

# Beyond Output Matching: Bidirectional Alignment for Enhanced In-Context Learning

Anonymous ACL submission

## Abstract

Large language models (LLMs) have shown impressive few-shot generalization on many tasks via in-context learning (ICL). Despite their success in showing such emergent abilities, the scale and complexity of larger models also lead to unprecedentedly high computational demands and deployment challenges. In reaction, researchers explore transferring the powerful capabilities of larger models to more efficient and compact models by typically aligning the *output* of smaller (student) models with that of larger (teacher) models. Existing methods either train student models on the generated outputs of teacher models or imitate their token-level probability distributions. However, these distillation methods pay little to no attention to the *input*, which also plays a crucial role in ICL. Based on the finding that the performance of ICL is highly sensitive to the selection of demonstration examples, we propose Bidirectional Alignment (BiAlign) to fully leverage the models’ preferences for ICL examples to improve the ICL abilities of student models. Specifically, we introduce the alignment of input preferences between student and teacher models by incorporating a novel ranking loss, in addition to aligning the token-level output distribution. With extensive experiments and analysis, we demonstrate that BiAlign can consistently outperform existing baselines on various tasks involving language understanding, reasoning, and coding.

## 1 Introduction

With the recent advancements in model scale and pretraining data, large language models (LLMs) have demonstrated impressive few-shot learning capabilities via in-context learning (ICL). With ICL, the LLM generates an output for a given query by conditioning on a few demonstration examples and optionally a task description, and it does so without any parameter updates (Brown et al., 2020). Despite the success of ICL in few-shot

generalization, the high computational demands and deployment challenges posed by the size of the LLMs hinder their widespread application. Serving an LLM with 175B parameters requires at least 350GB GPU memory (Hsieh et al., 2023), which is far beyond what is affordable in most real-world settings. Also, the serving cost increases with model size – it costs 1-2 FLOPs per parameter to infer on one token (Kaplan et al., 2020).

To alleviate this issue, researchers have proposed a number of methods to transfer the emergent capabilities of larger (teacher) models to more efficient and compact smaller (student) models, an approach commonly known as knowledge distillation (Hinton et al., 2015). In this approach, the student models are trained to align their *output* space with that of the teachers. This is typically achieved by either training on the generated outputs of the teacher models (Hsieh et al., 2023; Wang et al., 2022; Xu et al., 2023a) or by imitating their token-level probability distributions (Agarwal et al., 2023; Huang et al., 2023b; Gu et al., 2024).<sup>1</sup>

While existing distillation methods demonstrate improved ICL results, they pay little attention to the *input*, specifically the demonstrations, which have been shown to have a significant impact on the performance of ICL (Zhao et al., 2021; Xie et al., 2022; Qin et al., 2023). Indeed, selecting different sets of demonstration examples can yield performance ranging from almost random to better than state-of-the-art fine-tuned models (Gao et al., 2021; Lu et al., 2022), indicating that the model has different preferences for different inputs. Inspired by this finding, we propose **Bidirectional Alignment** (BiAlign), a simple yet effective framework for improving the ICL abilities

<sup>1</sup>Different from the conventional *strong-to-weak* generalization, Burns et al. (2023) recently introduce *weak-to-strong* generalization, which explores leveraging weaker (smaller) models to elicit “superalignment” from the stronger (larger) models. This paper however considers the conventional *strong-to-weak* approach.

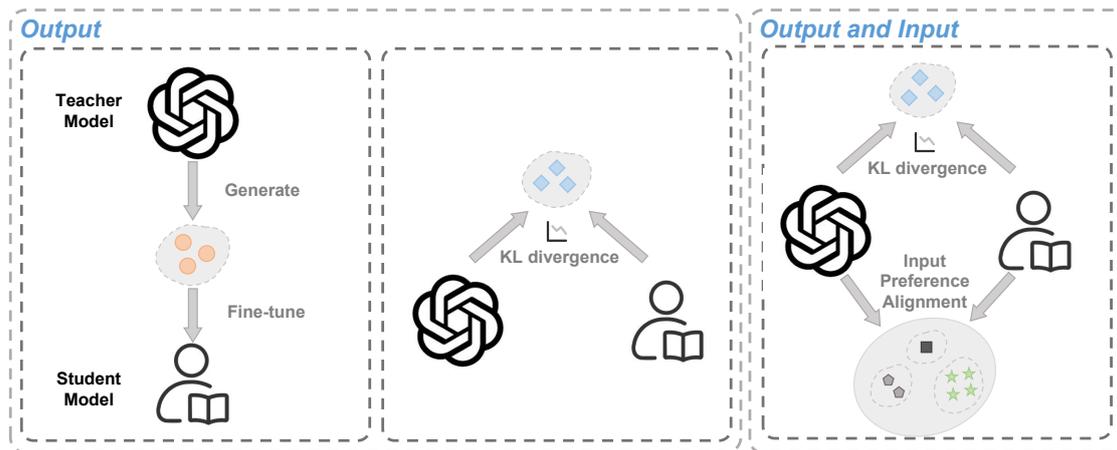


Figure 1: Comparison between different types of approaches to aligning student models. Existing methods typically fine-tune student models on generated outputs of teacher models or to match their token-level output probability distributions (*left part*). In contrast, our method (BiAlign) considers the models’ preferences for different inputs (the more helpful an input is for generating the target, the more the model prefers that input) to achieve input preference alignment (*right part*).

of student models (Figure 1). Specifically, BiAlign introduces the alignment of input preferences between student and teacher models through the incorporation of a novel ranking loss, in addition to aligning the token-level output distributions. Our main hypothesis is that for an effective knowledge distillation, the student model should align with not only the teacher model’s output distribution but also its input preference (i.e., the more helpful an input is for generating the target, the more the model prefers that input).<sup>2</sup> BiAlign allows student models to obtain more fine-grained supervision from teacher models by fully leveraging their preferences for different demonstrations in ICL. Empirical results on tasks spanning language understanding, symbolic reasoning, mathematical reasoning, logical reasoning, and coding show that BiAlign can consistently outperform previous baselines. In summary, our main contributions are:

- To the best of our knowledge, we for the first time consider aligning student models with teacher models from an *input preference* perspective. We propose Bidirectional Alignment (BiAlign) to fully leverage the models’ preferences for different demonstration examples to improve the ICL capabilities of student models.
- With extensive experiments and analysis, we demonstrate the effectiveness of BiAlign on a

<sup>2</sup>Our hypothesis is closely related to preference learning in RLHF, where the reward model learns ‘which outputs should be preferred’. After learning, a well-trained reward model can rank model responses with expertise comparable to humans.

variety of tasks. For example, it brings about 20% relative improvement on GSM8K (Cobbe et al., 2021) and 18% on LogiQA (Liu et al., 2020). Our code base is available at <redacted>.

## 2 Related Work

This work concerns how to improve the ICL ability of student models by aligning the student and teacher models’ preferences for different few-shot demonstrations. In light of this, we review three lines of work that form the basis of this work: few-shot learning, in-context learning, and alignment.

### 2.1 Few-shot Learning

Few-shot learning (FSL) aims to learn tasks with only a few labeled examples, which faces the challenge of over-fitting due to the scarcity of labeled training data. Existing methods to address this challenge can be mainly divided into three categories: (i) reducing the hypothesis space with prior knowledge (Triantafillou et al., 2017; Hu et al., 2018), (ii) optimizing the strategy for searching the best hypothesis in whole space (Ravi and Larochelle, 2017; Finn et al., 2017), and (iii) augmenting the few-shot data (Gao et al., 2020; Qin and Joty, 2022; Ding et al., 2023). More recently, LLMs have achieved promising performance on various few-shot tasks via in-context learning (ICL).

### 2.2 In-context Learning (ICL)

By conditioning on a prompt that includes several demonstration examples and optionally a task

description, a frozen LLM, by virtue of ICL, showcases impressive few-shot generalization (Brown et al., 2020). ICL has drawn a great deal of attention from the research community in recent days. Chen et al. (2022); Min et al. (2022a); Wei et al. (2023a) have explored ways to enhance the ICL capabilities of language models by either self-supervised or supervised training. In parallel, extensive analytical studies have been conducted to understand factors influencing the performance of ICL (Zhao et al., 2021; Wei et al., 2022a; Yoo et al., 2022; Min et al., 2022b; Wei et al., 2023b; Zhang et al., 2024), as well as to elucidate the underlying mechanisms that contribute to the success of ICL (Olsson et al., 2022; Xie et al., 2022; Pan, 2023; Li et al., 2023a; Dai et al., 2023). Furthermore, there is a series of ongoing research dedicated to various aspects of ICL: (i) demonstration designing strategies, including demonstration organization (Liu et al., 2022; Rubin et al., 2022; Wang et al., 2023b; Qin et al., 2023; Wang et al., 2024) and demonstration formatting (Wei et al., 2022c; Wang et al., 2022; Zhang et al., 2023; Zhou et al., 2023), (ii) multi-modal ICL (Huang et al., 2023a; Wang et al., 2023c,a; Zhu et al., 2023), and (iii) applications of ICL (Ding et al., 2022; Meade et al., 2023; Zheng et al., 2023; Long et al., 2024).

## 2.3 Alignment

Existing work on alignment can be mainly divided into two parts based on the objectives: aligning with humans and aligning with teacher models. To align with humans, reinforcement learning from human feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022) explores how human feedback can be used to train language models to better align with human preferences and values using reinforcement learning algorithms such as PPO (Schulman et al., 2017). Several recent studies have introduced lightweight alternatives of PPO, including RRHF (Yuan et al., 2023), DPO (Rafailov et al., 2023), ReMax (Li et al., 2023b), IPO (Azar et al., 2024) and KTO (Ethayarajh et al., 2024). Alignment with teacher models, also known as distillation (Hinton et al., 2015), aims to transfer the powerful capabilities of large teacher models to more efficient and compact student counterparts. It has emerged as a powerful solution to reduce the high computational demands and serving challenges posed by large models. Current distillation methods typically train student models on generated outputs of teacher models

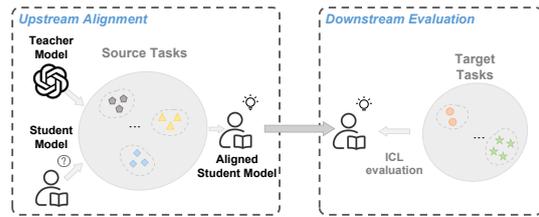


Figure 2: In the upstream ICL alignment stage, we align a student model with a teacher on the source tasks. Then in the downstream evaluation stage, we evaluate the ICL performance of the aligned student model on a held-out set of target tasks, which are different from the source tasks.

(Hsieh et al., 2023; Wang et al., 2022; Xu et al., 2023a) or to imitate teacher models’ token-level probability distributions (Sanh et al., 2019; Jiao et al., 2020; Agarwal et al., 2023; Huang et al., 2023b; Gu et al., 2024), i.e., these approaches focus on aligning the output of student models with that of teachers. However, they pay little attention to the input demonstrations which also significantly influence the performance of ICL (Qin et al., 2023). In contrast to these methods, our proposed method (BiAlign) leverages the models’ preferences for different in-context examples to achieve input preference alignment.

## 3 Methodology

### 3.1 Problem Setting

Given a training set  $\mathcal{D}_{\text{train}}$  consisting of a set of source tasks  $\mathcal{T}^{\text{src}}$ , the goal of ICL alignment is to align the ICL ability of a student model  $S$  with that of a teacher model  $T$ . Upon successful alignment, the model  $S$  is expected to show improved ICL ability on a held-out set of target tasks  $\mathcal{T}^{\text{tgt}}$ . We divide the whole process into two stages, as illustrated in Figure 2.

• **Upstream ICL alignment on  $\mathcal{T}^{\text{src}}$ :** In this alignment stage, the model has access to  $\mathcal{T}^{\text{src}}$ . We formalize samples in  $\mathcal{D}_{\text{train}}$  in the  $k$ -shot ICL format  $\{\hat{X}_i = (x_1, y_1), \dots, (x_k, y_k), (\hat{x}_i, \hat{y}_i)\}$ , where  $(x_j, y_j), 1 \leq j \leq k$  denotes the  $k$  demonstration examples and  $(\hat{x}_i, \hat{y}_i)$  is the test sample. We concatenate these examples to form an ICL training sample  $\hat{X}_i$ . We then align the student model  $S$  with the teacher model  $T$  on this formatted ICL data.

• **Downstream ICL evaluation on  $\mathcal{T}^{\text{tgt}}$ :** Following the upstream ICL alignment stage, we evaluate the ICL ability of the aligned model  $S^*$  on  $\mathcal{T}^{\text{tgt}}$ , where  $\mathcal{T}^{\text{tgt}}$  has no overlap with  $\mathcal{T}^{\text{src}}$ . For every

target task  $\mathcal{T}_k$ , we evaluate the model performance using both the default ICL demonstrations, as per convention, and their variants.

### 3.2 Bidirectional Alignment (BiAlign)

Based on the finding that the performance of ICL is highly sensitive to the selection of demonstration examples (Zhao et al., 2021), we propose Bidirectional Alignment (BiAlign) to fully leverage the models’ preferences for different demonstration examples with the goal of improving the ICL ability of the student model. Our approach is illustrated in Figure 3.

**Aligning Token-level Distributions** Given the ICL training examples in the concatenated form  $\{\hat{X}_i = (x_1, y_1), \dots, (x_k, y_k), (\hat{x}_i, \hat{y}_i)\}$  as discussed above, to achieve *token-level output distribution alignment* on  $\hat{X}_i$ , we minimize a KL divergence loss between the student model and teacher model for the *whole* sequence instead of only  $\hat{y}_i$  following Gu et al. (2023).<sup>3</sup> More formally,

$$\mathcal{L}^{\text{KL}} = \sum_{i=1}^m \sum_{j=1}^t D_{\text{KL}}(P_j(\mathcal{V}|\hat{X}_i, \theta_T) || P_j(\mathcal{V}|\hat{X}_i, \theta_S)) \quad (1)$$

where  $m$  is the number of ICL training samples in  $\mathcal{D}_{\text{train}}$ ,  $t$  is the number of tokens in  $\hat{X}_i$ ,  $\mathcal{V}$  is the models’ common vocabulary of tokens;  $\theta_T$  and  $\theta_S$  are the parameters of the teacher model and the student model, respectively.

**Aligning Preferences for Demonstrations** Intuitively, for the student and teacher models to be well-aligned, the demonstrations preferred by the teacher model should also be preferred by the student, i.e., to truly emulate the teacher model, the student needs to learn “what to output” as well as “which input demonstrations should be preferred” in order to generate high-quality outputs. This is similar in spirit to the scenario where a reward model is trained to align with preferences over model responses given by human experts (Ouyang et al., 2022). To this end, we introduce *input preference alignment* to align the student and teacher models’ preferences for different demonstrations.

For simplicity, let  $R_i = \{(x_1, y_1), \dots, (x_k, y_k)\}$  denote the  $k$ -shot demonstrations in each ICL training sample  $\hat{X}_i = (x_1, y_1), \dots, (x_k, y_k), (\hat{x}_i, \hat{y}_i)$ . To rank the model’s preferences for different

demonstration examples, we first need to obtain a set  $\mathcal{D}_{\text{rank}} = \{R_{ij}, (\hat{x}_i, \hat{y}_i)\}_{j=1}^N$ , where  $R_{ij}$  is a subset of  $R_i$  and  $N$  is the number of subsets considered for ranking. Modeling on the full subset space of  $R_i$  can be computationally prohibitive as it grows exponentially with  $|R_i|$ . Therefore, we set  $N \ll |\mathcal{P}(R_i)|$ , where  $\mathcal{P}(R_i)$  is the power set of  $R_i$ . Zhao et al. (2024) highlights the impact of similar examples in the demonstrations. Building on this insight, we categorize all demonstrations in  $R_i$  into two groups, namely  $G_{\text{sim}}$  and  $G_{\text{dissim}}$ , based on their similarity to the test example  $(\hat{x}_i, \hat{y}_i)$ . Subsequently, we sample  $N$  subsets from  $\mathcal{P}(R_i)$  with different numbers of similar examples.

We use both the student and teacher models to measure their preferences for each subset  $R_{ij}$ , which we estimate using the prediction probability of  $\hat{y}_i$  given  $R_{ij}$  and  $\hat{x}_i$  as input:<sup>4</sup>

$$Q^{\text{T}}(R_{ij}) = P(\hat{y}_i | R_{ij}, \hat{x}_i, \theta_T); Q^{\text{S}}(R_{ij}) = P(\hat{y}_i | R_{ij}, \hat{x}_i, \theta_S) \quad (2)$$

where  $Q^{\text{T}}$  and  $Q^{\text{S}}$  are the preference scores of the teacher and student models, respectively. Intuitively, the more helpful the subset  $R_{ij}$  is for generating the target  $\hat{y}_i$ , the more the model prefers this subset.

To align the preferences of the student and teacher models for different subsets, we introduce a novel ranking loss:

$$\begin{aligned} \mathcal{L}^{\text{rank}} = & \sum_{i=1}^m \sum_{R^+, R^- \in R_i^{\text{all}}} \max\{0, \\ & \underbrace{\frac{\log Q^{\text{S}}(R^-) - \log Q^{\text{S}}(R^+)}{\max_{R' \in R_i^{\text{all}}} \log Q^{\text{S}}(R') - \min_{R' \in R_i^{\text{all}}} \log Q^{\text{S}}(R')}}_{\text{Left}} \} \\ & + \underbrace{\frac{1}{N-1} (\text{rank}(Q^{\text{T}}(R^-)) - \text{rank}(Q^{\text{T}}(R^+)))}_{\text{Right}} \} \end{aligned} \quad (3)$$

where  $R_i^{\text{all}} = \{R_{ij}\}_{j=1}^N$  contains all subsets sampled for the test example  $(\hat{x}_i, \hat{y}_i)$ ,  $(R^+, R^-)$  refers to the pair of positive and negative subsets determined by the preference score of the teacher model (the subset with the higher preference score is considered as the positive one), and  $\text{rank}()$  stands for the function that measures the relative ranking of subset scores which ranges from 1 (most preferred) to  $N$  (least preferred). The left part of  $\mathcal{L}^{\text{rank}}$  measures the difference in preference scores

<sup>3</sup>Training on the whole sequence can ensure a large number of tokens in a batch, which is crucial to maintaining the basic in-weights capability (Chan et al., 2022).

<sup>4</sup>Under the assumption that the prior  $P(R_{ij}|\hat{x}_i, \theta)$  is uniform, it is easy to show using the Bayes rule:  $Q(R_{ij}) \propto P(R_{ij}|\hat{y}_i, \hat{x}_i, \theta) = \frac{P(\hat{y}_i|R_{ij}, \hat{x}_i, \theta)P(R_{ij}|\hat{x}_i, \theta)}{\sum_j P(\hat{y}_i|R_{ij}, \hat{x}_i, \theta)P(R_{ij}|\hat{x}_i, \theta)}$

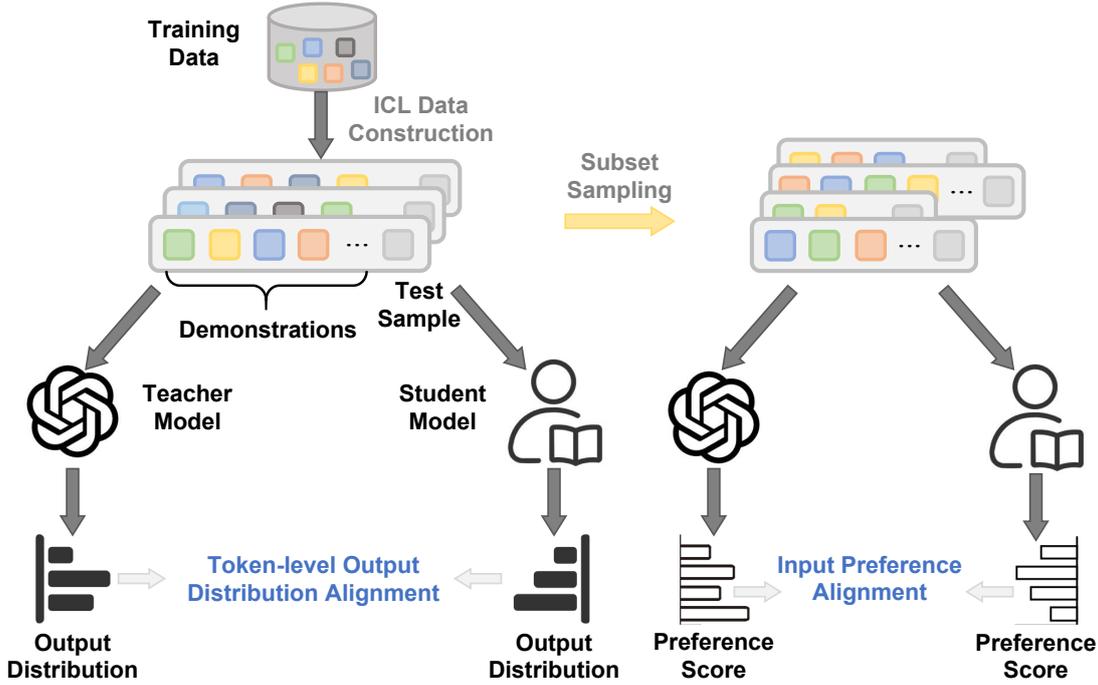


Figure 3: Illustration of our Bidirectional Alignment (BiAlign) framework. It attains *token-level output distribution alignment* by minimizing the KL divergence loss between the student and teacher models on the constructed ICL samples. Furthermore, after sampling several subsets from the set of all demonstrations, it optimizes a ranking loss for *input preference alignment* to align the student and teacher models’ preferences for different demonstration examples.

of the student model for the pair  $(R^+, R^-)$  and the right part reflects the relative ranking difference between  $R^+$  and  $R^-$  (see more analysis of  $\mathcal{L}^{\text{rank}}$  in Section 5.2). Therefore,  $\mathcal{L}^{\text{rank}}$  allows the student model to obtain more fine-grained supervision from the teacher model by *matching the relative ranking* of their preference scores for different demonstration examples in ICL.

The overall loss that BiAlign optimizes for alignment is:  $\mathcal{L} = \mathcal{L}^{\text{KL}} + \lambda \mathcal{L}^{\text{rank}}$ , where  $\lambda$  is the weight of the ranking loss. Besides, we illustrate the whole learning process in Appendix A.1.

## 4 Experimental Setup

In this section, we first describe the tasks and datasets, and then introduce methods compared in our work.

### 4.1 Tasks and Datasets

In this work, we use CrossFit (Ye et al., 2021), a large and diverse collection of few-shot tasks covering various types including classification, question answering and generation, as the source tasks  $\mathcal{T}^{\text{src}}$  (see Appendix A.2 for details of source tasks). For each task in CrossFit, we combine the

original training and validation data as the new training data which is then randomly partitioned into a set of ICL samples with  $4 \leq k \leq 10$  demonstration examples. For each ICL example, we sample  $N = 4$  subsets from the set of all demonstrations for calculating the ranking loss. After the preprocessing, we obtain 12K ICL examples in total.

We evaluate the ICL performance of the aligned model on 5 target tasks spanning language understanding, symbolic reasoning, mathematical reasoning, logical reasoning, and coding: MMLU (Hendrycks et al., 2021), BBH (Suzgun et al., 2022), GSM8K (Cobbe et al., 2021), LogiQA (Liu et al., 2020) and HumanEval (Chen et al., 2021). Note that there is no overlap between CrossFit and target tasks, and we obtain all outputs from the models using greedy decoding following Xu et al. (2023b). For each target task, we perform evaluations three times using different prompts and report the average results. Details of different target tasks and implementation are provided in Appendix A.3 and A.4, respectively.

### 4.2 Methods Compared

We mainly experiment with Llama 2-7B (Touvron

Method	MMLU	BBH	GSM8K	LogiQA	HumanEval	Average
<i>No Alignment Baselines</i>						
Vanilla	45.4 $\pm$ 0.6	39.5 $\pm$ 0.5	15.2 $\pm$ 0.3	30.3 $\pm$ 0.4	14.6 $\pm$ 0.4	29.0 $\pm$ 0.3
FT	46.4 $\pm$ 0.5	39.8 $\pm$ 0.5	15.6 $\pm$ 0.4	31.7 $\pm$ 0.3	14.2 $\pm$ 0.4	29.5 $\pm$ 0.4
C-Pretrain	46.0 $\pm$ 0.4	38.5 $\pm$ 0.6	15.9 $\pm$ 0.4	31.4 $\pm$ 0.4	13.4 $\pm$ 0.5	29.0 $\pm$ 0.4
Llama 2-13B Teacher						
Teacher	55.3 $\pm$ 0.5	47.8 $\pm$ 0.4	27.8 $\pm$ 0.3	37.8 $\pm$ 0.4	18.3 $\pm$ 0.3	37.4 $\pm$ 0.3
Output-Align	46.3 $\pm$ 0.4	39.3 $\pm$ 0.4	15.4 $\pm$ 0.2	32.2 $\pm$ 0.3	14.0 $\pm$ 0.2	29.4 $\pm$ 0.2
BiAlign	<b>47.5</b> $\pm$ 0.4	<b>41.0</b> $\pm$ 0.3	<b>16.8</b> $\pm$ 0.3	<b>33.9</b> $\pm$ 0.4	<b>15.6</b> $\pm$ 0.4	<b>31.0</b> $\pm$ 0.3
Llama 2-70B Teacher						
Teacher	67.2 $\pm$ 0.6	64.2 $\pm$ 0.4	53.3 $\pm$ 0.4	48.0 $\pm$ 0.5	26.8 $\pm$ 0.4	51.9 $\pm$ 0.4
Output-Align	47.1 $\pm$ 0.5	39.8 $\pm$ 0.4	16.4 $\pm$ 0.3	33.2 $\pm$ 0.3	14.6 $\pm$ 0.4	30.2 $\pm$ 0.3
BiAlign	<b>49.5</b> $\pm$ 0.3	<b>43.2</b> $\pm$ 0.5	<b>18.3</b> $\pm$ 0.4	<b>35.7</b> $\pm$ 0.4	<b>16.6</b> $\pm$ 0.3	<b>32.7</b> $\pm$ 0.3

Table 1: Performance (%) of different methods on 5 target tasks. We use Llama 2-7B as a student and Llama 2-13B or 70B as a teacher model. The rows with “Teacher” (grey) indicate the corresponding teacher model’s performance on the target tasks. **Bold** indicates the best result for Llama 2-7B (student). BiAlign is consistently better than all previous baselines.

et al., 2023) as the student model and Llama 2-13B or 70B as the teacher model. For Llama 2-70B, we use the quantized version TheBloke/Llama-2-70B-GPTQ (TheBloke, 2023) due to resource constraints. We compare BiAlign with the following methods:

- **Vanilla** simply evaluates the ICL performance of the student model on target tasks without any alignment, serving as the baseline for all other approaches.
- **Fine-tuning (FT)** tunes the student model on the 12K ICL examples constructed from CrossFit using a multi-task learning scheme, which is indeed the meta-training in Min et al. (2022a).
- **Continual Pretraining (C-Pretrain)** simply performs continual pretraining, *i.e.*, next token prediction for the whole sequence, of the student model on the 12K samples.
- **Output Alignment (Output-Align)** trains the student model to align token-level output distributions with the teacher model (Huang et al., 2023b; Gu et al., 2024).

We additionally show the connection between BiAlign and In-Context Pretraining (Shi et al., 2024) in Section 5.2.

## 5 Results and Analysis

### 5.1 Main Results

Table 1 shows the performance scores of different methods on all investigated target tasks. From the

	ASDiv	SVAMP	GSM8K	AQUA-RAT
Vanilla	46.6	41.2	15.2	24.4
BiAlign	<b>49.4</b>	<b>43.5</b>	<b>16.8</b>	<b>27.2</b>
Relative Gain	6.0	5.6	10.5	11.5

Table 2: Relative gain (%) of BiAlign on math reasoning tasks of varying difficulty levels.

results, we can observe that

- Our proposed BiAlign consistently outperforms baseline approaches on all datasets with different sizes of teacher models, demonstrating its superiority. Simply pretraining the model on source tasks does not improve the average performance since there is no overlap between source and target tasks. While fine-tuning brings marginal improvement, token-level output distribution alignment with a stronger (70B) teacher model can achieve better performance. Thanks to incorporating input preference alignment (see Section 5.2 for analysis of computational overhead), BiAlign yields about 2.0% performance boost on average when using a 13B teacher model, and this gain is 3.7% for a 70B teacher. Besides, when examining the effects of scaling up the teacher model, the performance of BiAlign sees an improvement on all tasks.
- In particular, BiAlign using a 13B teacher model achieves relative performance improvements of 11.9% on LogiQA and 10.5% on GSM8K compared with Vanilla, while using the 70B teacher, it achieves 17.8% on LogiQA and 20.4% on GSM8K. These results indicate that BiAlign can

Method	7B	13B
Output-Align	30.2	38.8
BiAlign	<b>32.7</b>	<b>40.9</b>

Table 3: Average results (%) of Output-Align and BiAlign with different sizes of student models (Llama 2-70B as the teacher).

Method	Vanilla	FT	C-Pretrain	Output-Align	BiAlign
Llama 3-8B	60.4	61.0	60.5	61.7	<b>63.9</b>
Phi-3-mini (3.8B)	66.7	67.1	66.5	67.4	<b>69.1</b>

Table 4: Average results (%) across 5 tasks of all methods with two different backbones. We use Llama 3-70B as the teacher for Llama 3-8B and Phi-3-medium (14B) as the teacher for Phi-3-mini (3.8B).

410 better improve the performance of tasks requiring  
 411 more fine-grained reasoning; see appendix A.14 for  
 412 an example in LogiQA. This is because BiAlign  
 413 allows the student model to obtain more fine-  
 414 grained supervision from the teacher model by fully  
 415 leveraging their preferences for different inputs.

416 To better support our claim, we further conduct  
 417 experiments on four mathematical reasoning tasks  
 418 ranging from low to high difficulty: ASDiv (Miao  
 419 et al., 2020), SVAMP (Patel et al., 2021), GSM8K  
 420 (Cobbe et al., 2021), and AQUA-RAT (Ling et al.,  
 421 2017a). The comparison between BiAlign and  
 422 Vanilla, as illustrated in Table 2, demonstrates  
 423 that BiAlign is indeed more beneficial for more  
 424 complex reasoning tasks.

425 • Both fine-tuning and output alignment  
 426 sometimes hurt the zero-shot learning capability  
 427 of the model as shown by the performance on  
 428 HumanEval. In contrast, BiAlign brings an  
 429 average relative improvement of about 10.3% on  
 430 HumanEval. We speculate that this is due to the  
 431 subset sampling in input preference alignment,  
 432 which helps the model generalize better to the  
 433 unseen zero-shot setting.

## 434 5.2 Analysis

435 **Larger Student Model** We further experiment  
 436 with a larger student model to verify the  
 437 effectiveness of BiAlign. Specifically, we use  
 438 Llama 2-13B as the student model and Llama  
 439 2-70B as the teacher model. We employ  
 440 QLoRA (Dettmers et al., 2023) to fine-tune the  
 441 student model with consideration of computational  
 442 resource limitations. The results averaged over the  
 443 5 tasks are reported in Table 3, which demonstrate  
 444 the consistent superiority of BiAlign across model

	Default	Variant
BiAlign	<b>31.0</b>	30.5

Table 5: Average results (%) of BiAlign with different ranking loss formulations.

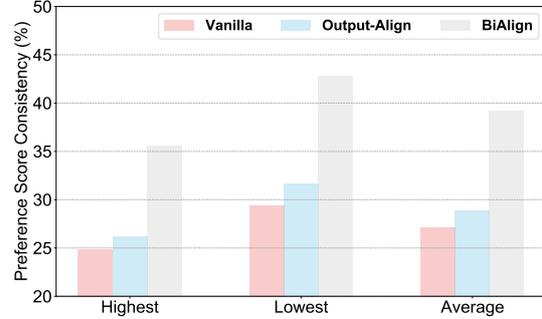


Figure 4: Preference score consistency (%) of different methods.

445 scales.

446 **Different Backbone Models** Our experiments  
 447 and analysis so far use Llama 2 as the backbone  
 448 model. To verify whether the performance gain  
 449 of BiAlign is consistent across different backbone  
 450 models, we extend the experiments to Llama 3  
 451 (Dubey et al., 2024) and Phi 3 (Abdin et al., 2024).  
 452 For Llama 3, we use the 8B model as the student  
 453 and the 70B model as the teacher. For Phi 3, we use  
 454 Phi-3-mini (3.8B) as the student and Phi-3-medium  
 455 (14B) as the teacher. From the average results  
 456 shown in Table 4, we can see that BiAlign still  
 457 outperforms all baseline approaches when using  
 458 other language models as the backbone, showing  
 459 its robustness to model types.

460 **Comment on Training-time Computational**  
 461 **Overhead** Smaller models are a preferred choice  
 462 for resource-constrained deployments, where the  
 463 inference cost matters the most. BiAlign does  
 464 not introduce any additional cost during inference.  
 465 The additional computational overhead only occurs  
 466 once during model training. To quantify the  
 467 increase in computational overhead caused by the  
 468 ranking loss, we use DeepSpeed Flops Profiler  
 469 (Rasley et al., 2020) to calculate the training FLOPs  
 470 of Output-Align and BiAlign, which are  $3.3 \times 10^{17}$   
 471 and  $7.6 \times 10^{17}$  respectively (about 2.3 times).  
 472 Therefore, we further design two experiments to  
 473 compare BiAlign and Output-Align under the same  
 474 training FLOPs: (i) we combine the original ICL  
 475 training examples with the sampled subset data  
 476 and conduct Output-Align on the combined data

(roughly the same FLOPs as BiAlign), which performs (29.5) similarly to the original Output-Align method (29.4), verifying the superiority of BiAlign; (ii) we reduce the training epochs of BiAlign from 4 to 2 (roughly the same FLOPs as Output-Align) and assess the final checkpoint. There is no significant performance degradation (from 31.0 to 30.8), which also demonstrates that BiAlign can outperform baselines under the same training FLOPs.

**Different Ranking Loss Formulations** In the right part of Equation 3, we employ the  $\text{rank}()$  function to represent the relative ranking of the model’s preference scores instead of relying on the scores themselves. This approach is grounded in the idea that the primary objective of input preference alignment is to match the rankings of the subset scores, rather than their specific values. By focusing on rankings, we can reduce the impact of potential variations in score magnitudes, allowing the model to prioritize the relative ranking of preferences. We further conduct experiments with an alternative ranking loss formulation that does not incorporate  $\text{rank}()$ , while maintaining all other implementation details. The average results reported in Table 5 underscore the importance of using  $\text{rank}()$  for alignment.

**Connection with In-Context Pretraining** Shi et al. (2024) propose In-Context Pretraining (ICP) which pretrains language models on a sequence of related documents. BiAlign mainly differs from it in the following two aspects: (i) ICP focuses on the pretraining stage while BiAlign is specifically designed for more lightweight supervised fine-tuning. (ii) The objective of ICP is to design more effective pretraining data. In contrast, BiAlign leverages distillation to improve the capabilities of the student model. Therefore, BiAlign can be seamlessly integrated with ICP to further improve the ICL ability.

**Effect of Demonstration Numbers** As mentioned in Section 4.1, each constructed ICL training sample contains  $4 \leq k \leq 10$  demonstration examples, which could enhance the model’s ability to generalize to different numbers of demonstrations. To investigate the effect of demonstration numbers in source tasks, we further conduct training on examples containing a fixed number  $k \in \{5, 8, 10\}$  of demonstrations. The average results of the 5 target tasks are reported

Method	Demonstration number			
	Default ( $4 \leq k \leq 10$ )	5	8	10
BiAlign	<b>31.0</b>	30.8	30.4	30.5

Table 6: Average results (%) of BiAlign with different  $k$  (demonstration number) for constructed ICL training samples.

in Table 6. We can see that training with a fixed number of demonstrations results in performance degradation to a certain degree, justifying our training set construction strategy.

**Preference Score Consistency** As illustrated in Section 3.2,  $\mathcal{L}^{\text{rank}}$  enables the student model to match the relative ranking of the preference scores for different ICL demonstrations with that of the teacher model. To verify this, we report the *preference score consistency* comparison between BiAlign and Output-Align in Figure 4. Specifically, we randomly select 500 examples from MMLU. For each example, we randomly sample 5 subsets from the set of all demonstrations and obtain their preference scores using different models. The preference score consistency of different methods is then calculated as the proportion of the highest/lowest scoring subsets that are consistent between the corresponding student model and the teacher model. From the results, we can see that BiAlign can indeed achieve much higher preference score consistency than Output-Align, indicating the effectiveness of  $\mathcal{L}^{\text{rank}}$ .

In addition, for interested readers, we show the results with different subset sampling methods, different numbers of subsets and different source task selections, the analysis of KL divergence calculation, training steps and additional training data, the influence of ranking loss weight, the effect of contrastive pair selection, and a case study of model output in Appendix A.5 ~ A.13, respectively.

## 6 Conclusion

In this work, we have introduced Bidirectional Alignment (BiAlign) that can improve the ICL capabilities of student models by aligning the input preferences between student and teacher models in addition to aligning the token-level output distributions. Extensive experimental results and analysis show that BiAlign consistently outperforms previous baseline approaches.

## 568 Limitations

569 As the first work on input preference alignment,  
570 one limitation of our paper is the additional  
571 computational overhead introduced by the ranking  
572 loss. A further improvement could be to explore  
573 more efficient input alignment methods to improve  
574 the ICL capabilities of student models.

## 575 References

576 Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan,  
577 Jyoti Aneja, Ahmed Awadallah, Hany Awadalla,  
578 Nguyen Bach, Amit Bahree, Arash Bakhtiari,  
579 Harkirat Behl, et al. 2024. Phi-3 technical report:  
580 A highly capable language model locally on your  
581 phone. *arXiv preprint arXiv:2404.14219*.

582 Rishabh Agarwal, Nino Vieillard, Piotr Stanczyk,  
583 Sabela Ramos, Matthieu Geist, and Olivier Bachem.  
584 2023. Gkd: Generalized knowledge distillation for  
585 auto-regressive sequence models. *arXiv preprint*  
586 *arXiv:2306.13649*.

587 Tiago A. Almeida, José María G. Hidalgo, and Akebo  
588 Yamakami. 2011. [Contributions to the study of](#)  
589 [sms spam filtering: New collection and results](#).  
590 In *Proceedings of the 11th ACM Symposium on*  
591 *Document Engineering, DocEng '11*, page 259–262,  
592 New York, NY, USA. Association for Computing  
593 Machinery.

594 Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik  
595 Koncel-Kedziorski, Yejin Choi, and Hannaneh  
596 Hajishirzi. 2019. [MathQA: Towards interpretable](#)  
597 [math word problem solving with operation-](#)  
598 [based formalisms](#). In *Proceedings of the 2019*  
599 *Conference of the North American Chapter of the*  
600 *Association for Computational Linguistics: Human*  
601 *Language Technologies, Volume 1 (Long and Short*  
602 *Papers)*, pages 2357–2367, Minneapolis, Minnesota.  
603 Association for Computational Linguistics.

604 Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo,  
605 Bilal Piot, Remi Munos, Mark Rowland, Michal  
606 Valko, and Daniele Calandriello. 2024. A general  
607 theoretical paradigm to understand learning from  
608 human preferences. In *International Conference*  
609 *on Artificial Intelligence and Statistics*, pages 4447–  
610 4455. PMLR.

611 Francesco Barbieri, Jose Camacho-Collados, Luis  
612 Espinosa Anke, and Leonardo Neves. 2020.  
613 [TweetEval: Unified benchmark and comparative](#)  
614 [evaluation for tweet classification](#). In *Findings of the*  
615 *Association for Computational Linguistics: EMNLP*  
616 *2020*, pages 1644–1650, Online. Association for  
617 Computational Linguistics.

618 Max Bartolo, Alastair Roberts, Johannes Welbl,  
619 Sebastian Riedel, and Pontus Stenetorp. 2020. [Beat](#)  
620 [the AI: Investigating adversarial human annotation](#)  
621 [for reading comprehension](#). *Transactions of the*

*Association for Computational Linguistics*, 8:662–  
678. 622 623

Jonathan Berant, Andrew Chou, Roy Frostig, and  
624 Percy Liang. 2013. [Semantic parsing on Freebase](#)  
625 [from question-answer pairs](#). In *Proceedings of the*  
626 *2013 Conference on Empirical Methods in Natural*  
627 *Language Processing*, pages 1533–1544, Seattle,  
628 Washington, USA. Association for Computational  
629 Linguistics. 630

Chandra Bhagavatula, Ronan Le Bras, Chaitanya  
631 Malaviya, Keisuke Sakaguchi, Ari Holtzman,  
632 Hannah Rashkin, Doug Downey, Wen tau Yih,  
633 and Yejin Choi. 2020. [Abductive commonsense](#)  
634 [reasoning](#). In *International Conference on Learning*  
635 *Representations*. 636

Yonatan Bisk, Rowan Zellers, Ronan LeBras,  
637 Jianfeng Gao, and Yejin Choi. 2020. [PIQA:](#)  
638 [reasoning about physical commonsense in natural](#)  
639 [language](#). In *The Thirty-Fourth AAAI Conference on*  
640 *Artificial Intelligence, AAAI 2020, The Thirty-Second*  
641 *Innovative Applications of Artificial Intelligence*  
642 *Conference, IAAI 2020, The Tenth AAAI Symposium*  
643 *on Educational Advances in Artificial Intelligence,*  
644 *EAAI 2020, New York, NY, USA, February 7-12, 2020,*  
645 *pages 7432–7439*. AAAI Press. 646

Michael Boratko, Xiang Li, Tim O’Gorman, Rajarshi  
647 Das, Dan Le, and Andrew McCallum. 2020. [Proto-](#)  
648 [QA: A question answering dataset for prototypical](#)  
649 [common-sense reasoning](#). In *Proceedings of the*  
650 *2020 Conference on Empirical Methods in Natural*  
651 *Language Processing (EMNLP)*, pages 1122–1136,  
652 Online. Association for Computational Linguistics. 653

Jan A. Botha, Manaal Faruqui, John Alex, Jason  
654 Baldrige, and Dipanjan Das. 2018. [Learning to](#)  
655 [split and rephrase from Wikipedia edit history](#). In  
656 *Proceedings of the 2018 Conference on Empirical*  
657 *Methods in Natural Language Processing*, pages  
658 732–737, Brussels, Belgium. Association for  
659 Computational Linguistics. 660

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie  
661 Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind  
662 Neelakantan, Pranav Shyam, Girish Sastry, Amanda  
663 Askell, et al. 2020. Language models are few-shot  
664 learners. *arXiv preprint arXiv:2005.14165*. 665

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner,  
666 Bowen Baker, Leo Gao, Leopold Aschenbrenner,  
667 Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan  
668 Leike, et al. 2023. [Weak-to-strong generalization:](#)  
669 [Eliciting strong capabilities with weak supervision:](#)  
670 [arXiv preprint arXiv:2312.09390](#). 671

Stephanie C.Y. Chan, Adam Santoro, Andrew Kyle  
672 Lampinen, Jane X Wang, Aaditya K Singh,  
673 Pierre Harvey Richemond, James McClelland, and  
674 Felix Hill. 2022. [Data distributional properties drive](#)  
675 [emergent in-context learning in transformers](#). In  
676 *Advances in Neural Information Processing Systems*. 677

678	Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. <a href="#">SemEval-2019 task 3: EmoContext contextual emotion detection in text</a> . In <i>Proceedings of the 13th International Workshop on Semantic Evaluation</i> , pages 39–48, Minneapolis, Minnesota, USA. Association for Computational Linguistics.	736
679		737
680		738
681		
682		739
683		740
684		741
685	Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. <i>arXiv preprint arXiv:2107.03374</i> .	742
686		743
687		744
688		745
689		746
690		747
691	Michael Chen, Mike D’Arcy, Alisa Liu, Jared Fernandez, and Doug Downey. 2019. <a href="#">CODAH: An adversarially-authored question answering dataset for common sense</a> . In <i>Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP</i> , pages 63–69, Minneapolis, USA. Association for Computational Linguistics.	748
692		749
693		750
694		751
695		
696		
697		
698	Mingda Chen, Jingfei Du, Ramakanth Pasunuru, Todor Mihaylov, Sridi Iyer, Veselin Stoyanov, and Zornitsa Kozareva. 2022. <a href="#">Improving in-context few-shot learning via self-supervised training</a> . In <i>Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies</i> , pages 3558–3573, Seattle, United States. Association for Computational Linguistics.	752
699		753
700		754
701		755
702		756
703		757
704		
705		
706		
707	Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyu Zhou, and William Yang Wang. 2020. <a href="#">Tabfact: A large-scale dataset for table-based fact verification</a> . In <i>8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020</i> . OpenReview.net.	758
708		759
709		760
710		761
711		762
712		763
713		764
714	Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. <a href="#">Deep reinforcement learning from human preferences</a> . <i>Advances in neural information processing systems</i> , 30.	765
715		766
716		
717		
718		
719	Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. <a href="#">BoolQ: Exploring the surprising difficulty of natural yes/no questions</a> . In <i>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)</i> , pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.	767
720		768
721		769
722		770
723		771
724		772
725		
726		
727		
728		
729	Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. <a href="#">Think you have solved question answering? try arc, the ai2 reasoning challenge</a> . <i>ArXiv preprint</i> , abs/1803.05457.	767
730		768
731		769
732		770
733		771
734	Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias	772
735		773
		774
		775
		776
		777
		778
		779
		780
		781
		782
		783
		784
		785
		786
		787
		788
		789
		790
		791
	Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. <i>arXiv preprint arXiv:2110.14168</i> .	
	Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. <a href="#">Structural scaffolds for citation intent classification in scientific publications</a> . In <i>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)</i> , pages 3586–3596, Minneapolis, Minnesota. Association for Computational Linguistics.	
	Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In <i>Machine Learning Challenges Workshop</i> , pages 177–190. Springer.	
	Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2023. Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers. In <i>ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models</i> .	
	Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner. 2019. <a href="#">Quoref: A reading comprehension dataset with questions requiring coreferential reasoning</a> . In <i>Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)</i> , pages 5925–5932, Hong Kong, China. Association for Computational Linguistics.	
	Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In <i>Proceedings of the 11th International AAAI Conference on Web and Social Media, ICWSM ’17</i> , pages 512–515.	
	Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. <a href="#">Hate speech dataset from a white supremacy forum</a> . In <i>Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)</i> , pages 11–20, Brussels, Belgium. Association for Computational Linguistics.	
	Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2019. <a href="#">The commitmentbank: Investigating projection in naturally occurring discourse</a> . <i>Proceedings of Sinn und Bedeutung</i> , 23(2):107–124.	
	Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. <i>arXiv preprint arXiv:2305.14314</i> .	
	T. Diggelmann, Jordan L. Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. <a href="#">Climate-fever: A dataset for verification of real-world climate claims</a> . <i>ArXiv preprint</i> , abs/2012.00614.	

792	Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. <a href="#">Wizard of wikipedia: Knowledge-powered conversational agents</a> . In <i>7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019</i> . OpenReview.net.	Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. <a href="#">Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model</a> . In <i>Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics</i> , pages 1074–1084, Florence, Italy. Association for Computational Linguistics.	846 847 848 849 850 851 852 853
798	Bosheng Ding, Chengwei Qin, Linlin Liu, Lidong Bing, Shafiq Joty, and Boyang Li. 2022. <a href="#">Is gpt-3 a good data annotator?</a> <i>arXiv preprint arXiv:2212.10450</i> .	Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. <a href="#">ELI5: Long form question answering</a> . In <i>Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics</i> , pages 3558–3567, Florence, Italy. Association for Computational Linguistics.	854 855 856 857 858 859 860
801	Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. <a href="#">Is GPT-3 a good data annotator?</a> In <i>Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 11173–11195, Toronto, Canada. Association for Computational Linguistics.	Manaaf Faruqui and Dipanjan Das. 2018. <a href="#">Identifying well-formed natural language questions</a> . In <i>Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing</i> , pages 798–803, Brussels, Belgium. Association for Computational Linguistics.	861 862 863 864 865 866
802	William B. Dolan and Chris Brockett. 2005. <a href="#">Automatically constructing a corpus of sentential paraphrases</a> . In <i>Proceedings of the Third International Workshop on Paraphrasing (IWP2005)</i> .	Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. <a href="#">Model-agnostic meta-learning for fast adaptation of deep networks</a> . In <i>Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017</i> , volume 70 of <i>Proceedings of Machine Learning Research</i> , pages 1126–1135. PMLR.	867 868 869 870 871 872 873
803	Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. <a href="#">The llama 3 herd of models</a> . <i>arXiv preprint arXiv:2407.21783</i> .	Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. <a href="#">Making pre-trained language models better few-shot learners</a> . In <i>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)</i> , pages 3816–3830, Online. Association for Computational Linguistics.	874 875 876 877 878 879 880 881
804	Matthew Dunn, Levent Sagun, Mike Higgins, V. U. Güney, Volkan Cirik, and Kyunghyun Cho. 2017. <a href="#">Searchqa: A new q&amp;a dataset augmented with context from a search engine</a> . <i>ArXiv preprint</i> , abs/1704.05179.	Tianyu Gao, Xu Han, Ruobing Xie, Zhiyuan Liu, Fen Lin, Leyu Lin, and Maosong Sun. 2020. <a href="#">Neural snowball for few-shot relation learning</a> . In <i>The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020</i> , pages 7772–7779. AAAI Press.	882 883 884 885 886 887 888 889 890 891
805	Ondřej Dušek, David M. Howcroft, and Verena Rieser. 2019. <a href="#">Semantic noise matters for neural natural language generation</a> . In <i>Proceedings of the 12th International Conference on Natural Language Generation</i> , pages 421–426, Tokyo, Japan. Association for Computational Linguistics.	Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. <a href="#">SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization</a> . In <i>Proceedings of the 2nd Workshop on New Frontiers in Summarization</i> , pages 70–79, Hong Kong, China. Association for Computational Linguistics.	892 893 894 895 896 897 898
806	Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2020. <a href="#">Evaluating the State-of-the-Art of End-to-End Natural Language Generation: The E2E NLG Challenge</a> . <i>Computer Speech &amp; Language</i> , 59:123–156.	Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. <a href="#">SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning</a> . In <i>*SEM 2012: The First Joint Conference on Lexical and Computational</i>	899 900 901 902 903
807	Hady Elsahar, Pavlos Vougiouklis, Arslan Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. <a href="#">T-REX: A large scale alignment of natural language with knowledge base triples</a> . In <i>Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)</i> , Miyazaki, Japan. European Language Resources Association (ELRA).		
808	Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. <a href="#">Kto: Model alignment as prospect theoretic optimization</a> . <i>arXiv preprint arXiv:2402.01306</i> .		

904	<i>Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)</i> , pages 394–398, Montréal, Canada. Association for Computational Linguistics.	
905		
906		
907		
908		
909		
910	Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2023. <a href="#">Pre-training to learn in context</a> . In <i>Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 4849–4870, Toronto, Canada. Association for Computational Linguistics.	
911		
912		
913		
914		
915		
916	Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. <a href="#">MiniLLM: Knowledge distillation of large language models</a> . In <i>The Twelfth International Conference on Learning Representations</i> .	
917		
918		
919		
920	Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. <a href="#">Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports</a> . <i>Journal of Biomedical Informatics</i> , 45(5):885–892.	
921		
922		
923		
924		
925		
926		
927		
928	Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. <a href="#">Question-answer driven semantic role labeling: Using natural language to annotate natural language</a> . In <i>Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing</i> , pages 643–653, Lisbon, Portugal. Association for Computational Linguistics.	
929		
930		
931		
932		
933		
934		
935	Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. <a href="#">Measuring massive multitask language understanding</a> . In <i>International Conference on Learning Representations</i> .	
936		
937		
938		
939		
940	Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. <a href="#">Distilling the knowledge in a neural network</a> . <i>arXiv preprint arXiv:1503.02531</i> .	
941		
942		
943	Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. <a href="#">Robust disambiguation of named entities in text</a> . In <i>Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing</i> , pages 782–792, Edinburgh, Scotland, UK. Association for Computational Linguistics.	
944		
945		
946		
947		
948		
949		
950		
951	Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Chin-Yew Lin, and Deepak Ravichandran. 2001. <a href="#">Toward semantics-based answer pinpointing</a> . In <i>Proceedings of the First International Conference on Human Language Technology Research</i> .	
952		
953		
954		
955		
956	Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. <a href="#">Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes</a> . In <i>Findings of the Association for Computational Linguistics: ACL 2023</i> , pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.	961
960		962
		963
		964
	Zikun Hu, Xiang Li, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2018. <a href="#">Few-shot charge prediction with discriminative legal attributes</a> . In <i>Proceedings of the 27th International Conference on Computational Linguistics</i> , pages 487–498, Santa Fe, New Mexico, USA. Association for Computational Linguistics.	965
		966
		967
		968
		969
		970
	Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. <a href="#">Cosmos QA: Machine reading comprehension with contextual commonsense reasoning</a> . In <i>Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)</i> , pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.	971
		972
		973
		974
		975
		976
		977
		978
		979
	Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. 2023a. <a href="#">Language is not all you need: Aligning perception with language models</a> . <i>arXiv preprint arXiv:2302.14045</i> .	980
		981
		982
		983
		984
		985
	Yukun Huang, Yanda Chen, Zhou Yu, and Kathleen McKeown. 2023b. <a href="#">In-context learning distillation: Transferring few-shot learning ability of pre-trained language models</a> . <i>arXiv preprint</i> .	986
		987
		988
		989
	Kelvin Jiang, Dekun Wu, and Hui Jiang. 2019. <a href="#">FreebaseQA: A new factoid QA data set matching trivia-style question-answer pairs with Freebase</a> . In <i>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)</i> , pages 318–323, Minneapolis, Minnesota. Association for Computational Linguistics.	990
		991
		992
		993
		994
		995
		996
		997
		998
	Fangkai Jiao, Zhiyang Teng, Shafiq Joty, Bosheng Ding, Aixin Sun, Zhengyuan Liu, and Nancy F Chen. 2023. <a href="#">Logicllm: Exploring self-supervised logic-enhanced training for large language models</a> . <i>arXiv preprint arXiv:2305.13718</i> .	999
		1000
		1001
		1002
		1003
	Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. <a href="#">TinyBERT: Distilling BERT for natural language understanding</a> . In <i>Findings of the Association for Computational Linguistics: EMNLP 2020</i> , pages 4163–4174, Online. Association for Computational Linguistics.	1004
		1005
		1006
		1007
		1008
		1009
		1010
	Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. <a href="#">Scaling laws for neural language models</a> . <i>arXiv preprint arXiv:2001.08361</i> .	1011
		1012
		1013
		1014
		1015
	Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018.	1016
		1017

1018	Looking beyond the surface: A challenge set for reading comprehension over multiple sentences.	in <i>Natural Language Processing</i> , pages 1203–1213, Austin, Texas. Association for Computational Linguistics.	1076
1019			1077
1020			1078
1021			
1022		Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, D. Kontokostas, Pablo N. Mendes, Sebastian Hellmann, M. Morsey, Patrick van Kleef, S. Auer, and C. Bizer. 2015. Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia. <i>Semantic Web</i> , 6:167–195.	1079
1023			1080
1024			1081
1025			1082
1026			1083
1027	Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. <i>Proceedings of the AAAI Conference on Artificial Intelligence</i> , 34(05):8082–8090.		1084
1028			1085
1029			1086
1030			1087
1031			1088
1032			1089
1033	Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In <i>Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018</i> , pages 5189–5197. AAAI Press.		1090
1034			1091
1035			1092
1036			1093
1037			1094
1038			1095
1039			
1040			
1041			1096
1042			1097
1043			1098
1044	Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2019. Abstractive summarization of Reddit posts with multi-level memory networks. In <i>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)</i> , pages 2519–2531, Minneapolis, Minnesota. Association for Computational Linguistics.		1099
1045			1100
1046			1101
1047			1102
1048			
1049			
1050			1103
1051			1104
1052			1105
1053			1106
1054			
1055			
1056			1107
1057			1108
1058			1109
1059			1110
1060			1111
1061			1112
1062			1113
1063			
1064			
1065			1114
1066			1115
1067			1116
1068			1117
1069			1118
1070			1119
1071			1120
1072			1121
1073			
1074			
1075			
			1122
			1123
			1124
			1125
			1126
			1127
			1128
			1129
			1130
			1131



1246		<i>Human Language Technologies</i> , pages 2791–2809,	instructions with human feedback. <i>arXiv preprint</i>	1303
1247		Seattle, United States. Association for Computational	<i>arXiv:2203.02155</i> .	1304
1248		Linguistics.		
1249	Sewon Min, Xinxu Lyu, Ari Holtzman, Mikel		Jane Pan. 2023. <i>What In-Context Learning “Learns” In-</i>	1305
1250	Artetxe, Mike Lewis, Hannaneh Hajishirzi, and		<i>Context: Disentangling Task Recognition and Task</i>	1306
1251	Luke Zettlemoyer. 2022b. <i>Rethinking the role of</i>		<i>Learning</i> . Ph.D. thesis, Princeton University.	1307
1252	demonstrations: What makes in-context learning			
1253	work? In <i>Proceedings of the 2022 Conference on</i>		Bo Pang and Lillian Lee. 2005. <i>Seeing stars: Exploiting</i>	1308
1254	<i>Empirical Methods in Natural Language Processing</i> ,		<i>class relationships for sentiment categorization</i>	1309
1255	pages 11048–11064, Abu Dhabi, United Arab		<i>with respect to rating scales</i> . In <i>Proceedings of</i>	1310
1256	Emirates. Association for Computational Linguistics.		<i>the 43rd Annual Meeting of the Association for</i>	1311
			<i>Computational Linguistics (ACL’05)</i> , pages 115–124,	1312
1257	Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos,		Ann Arbor, Michigan. Association for Computational	1313
1258	and Grigorios Tsoumakos. 2020. <i>Ethos: an online</i>		Linguistics.	1314
1259	<i>hate speech detection dataset</i> . <i>ArXiv preprint</i> ,			
1260	abs/2006.08328.		Dimitris Pappas, Petros Stavropoulos, Ion Androu-	1315
			sopoulos, and Ryan McDonald. 2020. <i>BioMRC: A</i>	1316
1261	Nikita Nangia, Clara Vania, Rasika Bhalerao, and		<i>dataset for biomedical machine reading comprehen-</i>	1317
1262	Samuel R. Bowman. 2020. <i>CrowS-pairs: A</i>		<i>sion</i> . In <i>Proceedings of the 19th SIGBioMed Work-</i>	1318
1263	<i>challenge dataset for measuring social biases in</i>		<i>shop on Biomedical Language Processing</i> , pages	1319
1264	<i>masked language models</i> . In <i>Proceedings of the</i>		140–149, Online. Association for Computational	1320
1265	<i>2020 Conference on Empirical Methods in Natural</i>		Linguistics.	1321
1266	<i>Language Processing (EMNLP)</i> , pages 1953–1967,			
1267	Online. Association for Computational Linguistics.		Arkil Patel, Satwik Bhattamishra, and Navin Goyal.	1322
			2021. <i>Are NLP models really able to solve simple</i>	1323
1268	Courtney Napoles, Matthew Gormley, and Benjamin		<i>math word problems?</i> In <i>Proceedings of the 2021</i>	1324
1269	Van Durme. 2012. <i>Annotated Gigaword</i> . In		<i>Conference of the North American Chapter of the</i>	1325
1270	<i>Proceedings of the Joint Workshop on Automatic</i>		<i>Association for Computational Linguistics: Human</i>	1326
1271	<i>Knowledge Base Construction and Web-scale Knowl-</i>		<i>Language Technologies</i> , pages 2080–2094, Online.	1327
1272	<i>edge Extraction (AKBC-WEKEX)</i> , pages 95–100,		Association for Computational Linguistics.	1328
1273	Montréal, Canada. Association for Computational			
1274	Linguistics.		Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim	1329
			Rocktäschel, Yuxiang Wu, Alexander H. Miller,	1330
1275	Shashi Narayan, Shay B. Cohen, and Mirella Lapata.		and Sebastian Riedel. 2020. <i>How context affects</i>	1331
1276	2018. <i>Don’t give me the details, just the</i>		<i>language models’ factual predictions</i> . In <i>Automated</i>	1332
1277	<i>summary! topic-aware convolutional neural networks</i>		<i>Knowledge Base Construction</i> .	1333
1278	<i>for extreme summarization</i> . In <i>Proceedings of the</i>			
1279	<i>2018 Conference on Empirical Methods in Natural</i>		Fabio Petroni, Tim Rocktäschel, Sebastian Riedel,	1334
1280	<i>Language Processing</i> , pages 1797–1807, Brussels,		Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and	1335
1281	Belgium. Association for Computational Linguistics.		Alexander Miller. 2019. <i>Language models as</i>	1336
			<i>knowledge bases?</i> In <i>Proceedings of the</i>	1337
1282	Yixin Nie, Adina Williams, Emily Dinan, Mohit		<i>2019 Conference on Empirical Methods in Natural</i>	1338
1283	Bansal, Jason Weston, and Douwe Kiela. 2020.		<i>Language Processing and the 9th International</i>	1339
1284	<i>Adversarial NLI: A new benchmark for natural</i>		<i>Joint Conference on Natural Language Processing</i>	1340
1285	<i>language understanding</i> . In <i>Proceedings of the 58th</i>		<i>(EMNLP-IJCNLP)</i> , pages 2463–2473, Hong Kong,	1341
1286	<i>Annual Meeting of the Association for Computational</i>		China. Association for Computational Linguistics.	1342
1287	<i>Linguistics</i> , pages 4885–4901, Online. Association			
1288	for Computational Linguistics.		Mohammad Taher Pilehvar and Jose Camacho-	1343
			Collados. 2019. <i>WiC: the word-in-context</i>	1344
1289	Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas		<i>dataset for evaluating context-sensitive meaning</i>	1345
1290	Joseph, Nova DasSarma, Tom Henighan, Ben Mann,		<i>representations</i> . In <i>Proceedings of the 2019</i>	1346
1291	Amanda Askell, Yuntao Bai, Anna Chen, et al. 2022.		<i>Conference of the North American Chapter of the</i>	1347
1292	In-context learning and induction heads. <i>arXiv</i>		<i>Association for Computational Linguistics: Human</i>	1348
1293	<i>preprint arXiv:2209.11895</i> .		<i>Language Technologies, Volume 1 (Long and Short</i>	1349
			<i>Papers)</i> , pages 1267–1273, Minneapolis, Minnesota.	1350
1294	A. Othman and M. Jemni. 2012. English-asl		Association for Computational Linguistics.	1351
1295	gloss parallel corpus 2012: Aslg-pc12. In <i>5th</i>			
1296	<i>Workshop on the Representation and Processing of</i>		Amir Pouran Ben Veyseh, Franck Dernoncourt,	1352
1297	<i>Sign Languages: Interactions between Corpus and</i>		Quan Hung Tran, and Thien Huu Nguyen. 2020.	1353
1298	<i>Lexicon LREC</i> .		<i>What does this acronym mean? introducing a new</i>	1354
			<i>dataset for acronym identification and disambigua-</i>	1355
1299	Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida,		<i>tion</i> . In <i>Proceedings of the 28th International</i>	1356
1300	Carroll L Wainwright, Pamela Mishkin, Chong		<i>Conference on Computational Linguistics</i> , pages	1357
1301	Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray,		3285–3301, Barcelona, Spain (Online). International	1358
1302	et al. 2022. <i>Training language models to follow</i>		Committee on Computational Linguistics.	1359



1474	<i>Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)</i> , pages 3477–3486, Minneapolis, Minnesota. Association for Computational Linguistics.	1531
1475		1532
1476		1533
1477		1534
1478	Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. <a href="#">Recursive deep models for semantic compositionality over a sentiment treebank</a> . In <i>Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing</i> , pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.	1535
1481		1536
1482		1537
1483		1538
1484		1539
1485		
1486	Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. <a href="#">DREAM: A challenge data set and models for dialogue-based reading comprehension</a> . <i>Transactions of the Association for Computational Linguistics</i> , 7:217–231.	1540
1487		1541
1488		1542
1489		1543
1490		1544
1491	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. <i>arXiv preprint arXiv:2210.09261</i> .	1545
1492		1546
1493		1547
1494		1548
1495		1549
1496		1550
1497	Oyvind Tafjord, Peter Clark, Matt Gardner, Wen-tau Yih, and Ashish Sabharwal. 2019a. <a href="#">Quarel: A dataset and models for answering questions about qualitative relationships</a> . <i>Proceedings of the AAAI Conference on Artificial Intelligence</i> , 33(01):7063–7071.	1551
1498		1552
1499		1553
1500		1554
1501		1555
1502	Oyvind Tafjord, Matt Gardner, Kevin Lin, and Peter Clark. 2019b. <a href="#">QuaRTz: An open-domain dataset of qualitative relationship questions</a> . In <i>Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)</i> , pages 5941–5946, Hong Kong, China. Association for Computational Linguistics.	1556
1503		1557
1504		1558
1505		1559
1506		1560
1507		1561
1508		1562
1509		1563
1510		1564
1511	Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. <a href="#">CommonsenseQA: A question answering challenge targeting common-sense knowledge</a> . In <i>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)</i> , pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.	1565
1512		1566
1513		1567
1514		1568
1515		1569
1516		1570
1517		1571
1518		1572
1519		1573
1520	Niket Tandon, Bhavana Dalvi, Keisuke Sakaguchi, Peter Clark, and Antoine Bosselut. 2019. <a href="#">WIQA: A dataset for “what if...” reasoning over procedural text</a> . In <i>Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)</i> , pages 6076–6085, Hong Kong, China. Association for Computational Linguistics.	1574
1521		1575
1522		1576
1523		1577
1524		1578
1525		1579
1526		1580
1527		1581
1528		1582
1529	TheBloke. 2023. <a href="#">Thebloke/llama-2-70b-gptq: Gptq model for meta’s llama 2 70b</a> .	1583
1530		1584
	James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. <a href="#">FEVER: a large-scale dataset for fact extraction and VERification</a> . In <i>Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)</i> , pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.	1585
		1586
		1587
		1588
	Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. <a href="#">Llama 2: Open foundation and fine-tuned chat models</a> . <i>arXiv preprint arXiv:2307.09288</i> .	
	Eleni Triantafillou, Richard Zemel, and Raquel Urtasun. 2017. Few-shot learning through an information retrieval lens. <i>arXiv preprint arXiv:1707.02610</i> .	
	Sowmya Vajjala and Ivana Lučić. 2018. <a href="#">On-eStopEnglish corpus: A new corpus for automatic readability assessment and text simplification</a> . In <i>Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications</i> , pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics.	
	Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023a. Neural codec language models are zero-shot text to speech synthesizers. <i>arXiv preprint arXiv:2301.02111</i> .	
	Liang Wang, Nan Yang, and Furu Wei. 2024. <a href="#">Learning to retrieve in-context examples for large language models</a> . In <i>Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 1752–1767, St. Julian’s, Malta. Association for Computational Linguistics.	
	William Yang Wang. 2017. <a href="#">“liar, liar pants on fire”: A new benchmark dataset for fake news detection</a> . In <i>Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)</i> , pages 422–426, Vancouver, Canada. Association for Computational Linguistics.	
	Xinyi Wang, Wanrong Zhu, and William Yang Wang. 2023b. Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning. <i>arXiv preprint arXiv:2301.11916</i> .	
	Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khoshabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. <i>arXiv preprint arXiv:2212.10560</i> .	
	Zhendong Wang, Yifan Jiang, Yadong Lu, Yelong Shen, Pengcheng He, Weizhu Chen, Zhangyang Wang, and Mingyuan Zhou. 2023c. In-context learning unlocked for diffusion models. <i>arXiv preprint arXiv:2305.01115</i> .	

1589	Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. <a href="#">BLiMP: The benchmark of linguistic minimal pairs for English</a> . <i>Transactions of the Association for Computational Linguistics</i> , 8:377–392.	1646
1590		1647
1591		1648
1592		1649
1593		1650
1594		1651
1595	Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. <a href="#">Neural network acceptability judgments</a> . <i>Transactions of the Association for Computational Linguistics</i> , 7:625–641.	1652
1596		1653
1597		1654
1598		
1599	Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. <a href="#">Emergent abilities of large language models</a> . <i>Transactions on Machine Learning Research</i> . Survey Certification.	1655
1600		1656
1601		1657
1602		1658
1603		1659
1604		
1605		
1606		
1607	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022b. <a href="#">Chain of thought prompting elicits reasoning in large language models</a> . In <i>Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS 2022)</i> .	1660
1608		1661
1609		1662
1610		1663
1611		1664
1612		1665
1613	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022c. <a href="#">Chain-of-thought prompting elicits reasoning in large language models</a> . <i>Advances in Neural Information Processing Systems</i> , 35:24824–24837.	1666
1614		1667
1615		1668
1616		1669
1617		1670
1618		1671
1619	Jerry Wei, Le Hou, Andrew Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, et al. 2023a. <a href="#">Symbol tuning improves in-context learning in language models</a> . <i>arXiv preprint arXiv:2305.08298</i> .	1672
1620		1673
1621		1674
1622		1675
1623		1676
1624	Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. 2023b. <a href="#">Larger language models do in-context learning differently</a> . <i>arXiv preprint arXiv:2303.03846</i> .	1677
1625		1678
1626		1679
1627		1680
1628		1681
1629	Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. <a href="#">Crowdsourcing multiple choice science questions</a> . In <i>Proceedings of the 3rd Workshop on Noisy User-generated Text</i> , pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.	1682
1630		1683
1631		1684
1632		1685
1633		1686
1634	Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. <a href="#">A broad-coverage challenge corpus for sentence understanding through inference</a> . In <i>Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)</i> , pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.	1687
1635		1688
1636		1689
1637		1690
1638		1691
1639		1692
1640		1693
1641		1694
1642		1695
1643	Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz,	1696
1644		1697
1645		1698
		1699
		1700
		1701
		1702
	Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. <a href="#">Transformers: State-of-the-art natural language processing</a> . In <i>Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations</i> , pages 38–45, Online. Association for Computational Linguistics.	1646
		1647
		1648
		1649
		1650
		1651
		1652
		1653
		1654
	Tomer Wolfson, Mor Geva, Ankit Gupta, Matt Gardner, Yoav Goldberg, Daniel Deutch, and Jonathan Berant. 2020. <a href="#">Break it down: A question understanding benchmark</a> . <i>Transactions of the Association for Computational Linguistics</i> , 8:183–198.	1655
		1656
		1657
		1658
		1659
	Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. <a href="#">An explanation of in-context learning as implicit bayesian inference</a> . In <i>International Conference on Learning Representations</i> .	1660
		1661
		1662
		1663
	Wenhan Xiong, Jiawei Wu, Hong Wang, Vivek Kulkarni, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2019. <a href="#">TWEETQA: A social media focused question answering dataset</a> . In <i>Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics</i> , pages 5020–5031, Florence, Italy. Association for Computational Linguistics.	1664
		1665
		1666
		1667
		1668
		1669
		1670
		1671
	Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023a. <a href="#">Wizardlm: Empowering large language models to follow complex instructions</a> . <i>arXiv preprint arXiv:2304.12244</i> .	1672
		1673
		1674
		1675
		1676
	Yiheng Xu, Hongjin Su, Chen Xing, Boyu Mi, Qian Liu, Weijia Shi, Binyuan Hui, Fan Zhou, Yitao Liu, Tianbao Xie, et al. 2023b. <a href="#">Lemur: Harmonizing natural language and code for language agents</a> . <i>arXiv preprint arXiv:2310.06830</i> .	1677
		1678
		1679
		1680
		1681
	Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. <a href="#">WikiQA: A challenge dataset for open-domain question answering</a> . In <i>Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing</i> , pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.	1682
		1683
		1684
		1685
		1686
		1687
	Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. <a href="#">HotpotQA: A dataset for diverse, explainable multi-hop question answering</a> . In <i>Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing</i> , pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.	1688
		1689
		1690
		1691
		1692
		1693
		1694
		1695
	Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. <a href="#">CrossFit: A few-shot learning challenge for cross-task generalization in NLP</a> . In <i>Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing</i> , pages 7163–7189, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.	1696
		1697
		1698
		1699
		1700
		1701
		1702



1818 Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and  
1819 Mohamed Elhoseiny. 2023. Minigt-4: Enhancing  
1820 vision-language understanding with advanced large  
1821 language models. *arXiv preprint arXiv:2304.10592*.

---

**Algorithm 1** Learning process of BiAlign

---

**Input:** ICL training set  $\mathcal{D}_{\text{ICL}} = \{\hat{X}_i = (x_1, y_1), \dots, (x_k, y_k), (\hat{x}_i, \hat{y}_i)\}$ , teacher model  $\theta_T$ , student model  $\theta_S$ , number of subsets  $N$ , weight of ranking loss  $\lambda$

- 1: **for** mini-batch  $\mathcal{B}$  in  $\mathcal{D}_{\text{ICL}}$  **do**
  - 2:   CALCULATE the KL divergence loss  $\mathcal{L}^{\text{KL}}$  on  $\mathcal{B}$  using Equation 1
  - 3:   **for**  $\hat{X}_i \in \mathcal{B}$  **do**
  - 4:     SAMPLE  $N$  subsets  $\{R_{ij}\}_{j=1}^N$  for the test sample  $(\hat{x}_i, \hat{y}_i)$
  - 5:     MEASURE preferences  $Q^T$  and  $Q^S$  for  $\{R_{ij}\}_{j=1}^N$  using Equation 2
  - 6:   **end for**
  - 7:   CALCULATE the ranking loss  $\mathcal{L}^{\text{rank}}$  on  $\mathcal{B}$  using Equation 3
  - 8:   UPDATE  $\theta_S$  by back-propagating with  $\mathcal{L} = \mathcal{L}^{\text{KL}} + \lambda\mathcal{L}^{\text{rank}}$
  - 9: **end for**
- 

	CrossFit	MMLU	BBH	GSM8K	LogiQA	HumanEval
# Samples	12K	15K	6.5K	8.5K	651	164
# Shot	4~10	5	3	8	5	0

Table 7: Details of different datasets. # refers to ‘the number of’. CrossFit (Ye et al., 2021) is used to construct training data and other tasks are used for evaluation.

## A Appendix

### A.1 Algorithm

The learning process of BiAlign is illustrated in Algorithm 1.

### A.2 Details of Source Tasks

We report the full list of source tasks used in our work in Table 16. All tasks are taken from CrossFit (Ye et al., 2021).

### A.3 Details of Target Tasks

In this work, we construct the in-context learning evaluation suite based on the following datasets:

- **MMLU:** The MMLU (Massive Multitask Language Understanding) benchmark (Hendrycks et al., 2021) consists of 57 diverse tasks covering various fields like computer science, history and law, aiming to evaluate the knowledge obtained during pretraining. Following its original setup, we use 5-shot ICL demonstrations for evaluation.
- **BBH:** The BBH (BIG-Bench Hard) (Suzgun et al., 2022) includes several types of reasoning

---

	Default	Variant
BiAlign	<b>31.0</b>	30.3

---

Table 8: Comparison between different subset sampling methods.

---

Method	Subset number			
	3	4	5	6
BiAlign	30.7	31.0	30.8	<b>31.1</b>

---

Table 9: Average performance (%) of BiAlign with different numbers of subsets  $N$ .

tasks that are believed to be difficult for current language models. Following the guidelines in Suzgun et al. (2022), we conduct the evaluation using 3-shot ICL demonstration examples with chain-of-thought prompting (Wei et al., 2022b).

- **GSM8K:** The GSM8K (Cobbe et al., 2021) dataset encompasses 8.5K grade school math word problems, aiming to evaluate the multi-step mathematical reasoning capabilities. We evaluate the ICL performance on it using 8-shot in-context examples with chain-of-thought prompting.
- **LogiQA:** LogiQA (Liu et al., 2020) is a logical reasoning benchmark sourced from logical examination papers intended for reading comprehension. Following Jiao et al. (2023), we conduct 5-shot evaluation.
- **HumanEval:** HumanEval (Chen et al., 2021) consists of 164 programming challenges for evaluating coding capabilities. We follow the official zero-shot setting in Chen et al. (2021) to verify whether bidirectional alignment hurts the zero-shot learning ability of models.

We summarize the details of all used datasets in Table 7.

### A.4 Implementation Details

Our methods are implemented with the PyTorch and Transformers library (Wolf et al., 2020). Model training is conducted utilizing DeepSpeed (Rasley et al., 2020; Rajbhandari et al., 2020) on 4 NVIDIA A100 GPUs. During the training phase, we set the learning rate to  $1e-6$  and the batch size to 64. The weight  $\lambda$  for the ranking loss is set to 1.0. For evaluation, we train the student model on the constructed ICL data for 4 epochs and assess the final checkpoint.

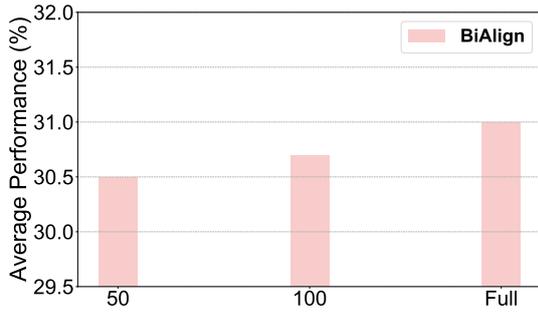


Figure 5: Average performance (%) of BiAlign with different numbers of source tasks.

Method	Type	
	Whole Sequence	Label Only
BiAlign	<b>31.0</b>	30.8

Table 10: Average performance (%) of BiAlign using different types of KL divergence calculation methods.

### A.5 Different Subset Sampling Methods

To investigate the influence of subset sampling methods, we replace the original method with ‘Randomly sample  $N$  subsets’ which does not consider similarity. The comparison between the two methods is shown in Table 8. We can observe a noticeable performance drop, highlighting the crucial role of incorporating example similarity in the sampling process.

### A.6 Different Numbers of Subsets

While we use  $N = 4$  subsets for calculating the ranking loss, we also evaluate the effectiveness of BiAlign with different  $N$ . Specifically, we conduct controlled experiments with  $\{3, 5, 6\}$  subsets and report the average results of the 5 target tasks in Table 9. We can observe that increasing the number of subsets does not always improve performance. BiAlign achieves the best performance (31.1) with 6 subsets and the performance with 4 subsets (31.0) is comparable. Besides, all variants consistently outperform baseline methods in Table 1, demonstrating the effectiveness of our designed input preference alignment.

### A.7 Different Source Task Selections

We hypothesize that the diversity of source tasks has a considerable influence on target task performance. To verify this, we study the effect of the number of source tasks by conducting controlled experiments on  $\{50, 100\}$  randomly selected source tasks. From the results in Figure 5,

Method	25%	50%	100%
Output-Align	29.1	29.3	29.4
BiAlign	<b>30.3</b>	<b>30.8</b>	<b>31.0</b>

Table 11: Comparison between BiAlign and Output-Align at different proportions of training steps.

we can observe that the performance of BiAlign keeps improving as the number of source tasks increases, indicating the importance of source task diversity.

### A.8 Whole Sequence vs. Label Only

To maintain the basic in-weights capability of the student model, we minimize the KL divergence loss for the whole sequence instead of only the label following Gu et al. (2023). In Table 10, we show the performance comparison between using the whole sequence and using only the label. We can see that using the whole sequence also results in slightly better average performance.

### A.9 Different Proportions of Training Steps

Table 11 reports the performance comparison between BiAlign and Output-Align at different proportions (roughly 25%, 50%, and 100%) of training steps. We can observe that BiAlign consistently outperforms Output-Align at different steps.

### A.10 Additional Training Data

The analysis in Section 5.2 shows that conducting Output-Align on the combination of the original ICL training examples and the sampled subset data achieves similar performance to the original Output-Align method. We further experiment with the fine-tuning approach. However, the performance becomes even worse (from 29.5 to 29.3), once again demonstrating that simply increasing training data does not necessarily lead to better performance.

### A.11 Ranking Loss Weights

To further investigate the influence of the ranking loss  $\mathcal{L}^{\text{rank}}$  (Equation 3), we conduct experiments with different weights  $\lambda$  and report the results in Table 12. All variants except the variant with  $\lambda = 5.0$  (too large) outperform baseline approaches by a large margin, which demonstrates the superiority of  $\mathcal{L}^{\text{rank}}$ .

Method	$\lambda$				
	0.2	0.5	1.0	2.0	5.0
BiAlign	30.8	<b>31.2</b>	31.0	30.9	29.9

Table 12: Average performance (%) of BiAlign with different  $\lambda$  for the ranking loss  $\mathcal{L}^{\text{rank}}$ .

Method	Pair number			
	3	4	5	All (6)
BiAlign	30.2	30.8	30.7	<b>31.0</b>

Table 13: Average results (%) of BiAlign with different numbers of contrastive pairs.

## 1946 A.12 Contrastive Pair Selection

1947 While we use all  $C(N, 2)$  ( $N = 4$  is the number of  
1948 subsets) pairs of positive and negative subsets for  
1949 input preference alignment, we also investigate the  
1950 effect of contrastive pair selection. Specifically,  
1951 we conduct controlled experiments on  $\{3, 4, 5\}$   
1952 randomly selected contrastive pairs and report the  
1953 average results in Table 13. The best performance  
1954 is achieved using all pairs, justifying our selection  
1955 strategy.

## 1956 A.13 Case Study

1957 We select GSM8K as a representative task and  
1958 show several examples of output in Table 14.  
1959 Compared with Vallina, BiAlign is able to generate  
1960 more precise and fine-grained reasoning paths, *e.g.*,  
1961 BiAlign can successfully understand the meaning  
1962 of ‘pair’ and generate the rationale ‘10 pairs of  
1963 socks is  $10 * 2 = 20$  socks’ while Vallina fails to do  
1964 so.

## 1965 A.14 Example in LogiQA

1966 We show the reasoning path required to solve a  
1967 sample in LogiQA in Table 15, which is quite fine-  
1968 grained.

---

Answer the following questions.

*Few-shot demonstrations*

**Question:** Lindsay is doing the laundry, and thinks she has missed some socks. There are 50 socks that need washing. If she washes 10 pairs of socks and 15 loose socks, how many socks has Lindsay missed?

**Answer:**

---

**Vanilla:** Lindsay started with 50 socks. She washed 10 pairs of socks and 15 loose socks. **So she had  $50 - 10 - 15 = 25$  socks left. So the answer is 25. ✗**

---

**BiAlign:** Lindsay started with 50 socks. She washed 10 pairs of socks and 15 loose socks. **10 pairs of socks is  $10 * 2 = 20$  socks. 15 loose socks is  $15 * 1 = 15$  socks. So she missed  $50 - 20 - 15 = 15$  socks. So the answer is 15. ✓**

---

Table 14: Output examples of Vanilla and BiAlign on GSM8K. The upper part is the input to LLMs, including few-shot demonstrations, and the lower part is the generated response. We color correct outputs in green and wrong outputs in red.

---

**Context:** There is no doubt that minors should be prohibited from smoking. However, we cannot explicitly ban the use of automatic cigarette vending machines in order to prevent minors from smoking. This ban is just like setting up roadblocks on the road to prohibit driving without a license. These roadblocks naturally prohibit driving without a license, but also block more than 99% of licensed drivers.

**Question:** In order to evaluate the above argument, which of the following questions is the most important?

**Options:**

A: Does the proportion of underage smokers in the total number of smokers exceed 1%?

B: How much inconvenience does the ban on the use of automatic vending machines bring to adult cigarette buyers?

C: Whether the proportion of unlicensed drivers in the total number of drivers really does not exceed 1%?

D: Is the harm of minor smoking really as serious as the public thinks?

---

**Reasoning path:**

To evaluate the argument effectively, the focus should be on understanding the impact and justification of the proposed ban on cigarette vending machines, especially in the context of preventing minors from smoking. The argument draws a parallel between the proposed ban and the hypothetical scenario of setting up roadblocks to prevent driving without a license, suggesting that while the measure may target a minority (in this case, underage smokers or unlicensed drivers), it disproportionately inconveniences the majority (licensed drivers or adult smokers).

*The most important question to evaluate the argument is: B: How much inconvenience does the ban on the use of automatic vending machines bring to adult cigarette buyers?*

This option directly addresses the central issue of the argument—the balance between the inconvenience caused to the majority by the ban and the intended benefit of preventing a minority (minors) from engaging in harmful behavior (smoking). It’s critical to assess whether the inconvenience to adult smokers (who are the majority of cigarette consumers) is justified by the potential benefits of reducing underage smoking. The comparison with roadblocks for unlicensed drivers emphasizes the argument’s concern with the proportionality and fairness of broad preventive measures.

Option A concerns the proportion of underage smokers but doesn’t directly address the balance of convenience versus benefit. Option C relates to the analogy used but does not directly help in evaluating the effectiveness and appropriateness of the vending machine ban. Option D questions the severity of the problem (minor smoking) but does not specifically help in assessing the argument’s concern about the consequences of the ban on the majority.

Therefore, understanding the extent of inconvenience to adult smokers is crucial in determining whether the proposed solution is proportionate and justified, making Option B the most relevant and important question for evaluating the argument.

---

Table 15: Reasoning path for an example in LogiQA.

Task Name	Reference
eli5-eli5	(Fan et al., 2019)
ethos-race	(Mollas et al., 2020)
tweet_qa	(Xiong et al., 2019)
tweet_eval-stance_hillary	(Barbieri et al., 2020)
piqa	(Bisk et al., 2020)
acronym_identification	(Pouan Ben Veysch et al., 2020)
wiki_split	(Botha et al., 2018)
scitail	(Khot et al., 2018)
emotion	(Saravia et al., 2018)
medical_questions_pairs	(McCreery et al., 2020)
blimp-anaphor_gender_agreement	(Warstadt et al., 2020)
sciq	(Welbl et al., 2017)
paws	(Zhang et al., 2019)
yelp_review_full	(Zhang et al., 2015); (link)
freebase_qa	(Jiang et al., 2019)
anli	(Nie et al., 2020)
quartz-with_knowledge	(Tafjord et al., 2019b)
hatexplain	(Mathew et al., 2020)
yahoo_answers_topics	(link)
search_qa	(Dunn et al., 2017)
tweet_eval-stance_feminist	(Barbieri et al., 2020)
codah	(Chen et al., 2019)
lama-squad	(Petroni et al., 2019, 2020)
superglue-record	(Zhang et al., 2018)
spider	(Yu et al., 2018)
mc_taco	(Zhou et al., 2019)
glue-mrpc	(Dolan and Brockett, 2005)
kilt_fever	(Thorne et al., 2018)
eli5-asks_qa	(Fan et al., 2019)
imdb	(Maas et al., 2011)
tweet_eval-stance_abortion	(Barbieri et al., 2020)
aqua_rat	(Ling et al., 2017b)
duorc	(Saha et al., 2018)
lama-trex	(Petroni et al., 2019, 2020)
tweet_eval-stance_atheism	(Barbieri et al., 2020)
ropes	(Lin et al., 2019)
squad-no_context	(Rajpurkar et al., 2016)
superglue-rte	(Dagan et al., 2005)
qasc	(Khot et al., 2020)
hate_speech_offensive	(Davidson et al., 2017)
trec-finegrained	(Li and Roth, 2002; Hovy et al., 2001)
glue-wnli	(Levesque et al., 2012)
yelp_polarity	(Zhang et al., 2015); (link)
kilt_hotpotqa	(Yang et al., 2018)
glue-sst2	(Socher et al., 2013)
xsum	(Narayan et al., 2018)
tweet_eval-offensive	(Barbieri et al., 2020)
aeslc	(Zhang and Tetreault, 2019)
emo	(Chatterjee et al., 2019)
hellaswag	(Zellers et al., 2019)
social_i_qa	(Sap et al., 2019)
kilt_wow	(Dinan et al., 2019)
scicite	(Cohan et al., 2019)
superglue-wsc	(Levesque et al., 2012)
hate_speech18	(de Gibert et al., 2018)
adversarialqa	(Bartolo et al., 2020)
break-QDMR	(Wolfson et al., 2020)
dream	(Sun et al., 2019)
circa	(Louis et al., 2020)
wiki_qa	(Yang et al., 2015)
ethos-directed_vs_generalized	(Mollas et al., 2020)
wiqa	(Tandon et al., 2019)
poem_sentiment	(Sheng and Uthus, 2020)
kilt_ay2	(Hoffart et al., 2011)
cosmos_qa	(Huang et al., 2019)
reddit_tifu-title	(Kim et al., 2019)
superglue-cb	(de Marneffe et al., 2019)
kilt_nq	(Kwiatkowski et al., 2019)
quarel	(Tafjord et al., 2019a)
race-high	(Lai et al., 2017)
wino_grande	(Sakaguchi et al., 2020)
break-QDMR-high-level	(Wolfson et al., 2020)
tweet_eval-irony	(Barbieri et al., 2020)
liar	(Wang, 2017)
openbookqa	(Mihaylov et al., 2018)
superglue-multirc	(Khashabi et al., 2018)
race-middle	(Lai et al., 2017)
quoref	(Dasigi et al., 2019)
cos_e	(Rajani et al., 2019)
reddit_tifu-tldr	(Kim et al., 2019)
ai2_arc	(Clark et al., 2018)
quail	(Rogers et al., 2020)
crawl_domain	(Zhang et al., 2020)
glue-cola	(Warstadt et al., 2019)

Task Name	Reference
art	(Bhagavatula et al., 2020)
rotten_tomatoes	(Pang and Lee, 2005)
tweet_eval-emoji	(Barbieri et al., 2020)
numer_sense	(Lin et al., 2020a)
blimp-existential_there_quantifiers_1	(Warstadt et al., 2020)
eli5-askh_qa	(Fan et al., 2019)
ethos-national_origin	(Mollas et al., 2020)
boolq	(Clark et al., 2019)
qa_srl	(He et al., 2015)
sms_spam	(Almeida et al., 2011)
samsun	(Gliwa et al., 2019)
ade_corpus_v2-classification	(Gurulingappa et al., 2012)
superglue-wic	(Pilehvar and Camacho-Collados, 2019)
ade_corpus_v2-dosage	(Gurulingappa et al., 2012)
tweet_eval-stance_climate	(Barbieri et al., 2020)
e2e_nlg_cleaned	(Dušek et al., 2020, 2019)
aslg_pc12	(Othman and Jemni, 2012)
ag_news	Gulli (link)
math_qa	(Amini et al., 2019)
commonsense_qa	(Talmor et al., 2019)
web_questions	(Berant et al., 2013)
biomrc	(Pappas et al., 2020)
swag	(Zellers et al., 2018)
blimp-determiner_noun_agreement_with_adj_irregular_1	(Warstadt et al., 2020)
glue-mnli	(Williams et al., 2018)
squad-with_context	(Rajpurkar et al., 2016)
blimp-ellipsis_n_bar_2	(Warstadt et al., 2020)
financial_phrasebank	(Malo et al., 2014)
sick	(Marelli et al., 2014)
ethos-religion	(Mollas et al., 2020)
hotpot_qa	(Yang et al., 2018)
tweet_eval-emotion	(Barbieri et al., 2020)
dbpedia_14	(Lehmann et al., 2015)
ethos-gender	(Mollas et al., 2020)
tweet_eval-hate	(Barbieri et al., 2020)
ethos-sexual_orientation	(Mollas et al., 2020)
health_fact	(Kotonya and Toni, 2020)
common_gen	(Lin et al., 2020b)
crows_pairs	(Nangia et al., 2020)
ade_corpus_v2-effect	(Gurulingappa et al., 2012)
blimp-sentential_negation_npi_scope	(Warstadt et al., 2020)
lama-conceptnet	(Petroni et al., 2019, 2020)
glue-qnli	(Rajpurkar et al., 2016)
quartz-no_knowledge	(Tafjord et al., 2019b)
google_wellformed_query	(Faruqui and Das, 2018)
kilt_trex	(Elsahar et al., 2018)
blimp-ellipsis_n_bar_1	(Warstadt et al., 2020)
trec	(Li and Roth, 2002; Hovy et al., 2001)
superglue-copa	(Gordon et al., 2012)
ethos-disability	(Mollas et al., 2020)
lama-google_re	(Petroni et al., 2019, 2020)
discovery	(Sileo et al., 2019)
blimp-anaphor_number_agreement	(Warstadt et al., 2020)
climate_fever	(Diggelmann et al., 2020)
blimp-irregular_past_participle_adjectives	(Warstadt et al., 2020)
tab_fact	(Chen et al., 2020)
gigaword	(Napoles et al., 2012)
glue-rte	(Dagan et al., 2005)
tweet_eval-sentiment	(Barbieri et al., 2020)
limit	(Manotas et al., 2020)
wikisql	(Zhong et al., 2017)
glue-qqp	(link)
onestop_english	(Vajjala and Lučić, 2018)
amazon_polarity	(McAuley and Leskovec, 2013)
blimp-wh_questions_object_gap	(Warstadt et al., 2020)
multi_news	(Fabbri et al., 2019)
proto_qa	(Boratko et al., 2020)
wiki_bio	(Lebret et al., 2016)
kilt_zsre	(Levy et al., 2017)
blimp-sentential_negation_npi_licensor_present	(Warstadt et al., 2020)

Table 16: List of all source tasks.