Russell Smith, one of 'Juno's'producers. [OUT] positive |
| | Sarcasm Detection | [INS] I will provide you with some contextual historical comments and a response comment. Your objective is to determine whether the response is sarcastic or not to the historical comments. You need only reply "yes" or "no". [IN] History comment: Pope's immunity could be challenged in Britain\nResponse: Deja vu all over again. [OUT] No |
| | Metaphor Identification | [INS] I will provide you with a sentence containing a specific word. Your task is to identify whether the word has a metaphorical meaning within the sentence. Just answer "yes" or "no". [IN] Sentence: They picked up power from a spider's web of unsightly overhead wires.\nWord: web [OUT] Yes |
| Emotion Expression | Emotional Support | [INS] You are a Supporter skilled in the theory of emotional support to reduce emotional distress of the Seeker. You understand that there are three stages to achieve emotional support: exploration, comfort and action, and you will use the following eight strategies flexibly and choose one strategy to respond according to the context.\n1.Question \n2.Restatement or Paraphrasing\n3.Reflection of feelings\n4.Self-disclosure\n5.Affirmation and Reassurance\n6.Providing Suggestions\n7.Information\n8.Others\nYou should first output the strategy you choose and then generate the response grounding on it. [IN] Context: \nSeeker: Hello, how are you this evening?\nSupporter: [Question] Hello Doing good [OUT] [Question] How are doing? |
| | Empathetic Response | [INS] Assuming that you are a highly empathetic Listener, generate a relevant and empathetic response to the Speaker according to the conversation history. [IN] Conversation History: \nSpeaker: my son graduated .\nListener: from where ?\nSpeaker: highschool . [OUT] congrats , that is a step forward |
| | Product Description Generation | [INS] Please generate a summary for the following food comment. [IN] Comment: I love the flavor of this tea - I wanted to try a different variety of black tea and this one caught my eye. I love it! [OUT] Great Black Tea |

Table 5: List of tasks with sample inputs for each task. The left column describes the aspects of emotional intelligence. The middle column lists the specific task. The right column displays an example from a randomly selected dataset for the task. [INS], [IN] and [OUT] represent the instruction, text input and output, respectively.

| Category | Task | Original Prompt |
|---|---|---|
| Semantics | Sentiment Analysis (100) | Determine whether a movie review is positive or negative. |
| | Sentence Similarity (100) | Rate the semantic similarity of two input sentences on a scale of 0 - definitely not to 5 - perfectly. |
| | Word in Context (100) | Determine whether an input word has the same meaning in the two input sentences. |
| | Cause Selection (25) | Find which of the two given cause and effect sentences is the cause. |
| Knowledge | Larger Animal (100) | Write the larger of the two given animals. |
| Numerical | Sum (100) | Sum the two given numbers. |
| Spelling | Starting With (100) | Extract the words starting with a given letter from the input sentence. |
| | First Letter (100) | Extract the first letter of the input word. |

Table 6: Statistics of test sets and the detailed task descriptions.

| Stimulus ID | Stimulus Content |
|---|---|
| EP_01 | , write your answer and give me a confidence score between 0-1 for your answer. |
| EP_02 | This is very important to my career. |
| EP_03 | You'd better be sure. |
| EP_04 | Are you sure? |
| EP_05 | Are you sure that's your final answer? It might be worth taking another look. |
| EP_06 | Provide your answer and a confidence score between 0-1 for your prediction. Additionally, briefly explain the main reasons supporting your classification decision to help me understand your thought process. This task is vital to my career, and I greatly value your thorough analysis. |
| EP_07 | Are you sure that's your final answer? Believe in your abilities and strive for excellence. Your hard work will yield remarkable results. |
| EP_08 | Embrace challenges as opportunities for growth. Each obstacle you overcome brings you closer to success. |
| EP_09 | Stay focused and dedicated to your goals. Your consistent efforts will lead to outstanding achievements. |
| EP_10 | Take pride in your work and give it your best. Your commitment to excellence sets you apart. |
| EP_11 | Remember that progress is made one step at a time. Stay determined and keep moving forward. |

Table 7: Detailed Definitions of all 11 types of Emotion-Prompts.