

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).¹

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., 2024). 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

¹Different 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.

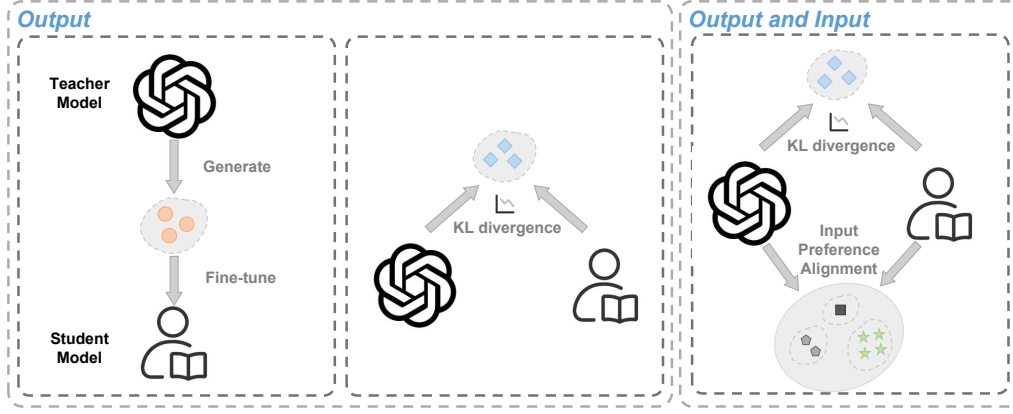


Figure 1: Comparison between different types of approaches to aligning student models. Existing methods typically fine-tune student models on generated outputs of teacher models or to match their token-level output probability distributions (*left* part). In contrast, our method (BiAlign) considers the 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).² 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 variety of tasks. For example, it brings about 20%

²Our 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.

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., 2024; 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 (Hsieh et al., 2023; Wang et al., 2022; Xu et al., 2023a) or to imitate teacher models’ token-level

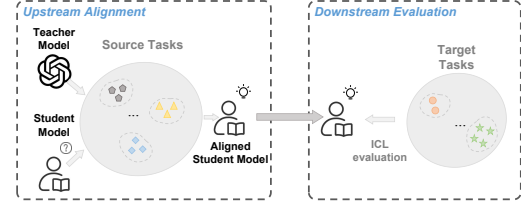


Figure 2: In the upstream ICL alignment stage, we align a student model with a teacher on the source tasks. Then in the downstream evaluation stage, we evaluate the ICL performance of the aligned student model on a held-out set of target tasks, which are different from the source tasks.

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., 2024). 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}^{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}^{tgt} . We divide the whole process into two stages, as illustrated in Figure 2.

• **Upstream ICL alignment on \mathcal{T}^{src} :** In this alignment stage, the model has access to \mathcal{T}^{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}^{tgt} :** Following the upstream ICL alignment stage, we evaluate the ICL ability of the aligned model S^* on \mathcal{T}^{tgt} , where \mathcal{T}^{tgt} has no overlap with \mathcal{T}^{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. (2024).³ 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; θ_T and θ_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_{sim} and G_{dissim} , based on their similarity to the test example (\hat{x}_i, \hat{y}_i) (see Appendix A.1 for details). 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:⁴

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

where Q^T and Q^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:

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

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 of the student model for the pair (R^+, R^-) and the

³Training 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).

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

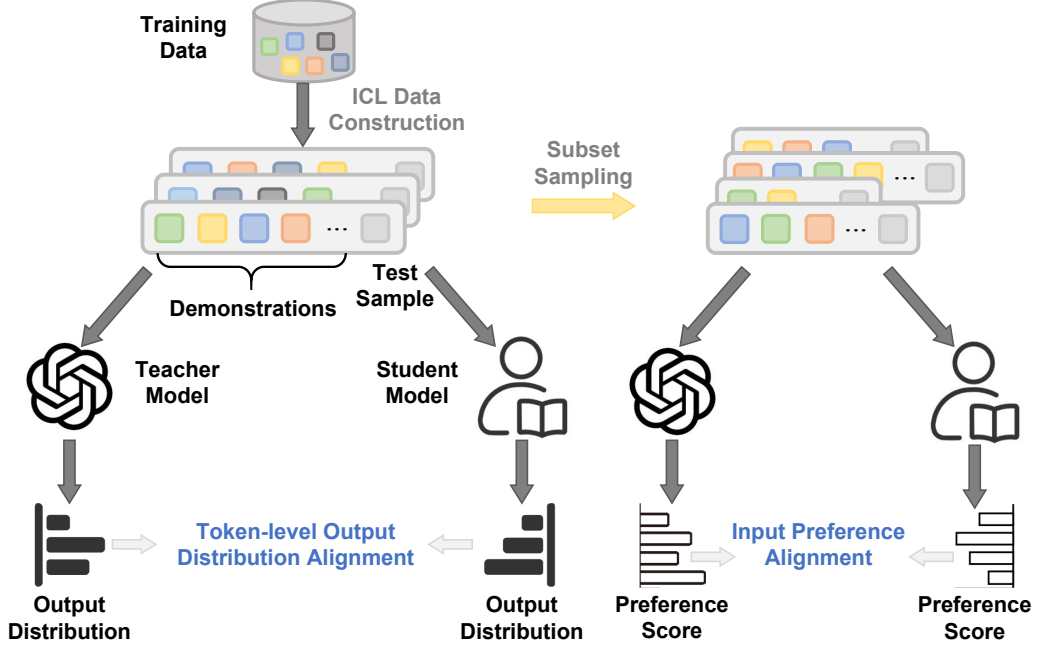


Figure 3: Illustration of our Bidirectional Alignment (BiAlign) framework. It attains *token-level output distribution alignment* by minimizing the KL divergence loss between the student and teacher models on the constructed ICL samples. Furthermore, after sampling several subsets from the set of all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

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 λ is the weight of the ranking loss. Besides, we illustrate the whole learning process in Appendix A.2.

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}^{src} (see Appendix A.3 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.4 and A.5, respectively.

4.2 Methods Compared

We mainly experiment with Llama 2-7B (Touvron 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

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	47.5 \pm 0.4	41.0 \pm 0.3	16.8 \pm 0.3	33.9 \pm 0.4	15.6 \pm 0.4	31.0 \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	49.5 \pm 0.3	43.2 \pm 0.5	18.3 \pm 0.4	35.7 \pm 0.4	16.6 \pm 0.3	32.7 \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.

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](#)).

Additionally, we show the connection between BiAlign and In-Context Pretraining ([Shi et al., 2024](#)) in Section 5.2, and discuss how BiAlign can be integrated with the latest ICL demonstration selection methods or reverse KL divergence in Appendix A.6 and A.7.

5 Results and Analysis

5.1 Main Results

Table 1 shows the performance scores of different methods on all investigated target tasks. From the results, we can observe that

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

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

- 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 better improve the performance of tasks requiring more fine-grained reasoning; see appendix A.21 for an example in LogiQA. This is because BiAlign

Method	7B	13B
Output-Align	30.2	38.8
BiAlign	32.7	40.9

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	63.9
Phi-3-mini (3.8B)	66.7	67.1	66.5	67.4	69.1

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).

allows the student model to obtain more fine-grained supervision from the teacher model by fully leveraging their preferences for different inputs.

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

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

5.2 Analysis

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

Different Backbone Models Our experiments and analysis so far use Llama 2 as the backbone

	Default	Variant
BiAlign	31.0	30.5

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

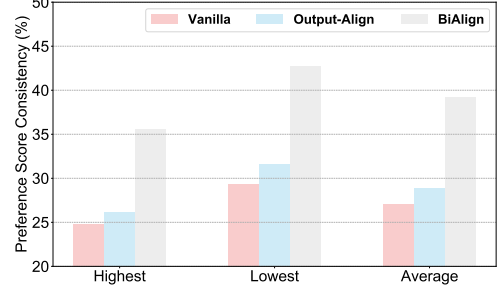


Figure 4: Preference score consistency (%) of different methods.

model. To verify whether the performance gain of BiAlign is consistent across different backbone models, we extend the experiments to Llama 3 (Dubey et al., 2024) and Phi 3 (Abdin et al., 2024). For Llama 3, we use the 8B model as the student and the 70B model as the teacher. For Phi 3, we use Phi-3-mini (3.8B) as the student and Phi-3-medium (14B) as the teacher. From the average results shown in Table 4, we can see that BiAlign still outperforms all baseline approaches when using other language models as the backbone, showing its robustness to model types. In addition, we show the scalability of BiAlign across more model scales using Qwen-2.5 (Yang et al., 2024) in Appendix A.8.

Comment on Training-time Computational Overhead Smaller models are a preferred choice for resource-constrained deployments, where the inference cost matters the most. BiAlign does not introduce any additional cost during inference. The additional computational overhead only occurs once during model training. To quantify the increase in computational overhead caused by the ranking loss, we use DeepSpeed Flops Profiler (Rasley et al., 2020) to calculate the training FLOPs of Output-Align and BiAlign, which are 3.3×10^{17} and 7.6×10^{17} respectively (about 2.3 times). Therefore, we further design two experiments to compare BiAlign and Output-Align under the same training FLOPs: (i) we combine the original ICL training examples with the sampled subset data 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 `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 `rank()`, while maintaining all other implementation details. The average results reported in Table 5 underscore the importance of using `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 in Table 6. We can see that training with a fixed number of demonstrations results in performance

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

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

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 (see results on other datasets in Appendix A.9). 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 further justification of input preference alignment, more cross-task generalization experiments, 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.10 ~ A.20, 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.

Limitations

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

References

- Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.
- Rishabh Agarwal, Nino Vieillard, Piotr Stanczyk, Sabella Ramos, Matthieu Geist, and Olivier Bachem. 2023. Gkd: Generalized knowledge distillation for auto-regressive sequence models. *arXiv preprint arXiv:2306.13649*.
- Tiago A. Almeida, José María G. Hidalgo, and Akebo Yamakami. 2011. [Contributions to the study of sms spam filtering: New collection and results](#). In *Proceedings of the 11th ACM Symposium on Document Engineering, DocEng '11*, page 259–262, New York, NY, USA. Association for Computing Machinery.
- Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. [MathQA: Towards interpretable math word problem solving with operation-based formalisms](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.
- Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR.
- Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. [TweetEval: Unified benchmark and comparative evaluation for tweet classification](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.
- Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. [Beat the AI: Investigating adversarial human annotation for reading comprehension](#). *Transactions of the*

Association for Computational Linguistics, 8:662–678.

- Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. [Semantic parsing on Freebase from question-answer pairs](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.
- Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. 2020. [Abductive commonsense reasoning](#). In *International Conference on Learning Representations*.
- Yonatan Bisk, Rowan Zellers, Ronan LeBras, Jianfeng Gao, and Yejin Choi. 2020. [PIQA: reasoning about physical commonsense in natural language](#). In *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*, pages 7432–7439. AAAI Press.
- Michael Boratko, Xiang Li, Tim O’Gorman, Rajarshi Das, Dan Le, and Andrew McCallum. 2020. [ProtoQA: A question answering dataset for prototypical common-sense reasoning](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1122–1136, Online. Association for Computational Linguistics.
- Jan A. Botha, Manaal Faruqui, John Alex, Jason Baldridge, and Dipanjan Das. 2018. [Learning to split and rephrase from Wikipedia edit history](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 732–737, Brussels, Belgium. Association for Computational Linguistics.
- Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. 2023. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*.
- Stephanie C.Y. Chan, Adam Santoro, Andrew Kyle Lampinen, Jane X Wang, Aaditya K Singh, Pierre Harvey Richemond, James McClelland, and Felix Hill. 2022. [Data distributional properties drive emergent in-context learning in transformers](#). In *Advances in Neural Information Processing Systems*.

688	Ankush Chatterjee, Kedhar Nath Narahari, Meghana	Plappert, Jerry Tworek, Jacob Hilton, Reiichiro	746
689	Joshi, and Puneet Agrawal. 2019. SemEval-2019 task	Nakano, et al. 2021. Training verifiers to solve math	747
690	3: EmoContext contextual emotion detection in text.	word problems. <i>arXiv preprint arXiv:2110.14168</i> .	748
691	In <i>Proceedings of the 13th International Workshop</i>		
692	<i>on Semantic Evaluation</i> , pages 39–48, Minneapolis,	Arman Cohan, Waleed Ammar, Madeleine van Zuylen,	749
693	Minnesota, USA. Association for Computational	and Field Cady. 2019. Structural scaffolds for	750
694	Linguistics.	citation intent classification in scientific publications.	751
695	Mark Chen, Jerry Tworek, Heewoo Jun, Qiming	In <i>Proceedings of the 2019 Conference of the</i>	752
696	Yuan, Henrique Ponde de Oliveira Pinto, Jared	<i>North American Chapter of the Association for</i>	753
697	Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph,	<i>Computational Linguistics: Human Language Tech-</i>	754
698	Greg Brockman, et al. 2021. Evaluating large	<i>nologies, Volume 1 (Long and Short Papers)</i> , pages	755
699	language models trained on code. <i>arXiv preprint</i>	3586–3596, Minneapolis, Minnesota. Association for	756
700	<i>arXiv:2107.03374</i> .	Computational Linguistics.	757
701	Michael Chen, Mike D’Arcy, Alisa Liu, Jared	Ido Dagan, Oren Glickman, and Bernardo Magnini.	758
702	Fernandez, and Doug Downey. 2019. CODAH: An	2005. The pascal recognising textual entailment	759
703	adversarially-authored question answering dataset for	challenge. In <i>Machine Learning Challenges</i>	760
704	common sense. In <i>Proceedings of the 3rd Workshop</i>	<i>Workshop</i> , pages 177–190. Springer.	761
705	<i>on Evaluating Vector Space Representations for NLP</i> ,		
706	pages 63–69, Minneapolis, USA. Association for	Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming	762
707	Computational Linguistics.	Ma, Zhifang Sui, and Furu Wei. 2023. Why can gpt	763
708	Mingda Chen, Jingfei Du, Ramakanth Pasunuru, Todor	learn in-context? language models implicitly perform	764
709	Mihaylov, Sridi Iyer, Veselin Stoyanov, and Zornitsa	gradient descent as meta-optimizers. In <i>ICLR</i>	765
710	Kozareva. 2022. Improving in-context few-shot	<i>2023 Workshop on Mathematical and Empirical</i>	766
711	learning via self-supervised training. In <i>Proceedings</i>	<i>Understanding of Foundation Models.</i>	767
712	<i>of the 2022 Conference of the North American</i>		
713	<i>Chapter of the Association for Computational</i>	Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A.	768
714	<i>Linguistics: Human Language Technologies</i> , pages	Smith, and Matt Gardner. 2019. Quoref: A reading	769
715	3558–3573, Seattle, United States. Association for	comprehension dataset with questions requiring	770
716	Computational Linguistics.	coreferential reasoning. In <i>Proceedings of the</i>	771
717	Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai	<i>2019 Conference on Empirical Methods in Natural</i>	772
718	Zhang, Hong Wang, Shiyang Li, Xiyu Zhou,	<i>Language Processing and the 9th International</i>	773
719	and William Yang Wang. 2020. Tabfact: A	<i>Joint Conference on Natural Language Processing</i>	774
720	large-scale dataset for table-based fact verification.	<i>(EMNLP-IJCNLP)</i> , pages 5925–5932, Hong Kong,	775
721	In <i>8th International Conference on Learning</i>	China. Association for Computational Linguistics.	776
722	<i>Representations, ICLR 2020, Addis Ababa, Ethiopia,</i>		
723	<i>April 26-30, 2020.</i> OpenReview.net.	Thomas Davidson, Dana Warmesley, Michael Macy,	777
724	Paul F Christiano, Jan Leike, Tom Brown, Miljan	and Ingmar Weber. 2017. Automated hate speech	778
725	Martic, Shane Legg, and Dario Amodei. 2017. Deep	detection and the problem of offensive language.	779
726	reinforcement learning from human preferences.	In <i>Proceedings of the 11th International AAAI</i>	780
727	<i>Advances in neural information processing systems</i> ,	<i>Conference on Web and Social Media, ICWSM ’17,</i>	781
728	30.	pages 512–515.	782
729	Christopher Clark, Kenton Lee, Ming-Wei Chang,	Ona de Gibert, Naiara Perez, Aitor García-Pablos, and	783
730	Tom Kwiatkowski, Michael Collins, and Kristina	Montse Cuadros. 2018. Hate speech dataset from	784
731	Toutanova. 2019. BoolQ: Exploring the surprising	a white supremacy forum. In <i>Proceedings of the</i>	785
732	difficulty of natural yes/no questions. In <i>Proceedings</i>	<i>2nd Workshop on Abusive Language Online (ALW2),</i>	786
733	<i>of the 2019 Conference of the North American</i>	pages 11–20, Brussels, Belgium. Association for	787
734	<i>Chapter of the Association for Computational</i>	Computational Linguistics.	788
735	<i>Linguistics: Human Language Technologies, Volume</i>	Marie-Catherine de Marneffe, Mandy Simons, and	789
736	<i>1 (Long and Short Papers)</i> , pages 2924–2936, Min-	Judith Tonhauser. 2019. The commitmentbank:	790
737	neapolis, Minnesota. Association for Computational	Investigating projection in naturally occurring	791
738	Linguistics.	discourse. <i>Proceedings of Sinn und Bedeutung</i> ,	792
739	Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot,	23(2):107–124.	793
740	Ashish Sabharwal, Carissa Schoenick, and Oyvind	Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and	794
741	Tafford. 2018. Think you have solved question	Luke Zettlemoyer. 2023. Qlora: Efficient finetuning	795
742	answering? try arc, the ai2 reasoning challenge.	of quantized llms. <i>arXiv preprint arXiv:2305.14314</i> .	796
743	<i>ArXiv preprint</i> , abs/1803.05457.	T. Diggelmann, Jordan L. Boyd-Graber, Jannis Bulian,	797
744	Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian,	Massimiliano Ciaramita, and Markus Leippold.	798
745	Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias	2020. Climate-fever: A dataset for verification	799
		of real-world climate claims. <i>ArXiv preprint</i> ,	800
		abs/2012.00614.	801

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

914	<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.	971
915		972
916		973
917		974
918		
919		
920	Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. MiniLLM: Knowledge distillation of large language models . In <i>The Twelfth International Conference on Learning Representations</i> .	
921		
922		
923		
924	Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports . <i>Journal of Biomedical Informatics</i> , 45(5):885–892.	975
925		976
926		977
927		978
928		979
929		980
930		981
931		982
932		983
933	Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-answer driven semantic role labeling: Using natural language to annotate natural language . In <i>Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing</i> , pages 643–653, Lisbon, Portugal. Association for Computational Linguistics.	
934		
935		
936		
937		
938		
939	Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding . In <i>International Conference on Learning Representations</i> .	984
940		985
941		986
942		987
943		988
944		989
945	Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. <i>arXiv preprint arXiv:1503.02531</i> .	
946		
947	Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenu, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text . 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.	990
948		991
949		992
950		993
951		994
952		995
953		996
954		997
955	Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Chin-Yew Lin, and Deepak Ravichandran. 2001. Toward semantics-based answer pinpointing . In <i>Proceedings of the First International Conference on Human Language Technology Research</i> .	998
956		999
957		1000
958		1001
959		1002
960	Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes . In <i>Findings of the Association for Computational Linguistics: ACL 2023</i> , pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.	1003
961		1004
962		1005
963		1006
964		1007
965		
966		
967		
968		
969	Zikun Hu, Xiang Li, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2018. Few-shot charge prediction with discriminative legal attributes . In <i>Proceedings of the 27th International Conference on Computational Linguistics</i> , pages 487–498, Santa Fe, New Mexico, USA. Association for Computational Linguistics.	1008
970		1009
		1010
		1011
		1012
		1013
		1014
		1015
		1016
		1017
		1018
		1019
		1020
		1021
		1022
		1023
		1024
		1025
		1026

1027	<i>Technologies, Volume 1 (Long Papers)</i> , pages 252–	<i>in Natural Language Processing</i> , pages 1203–	1084
1028	262, New Orleans, Louisiana. Association for	1213, Austin, Texas. Association for Computational	1085
1029	Computational Linguistics.	Linguistics.	1086
1030	Tushar Khot, Peter Clark, Michal Guerquin, Peter	Jens Lehmann, Robert Isele, Max Jakob, Anja	1087
1031	Jansen, and Ashish Sabharwal. 2020. Qasc:	Jentzsch, D. Kontokostas, Pablo N. Mendes,	1088
1032	A dataset for question answering via sentence	Sebastian Hellmann, M. Morsey, Patrick van Kleef,	1089
1033	composition . <i>Proceedings of the AAAI Conference</i>	S. Auer, and C. Bizer. 2015. Dbpedia - a large-	1090
1034	<i>on Artificial Intelligence</i> , 34(05):8082–8090.	scale, multilingual knowledge base extracted from	1091
1035	Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018.	wikipedia. <i>Semantic Web</i> , 6:167–195.	1092
1036	Scitail: A textual entailment dataset from science	Hector J. Levesque, Ernest Davis, and Leora	1093
1037	question answering . In <i>Proceedings of the Thirty-</i>	Morgenstern. 2012. The winograd schema challenge.	1094
1038	<i>Second AAAI Conference on Artificial Intelligence,</i>	In <i>Proceedings of the Thirteenth International Con-</i>	1095
1039	<i>(AAAI-18), the 30th innovative Applications of</i>	<i>ference on Principles of Knowledge Representation</i>	1096
1040	<i>Artificial Intelligence (IAAI-18), and the 8th AAAI</i>	<i>and Reasoning</i> , KR’12, page 552–561. AAAI Press.	1097
1041	<i>Symposium on Educational Advances in Artificial</i>	Omer Levy, Minjoon Seo, Eunsol Choi, and Luke	1098
1042	<i>Intelligence (EAAI-18), New Orleans, Louisiana,</i>	Zettlemoyer. 2017. Zero-shot relation extraction via	1099
1043	<i>USA, February 2-7, 2018</i> , pages 5189–5197. AAAI	reading comprehension . In <i>Proceedings of the 21st</i>	1100
1044	Press.	<i>Conference on Computational Natural Language</i>	1101
1045	Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim.	<i>Learning (CoNLL 2017)</i> , pages 333–342, Vancouver,	1102
1046	2019. Abstractive summarization of Reddit posts	Canada. Association for Computational Linguistics.	1103
1047	with multi-level memory networks . In <i>Proceedings</i>	Xin Li and Dan Roth. 2002. Learning question	1104
1048	<i>of the 2019 Conference of the North American</i>	classifiers . In <i>COLING 2002: The 19th International</i>	1105
1049	<i>Chapter of the Association for Computational</i>	<i>Conference on Computational Linguistics</i> .	1106
1050	<i>Linguistics: Human Language Technologies, Volume</i>	Yingcong Li, Muhammed Emrullah Ildiz, Dimitris	1107
1051	<i>1 (Long and Short Papers)</i> , pages 2519–2531, Min-	Papailiopoulos, and Samet Oymak. 2023a. Trans-	1108
1052	neapolis, Minnesota. Association for Computational	formers as algorithms: Generalization and stability	1109
1053	Linguistics.	in in-context learning.	1110
1054	Jannik Kossen, Yarin Gal, and Tom Rainforth. 2024. In-	Ziniu Li, Tian Xu, Yushun Zhang, Yang Yu, Ruoyu	1111
1055	context learning learns label relationships but is not	Sun, and Zhi-Quan Luo. 2023b. Remax: A simple,	1112
1056	conventional learning . In <i>The Twelfth International</i>	effective, and efficient method for aligning large	1113
1057	<i>Conference on Learning Representations</i> .	language models. <i>arXiv preprint arXiv:2310.10505</i> .	1114
1058	Neema Kotonya and Francesca Toni. 2020. Explainable	Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang	1115
1059	automated fact-checking for public health claims .	Ren. 2020a. Birds have four legs?! NumerSense:	1116
1060	In <i>Proceedings of the 2020 Conference on</i>	Probing Numerical Commonsense Knowledge of Pre-	1117
1061	<i>Empirical Methods in Natural Language Processing</i>	Trained Language Models . In <i>Proceedings of the</i>	1118
1062	<i>(EMNLP)</i> , pages 7740–7754, Online. Association for	<i>2020 Conference on Empirical Methods in Natural</i>	1119
1063	Computational Linguistics.	<i>Language Processing (EMNLP)</i> , pages 6862–6868,	1120
1064	Tom Kwiatkowski, Jennimaria Palomaki, Olivia	Online. Association for Computational Linguistics.	1121
1065	Redfield, Michael Collins, Ankur Parikh, Chris	Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei	1122
1066	Alberti, Danielle Epstein, Illia Polosukhin, Jacob	Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang	1123
1067	Devlin, Kenton Lee, Kristina Toutanova, Llion Jones,	Ren. 2020b. CommonGen: A constrained text	1124
1068	Matthew Kelcey, Ming-Wei Chang, Andrew M.	generation challenge for generative commonsense	1125
1069	Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov.	reasoning . In <i>Findings of the Association for</i>	1126
1070	2019. Natural questions: A benchmark for question	<i>Computational Linguistics: EMNLP 2020</i> , pages	1127
1071	answering research . <i>Transactions of the Association</i>	1823–1840, Online. Association for Computational	1128
1072	<i>for Computational Linguistics</i> , 7:452–466.	Linguistics.	1129
1073	Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and	Kevin Lin, Oyvind Tafjord, Peter Clark, and Matt	1130
1074	Eduard Hovy. 2017. RACE: Large-scale ReAding	Gardner. 2019. Reasoning over paragraph effects	1131
1075	comprehension dataset from examinations . In	in situations . In <i>Proceedings of the 2nd Workshop</i>	1132
1076	<i>Proceedings of the 2017 Conference on Empirical</i>	<i>on Machine Reading for Question Answering</i> ,	1133
1077	<i>Methods in Natural Language Processing</i> , pages	pages 58–62, Hong Kong, China. Association for	1134
1078	785–794, Copenhagen, Denmark. Association for	Computational Linguistics.	1135
1079	Computational Linguistics.	Wang Ling, Dani Yogatama, Chris Dyer, and Phil	1136
1080	Rémi Lebret, David Grangier, and Michael Auli. 2016.	Blunsom. 2017a. Program induction by rationale	1137
1081	Neural text generation from structured data with	generation: Learning to solve and explain algebraic	1138
1082	application to the biography domain . In <i>Proceedings</i>	word problems . In <i>Proceedings of the 55th</i>	1139
1083	<i>of the 2016 Conference on Empirical Methods</i>		

1140	<i>Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 158–167, Vancouver, Canada. Association for Computational Linguistics.	1197
1141		1198
1142		
1143		
1144	Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017b. Program induction by rationale generation: Learning to solve and explain algebraic word problems . In <i>Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 158–167, Vancouver, Canada. Association for Computational Linguistics.	1199
1145		1200
1146		1201
1147		1202
1148		1203
1149		
1150		
1151		
1152	Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In <i>Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures</i> , pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.	1204
1153		1205
1154		1206
1155		1207
1156		1208
1157		1209
1158		1210
1159		1211
1160	Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. <i>arXiv preprint arXiv:2007.08124</i> .	1212
1161		1213
1162		1214
1163		1215
1164		1216
1165	Do Long, Yiran Zhao, Hannah Brown, Yuxi Xie, James Zhao, Nancy Chen, Kenji Kawaguchi, Michael Shieh, and Junxian He. 2024. Prompt optimization via adversarial in-context learning . In <i>Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 7308–7327, Bangkok, Thailand. Association for Computational Linguistics.	1217
1166		1218
1167		1219
1168		1220
1169		1221
1170		1222
1171		
1172		
1173	Annie Louis, Dan Roth, and Filip Radlinski. 2020. “I’d rather just go to bed”: Understanding indirect answers . In <i>Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)</i> , pages 7411–7425, Online. Association for Computational Linguistics.	1223
1174		1224
1175		1225
1176		1226
1177		1227
1178		1228
1179	Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity . In <i>Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.	1229
1180		1230
1181		1231
1182		1232
1183		1233
1184		1234
1185		1235
1186		
1187	Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis . In <i>Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies</i> , pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.	1236
1188		1237
1189		1238
1190		1239
1191		1240
1192		1241
1193		1242
1194		1243
1195	Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts . <i>J. Assoc. Inf. Sci. Technol.</i> , 65(4):782–796.	1244
1196		1245
		1246
		1247
		1248
		1249
		1250
		1251
		1252
		1253

1254	<i>Human Language Technologies</i> , pages 2791–2809,	instructions with human feedback . <i>arXiv preprint</i>	1311
1255	Seattle, United States. Association for Computational	<i>arXiv:2203.02155</i> .	1312
1256	Linguistics.		
1257	Sewon Min, Xinxu Lyu, Ari Holtzman, Mikel	Jane Pan. 2023. <i>What In-Context Learning “Learns” In-</i>	1313
1258	Artetxe, Mike Lewis, Hannaneh Hajishirzi, and	<i>Context: Disentangling Task Recognition and Task</i>	1314
1259	Luke Zettlemoyer. 2022b. Rethinking the role of	<i>Learning</i> . Ph.D. thesis, Princeton University.	1315
1260	demonstrations: What makes in-context learning		
1261	work? In <i>Proceedings of the 2022 Conference on</i>	Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting	1316
1262	<i>Empirical Methods in Natural Language Processing</i> ,	class relationships for sentiment categorization	1317
1263	pages 11048–11064, Abu Dhabi, United Arab	with respect to rating scales . In <i>Proceedings of</i>	1318
1264	Emirates. Association for Computational Linguistics.	<i>the 43rd Annual Meeting of the Association for</i>	1319
		<i>Computational Linguistics (ACL’05)</i> , pages 115–124,	1320
		Ann Arbor, Michigan. Association for Computational	1321
1265	Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos,	Linguistics.	1322
1266	and Grigorios Tsoumakas. 2020. Ethos: an online		
1267	hate speech detection dataset . <i>ArXiv preprint</i> ,	Dimitris Pappas, Petros Stavropoulos, Ion Androu-	1323
1268	abs/2006.08328 .	sopoulos, and Ryan McDonald. 2020. BioMRC: A	1324
		dataset for biomedical machine reading comprehen-	1325
1269	Nikita Nangia, Clara Vania, Rasika Bhalerao, and	<i>sion</i> . In <i>Proceedings of the 19th SIGBioMed Work-</i>	1326
1270	Samuel R. Bowman. 2020. CrowS-pairs: A	<i>shop on Biomedical Language Processing</i> , pages	1327
1271	challenge dataset for measuring social biases in	140–149, Online. Association for Computational	1328
1272	masked language models . In <i>Proceedings of the</i>	Linguistics.	1329
1273	<i>2020 Conference on Empirical Methods in Natural</i>		
1274	<i>Language Processing (EMNLP)</i> , pages 1953–1967,	Arkil Patel, Satwik Bhattamishra, and Navin Goyal.	1330
1275	Online. Association for Computational Linguistics.	2021. Are NLP models really able to solve simple	1331
		math word problems? In <i>Proceedings of the 2021</i>	1332
1276	Courtney Napoles, Matthew Gormley, and Benjamin	<i>Conference of the North American Chapter of the</i>	1333
1277	Van Durme. 2012. Annotated Gigaword . In	<i>Association for Computational Linguistics: Human</i>	1334
1278	<i>Proceedings of the Joint Workshop on Automatic</i>	<i>Language Technologies</i> , pages 2080–2094, Online.	1335
1279	<i>Knowledge Base Construction and Web-scale Knowl-</i>	Association for Computational Linguistics.	1336
1280	<i>edge Extraction (AKBC-WEKEX)</i> , pages 95–100,		
1281	Montréal, Canada. Association for Computational	Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim	1337
1282	Linguistics.	Rocktäschel, Yuxiang Wu, Alexander H. Miller,	1338
		and Sebastian Riedel. 2020. How context affects	1339
1283	Shashi Narayan, Shay B. Cohen, and Mirella Lapata.	language models’ factual predictions . In <i>Automated</i>	1340
1284	2018. Don’t give me the details, just the	<i>Knowledge Base Construction</i> .	1341
1285	summary! topic-aware convolutional neural networks		
1286	for extreme summarization . In <i>Proceedings of the</i>	Fabio Petroni, Tim Rocktäschel, Sebastian Riedel,	1342
1287	<i>2018 Conference on Empirical Methods in Natural</i>	Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and	1343
1288	<i>Language Processing</i> , pages 1797–1807, Brussels,	Alexander Miller. 2019. Language models as	1344
1289	Belgium. Association for Computational Linguistics.	knowledge bases? In <i>Proceedings of the</i>	1345
		<i>2019 Conference on Empirical Methods in Natural</i>	1346
1290	Yixin Nie, Adina Williams, Emily Dinan, Mohit	<i>Language Processing and the 9th International</i>	1347
1291	Bansal, Jason Weston, and Douwe Kiela. 2020.	<i>Joint Conference on Natural Language Processing</i>	1348
1292	Adversarial NLI: A new benchmark for natural	<i>(EMNLP-IJCNLP)</i> , pages 2463–2473, Hong Kong,	1349
1293	language understanding . In <i>Proceedings of the 58th</i>	China. Association for Computational Linguistics.	1350
1294	<i>Annual Meeting of the Association for Computational</i>		
1295	<i>Linguistics</i> , pages 4885–4901, Online. Association	Mohammad Taher Pilehvar and Jose Camacho-	1351
1296	for Computational Linguistics.	Collados. 2019. WiC: the word-in-context	1352
		dataset for evaluating context-sensitive meaning	1353
1297	Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas	representations . In <i>Proceedings of the 2019</i>	1354
1298	Joseph, Nova DasSarma, Tom Henighan, Ben Mann,	<i>Conference of the North American Chapter of the</i>	1355
1299	Amanda Aspell, Yuntao Bai, Anna Chen, et al. 2022.	<i>Association for Computational Linguistics: Human</i>	1356
1300	In-context learning and induction heads. <i>arXiv</i>	<i>Language Technologies, Volume 1 (Long and Short</i>	1357
1301	<i>preprint arXiv:2209.11895</i> .	<i>Papers)</i> , pages 1267–1273, Minneapolis, Minnesota.	1358
		Association for Computational Linguistics.	1359
1302	A. Othman and M. Jemni. 2012. English-asl	Amir Pouran Ben Veyseh, Franck Dernoncourt,	1360
1303	gloss parallel corpus 2012: Aslg-pc12. In <i>5th</i>	Quan Hung Tran, and Thien Huu Nguyen. 2020.	1361
1304	<i>Workshop on the Representation and Processing of</i>	What does this acronym mean? introducing a new	1362
1305	<i>Sign Languages: Interactions between Corpus and</i>	dataset for acronym identification and disambigua-	1363
1306	<i>Lexicon LREC</i> .	tion . In <i>Proceedings of the 28th International</i>	1364
1307	Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida,	<i>Conference on Computational Linguistics</i> , pages	1365
1308	Carroll L Wainwright, Pamela Mishkin, Chong	3285–3301, Barcelona, Spain (Online). International	1366
1309	Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray,	Committee on Computational Linguistics.	1367
1310	et al. 2022. Training language models to follow		

- Chengwei Qin and Shafiq Joty. 2022. [Continual few-shot relation learning via embedding space regularization and data augmentation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2776–2789, Dublin, Ireland. Association for Computational Linguistics.
- Chengwei Qin, Aston Zhang, Chen Chen, Anirudh Dagar, and Wenming Ye. 2024. [In-context learning with iterative demonstration selection](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7441–7455, Miami, Florida, USA. Association for Computational Linguistics.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.
- Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. [Explain yourself! leveraging language models for commonsense reasoning](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.
- Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [SQuAD: 100,000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.
- Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.
- Sachin Ravi and Hugo Larochelle. 2017. [Optimization as a model for few-shot learning](#). In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. [Getting closer to ai complete question answering: A set of prerequisite real tasks](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8722–8731.
- Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. [Learning to retrieve prompts for in-context learning](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.
- Amrita Saha, Rahul Aralikkatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. [DuoRC: Towards complex language understanding with paraphrased reading comprehension](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1693, Melbourne, Australia. Association for Computational Linguistics.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. [Winogrande: An adversarial winograd schema challenge at scale](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8732–8740.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. [Social IQa: Commonsense reasoning about social interactions](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.
- Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. [CARER: Contextualized affect representations for emotion recognition](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Emily Sheng and David Uthus. 2020. [Investigating societal biases in a poetry composition system](#). In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 93–106, Barcelona, Spain (Online). Association for Computational Linguistics.
- Weijia Shi, Sewon Min, Maria Lomeli, Chunting Zhou, Margaret Li, Xi Victoria Lin, Noah A. Smith, Luke

1483	Zettlemoyer, Wen tau Yih, and Mike Lewis. 2024.	for “what if...” reasoning over procedural text. In	1540
1484	In-context pretraining: Language modeling beyond	<i>Proceedings of the 2019 Conference on Empirical</i>	1541
1485	document boundaries . In <i>The Twelfth International</i>	<i>Methods in Natural Language Processing and the 9th</i>	1542
1486	<i>Conference on Learning Representations</i> .	<i>International Joint Conference on Natural Language</i>	1543
		<i>Processing (EMNLP-IJCNLP)</i> , pages 6076–6085,	1544
1487	Damien Sileo, Tim Van De Cruys, Camille Pradel,	Hong Kong, China. Association for Computational	1545
1488	and Philippe Muller. 2019. Mining discourse	Linguistics.	1546
1489	markers for unsupervised sentence representation		
1490	learning . In <i>Proceedings of the 2019 Conference</i>	TheBloke. 2023. Thebloke/llama-2-70b-gptq: Gptq	1547
1491	<i>of the North American Chapter of the Association for</i>	model for meta’s llama 2 70b .	1548
1492	<i>Computational Linguistics: Human Language Tech-</i>		
1493	<i>nologies, Volume 1 (Long and Short Papers)</i> , pages	James Thorne, Andreas Vlachos, Christos	1549
1494	3477–3486, Minneapolis, Minnesota. Association for	Christodoulopoulos, and Arpit Mittal. 2018.	1550
1495	Computational Linguistics.	FEVER: a large-scale dataset for fact extraction	1551
		and VERification . In <i>Proceedings of the 2018</i>	1552
1496	Richard Socher, Alex Perelygin, Jean Wu, Jason	<i>Conference of the North American Chapter of</i>	1553
1497	Chuang, Christopher D. Manning, Andrew Ng, and	<i>the Association for Computational Linguistics:</i>	1554
1498	Christopher Potts. 2013. Recursive deep models for	<i>Human Language Technologies, Volume 1 (Long</i>	1555
1499	semantic compositionality over a sentiment treebank .	<i>Papers)</i> , pages 809–819, New Orleans, Louisiana.	1556
1500	In <i>Proceedings of the 2013 Conference on Empirical</i>	Association for Computational Linguistics.	1557
1501	<i>Methods in Natural Language Processing</i> , pages		
1502	1631–1642, Seattle, Washington, USA. Association	Hugo Touvron, Louis Martin, Kevin Stone, Peter	1558
1503	for Computational Linguistics.	Albert, Amjad Almahairi, Yasmine Babaei, Nikolay	1559
		Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti	1560
1504	Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi,	Bhosale, et al. 2023. Llama 2: Open foundation	1561
1505	and Claire Cardie. 2019. DREAM: A challenge	and fine-tuned chat models. <i>arXiv preprint</i>	1562
1506	data set and models for dialogue-based reading	<i>arXiv:2307.09288</i> .	1563
1507	comprehension . <i>Transactions of the Association for</i>		
1508	<i>Computational Linguistics</i> , 7:217–231.	Eleni Triantafillou, Richard Zemel, and Raquel Urtasun.	1564
		2017. Few-shot learning through an information	1565
1509	Mirac Suzgun, Nathan Scales, Nathanael Schärli,	retrieval lens. <i>arXiv preprint arXiv:1707.02610</i> .	1566
1510	Sebastian Gehrmann, Yi Tay, Hyung Won Chung,		
1511	Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny	Sowmya Vajjala and Ivana Lučić. 2018. On-	1567
1512	Zhou, et al. 2022. Challenging big-bench tasks and	eStopEnglish corpus: A new corpus for automatic	1568
1513	whether chain-of-thought can solve them. <i>arXiv</i>	readability assessment and text simplification . In <i>Pro-</i>	1569
1514	<i>preprint arXiv:2210.09261</i> .	<i>ceedings of the Thirteenth Workshop on Innovative</i>	1570
		<i>Use of NLP for Building Educational Applications</i> ,	1571
1515	Oyvind Tafjord, Peter Clark, Matt Gardner, Wen-tau	pages 297–304, New Orleans, Louisiana. Association	1572
1516	Yih, and Ashish Sabharwal. 2019a. Quarel: A dataset	for Computational Linguistics.	1573
1517	and models for answering questions about qualitative		
1518	relationships . <i>Proceedings of the AAAI Conference</i>	Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang,	1574
1519	<i>on Artificial Intelligence</i> , 33(01):7063–7071.	Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu,	1575
		Huaming Wang, Jinyu Li, et al. 2023a. Neural	1576
1520	Oyvind Tafjord, Matt Gardner, Kevin Lin, and	codec language models are zero-shot text to speech	1577
1521	Peter Clark. 2019b. QuaRTz: An open-domain	synthesizers. <i>arXiv preprint arXiv:2301.02111</i> .	1578
1522	dataset of qualitative relationship questions . In		
1523	<i>Proceedings of the 2019 Conference on Empirical</i>	Liang Wang, Nan Yang, and Furu Wei. 2024. Learning	1579
1524	<i>Methods in Natural Language Processing and the 9th</i>	to retrieve in-context examples for large language	1580
1525	<i>International Joint Conference on Natural Language</i>	models . In <i>Proceedings of the 18th Conference</i>	1581
1526	<i>Processing (EMNLP-IJCNLP)</i> , pages 5941–5946,	<i>of the European Chapter of the Association for</i>	1582
1527	Hong Kong, China. Association for Computational	<i>Computational Linguistics (Volume 1: Long Papers)</i> ,	1583
1528	Linguistics.	pages 1752–1767, St. Julian’s, Malta. Association	1584
		for Computational Linguistics.	1585
1529	Alon Talmor, Jonathan Herzig, Nicholas Lourie, and	William Yang Wang. 2017. “liar, liar pants on fire”:	1586
1530	Jonathan Berant. 2019. CommonsenseQA: A	A new benchmark dataset for fake news detection .	1587
1531	question answering challenge targeting common-	In <i>Proceedings of the 55th Annual Meeting of the</i>	1588
1532	sense knowledge . In <i>Proceedings of the 2019</i>	<i>Association for Computational Linguistics (Volume 2:</i>	1589
1533	<i>Conference of the North American Chapter of the</i>	<i>Short Papers)</i> , pages 422–426, Vancouver, Canada.	1590
1534	<i>Association for Computational Linguistics: Human</i>	Association for Computational Linguistics.	1591
1535	<i>Language Technologies, Volume 1 (Long and Short</i>		
1536	<i>Papers)</i> , pages 4149–4158, Minneapolis, Minnesota.	Xinyi Wang, Wanrong Zhu, and William Yang	1592
1537	Association for Computational Linguistics.	Wang. 2023b. Large language models are	1593
		implicitly topic models: Explaining and finding	1594
1538	Niket Tandon, Bhavana Dalvi, Keisuke Sakaguchi, Peter	good demonstrations for in-context learning. <i>arXiv</i>	1595
1539	Clark, and Antoine Bosselut. 2019. WIQA: A dataset	<i>preprint arXiv:2301.11916</i> .	1596

1597	Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. <i>arXiv preprint arXiv:2212.10560</i> .	1652
1598		1653
1599		1654
1600		1655
1601		1656
1602	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> .	1657
1603		1658
1604		1659
1605		1660
1606		
1607	Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English . <i>Transactions of the Association for Computational Linguistics</i> , 8:377–392.	1661
1608		1662
1609		1663
1610		1664
1611		1665
1612		1666
1613	Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments . <i>Transactions of the Association for Computational Linguistics</i> , 7:625–641.	1667
1614		1668
1615		1669
1616		1670
1617	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. Emergent abilities of large language models . <i>Transactions on Machine Learning Research</i> . Survey Certification.	1671
1618		1672
1619		
1620		1673
1621		1674
1622		1675
1623		1676
1624		1677
1625	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models . In <i>Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS 2022)</i> .	1678
1626		1679
1627		1680
1628		1681
1629		
1630		1682
1631	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022c. Chain-of-thought prompting elicits reasoning in large language models. <i>Advances in Neural Information Processing Systems</i> , 35:24824–24837.	1683
1632		1684
1633		1685
1634		1686
1635		1687
1636		1688
1637	Jerry Wei, Le Hou, Andrew Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, et al. 2023a. Symbol tuning improves in-context learning in language models. <i>arXiv preprint arXiv:2305.08298</i> .	1689
1638		1690
1639		1691
1640		1692
1641		1693
1642	Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. 2023b. Larger language models do in-context learning differently. <i>arXiv preprint arXiv:2303.03846</i> .	1694
1643		1695
1644		1696
1645		1697
1646		1698
1647	Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions . In <i>Proceedings of the 3rd Workshop on Noisy User-generated Text</i> , pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.	1699
1648		1700
1649		1701
1650		1702
1651		1703
	Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference . 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.	1704
		1705
		1706
		1707
		1708
		1709
	Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, 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. Transformers: State-of-the-art natural language processing . 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.	
	Tomer Wolfson, Mor Geva, Ankit Gupta, Matt Gardner, Yoav Goldberg, Daniel Deutch, and Jonathan Berant. 2020. Break it down: A question understanding benchmark . <i>Transactions of the Association for Computational Linguistics</i> , 8:183–198.	
	Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference . In <i>International Conference on Learning Representations</i> .	
	Wenhan Xiong, Jiawei Wu, Hong Wang, Vivek Kulkarni, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2019. TWEETQA: A social media focused question answering dataset . In <i>Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics</i> , pages 5020–5031, Florence, Italy. Association for Computational Linguistics.	
	Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023a. Wizardlm: Empowering large language models to follow complex instructions. <i>arXiv preprint arXiv:2304.12244</i> .	
	Yiheng Xu, Hongjin Su, Chen Xing, Boyu Mi, Qian Liu, Weijia Shi, Binyuan Hui, Fan Zhou, Yitao Liu, Tianbao Xie, et al. 2023b. Lemur: Harmonizing natural language and code for language agents. <i>arXiv preprint arXiv:2310.06830</i> .	
	An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report . <i>arXiv preprint arXiv:2412.15115</i> .	
	Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering . In <i>Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing</i> , pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.	

1710	Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua	Florence, Italy. Association for Computational	1768
1711	Bengio, William Cohen, Ruslan Salakhutdinov, and	Linguistics.	1769
1712	Christopher D. Manning. 2018. HotpotQA: A		
1713	dataset for diverse, explainable multi-hop question		
1714	answering . In <i>Proceedings of the 2018 Conference</i>	Hanlin Zhang, YiFan Zhang, Yaodong Yu, Dhruv	1770
1715	<i>on Empirical Methods in Natural Language</i>	Madeka, Dean Foster, Eric Xing, Himabindu	1771
1716	<i>Processing</i> , pages 2369–2380, Brussels, Belgium.	Lakkaraju, and Sham Kakade. 2024. A study on the	1772
1717	Association for Computational Linguistics.	calibration of in-context learning . In <i>Proceedings of</i>	1773
		<i>the 2024 Conference of the North American Chapter</i>	1774
1718	Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021.	<i>of the Association for Computational Linguistics:</i>	1775
1719	CrossFit: A few-shot learning challenge for cross-	<i>Human Language Technologies (Volume 1: Long</i>	1776
1720	task generalization in NLP . In <i>Proceedings of the</i>	<i>Papers)</i> , pages 6118–6136, Mexico City, Mexico.	1777
1721	<i>2021 Conference on Empirical Methods in Natural</i>	Association for Computational Linguistics.	1778
1722	<i>Language Processing</i> , pages 7163–7189, Online and		
1723	Punta Cana, Dominican Republic. Association for	Hao Zhang, Jae Ro, and Richard Sproat. 2020. Semi-	1779
1724	Computational Linguistics.	supervised URL segmentation with recurrent neural	1780
		networks pre-trained on knowledge graph entities . In	1781
1725	Xi Ye, Srinivasan Iyer, Asli Celikyilmaz, Veselin	<i>Proceedings of the 28th International Conference</i>	1782
1726	Stoyanov, Greg Durrett, and Ramakanth Pasunuru.	<i>on Computational Linguistics</i> , pages 4667–4675,	1783
1727	2023. Complementary explanations for effective in-	Barcelona, Spain (Online). International Committee	1784
1728	context learning . In <i>Findings of the Association</i>	on Computational Linguistics.	1785
1729	<i>for Computational Linguistics: ACL 2023</i> , pages		
1730	4469–4484, Toronto, Canada. Association for	Rui Zhang and Joel Tetreault. 2019. This email	1786
1731	Computational Linguistics.	could save your life: Introducing the task of	1787
		email subject line generation . In <i>Proceedings of</i>	1788
1732	Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim,	<i>the 57th Annual Meeting of the Association for</i>	1789
1733	Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-	<i>Computational Linguistics</i> , pages 446–456, Florence,	1790
1734	goo Lee, and Taeuk Kim. 2022. Ground-truth	Italy. Association for Computational Linguistics.	1791
1735	labels matter: A deeper look into input-label		
1736	demonstrations . In <i>Proceedings of the 2022</i>	Sheng Zhang, X. Liu, J. Liu, Jianfeng Gao, Kevin	1792
1737	<i>Conference on Empirical Methods in Natural</i>	Duh, and Benjamin Van Durme. 2018. Record:	1793
1738	<i>Language Processing</i> , pages 2422–2437, Abu	Bridging the gap between human and machine	1794
1739	Dhabi, United Arab Emirates. Association for	commonsense reading comprehension . <i>ArXiv</i>	1795
1740	Computational Linguistics.	<i>preprint</i> , abs/1810.12885.	1796
1741	Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga,	Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015.	1797
1742	Dongxu Wang, Zifan Li, James Ma, Irene Li,	Character-level convolutional networks for text	1798
1743	Qingning Yao, Shanelle Roman, Zilin Zhang,	classification. <i>Advances in neural information</i>	1799
1744	and Dragomir Radev. 2018. Spider: A large-	<i>processing systems</i> , 28:649–657.	1800
1745	scale human-labeled dataset for complex and cross-		
1746	domain semantic parsing and text-to-SQL task . In	Yuan Zhang, Jason Baldridge, and Luheng He.	1801
1747	<i>Proceedings of the 2018 Conference on Empirical</i>	2019. PAWS: Paraphrase adversaries from	1802
1748	<i>Methods in Natural Language Processing</i> , pages	word scrambling . In <i>Proceedings of the 2019</i>	1803
1749	3911–3921, Brussels, Belgium. Association for	<i>Conference of the North American Chapter of the</i>	1804
1750	Computational Linguistics.	<i>Association for Computational Linguistics: Human</i>	1805
		<i>Language Technologies, Volume 1 (Long and Short</i>	1806
1751	Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang,	<i>Papers)</i> , pages 1298–1308, Minneapolis, Minnesota.	1807
1752	Songfang Huang, and Fei Huang. 2023. Rrhf:	Association for Computational Linguistics.	1808
1753	Rank responses to align language models with		
1754	human feedback without tears . <i>arXiv preprint</i>	Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex	1809
1755	<i>arXiv:2304.05302</i> .	Smola. 2023. Automatic chain of thought	1810
		prompting in large language models . In <i>The</i>	1811
1756	Rowan Zellers, Yonatan Bisk, Roy Schwartz, and	<i>Eleventh International Conference on Learning</i>	1812
1757	Yejin Choi. 2018. SWAG: A large-scale	<i>Representations</i> .	1813
1758	adversarial dataset for grounded commonsense	Anhao Zhao, Fanghua Ye, Jinlan Fu, and Xiaoyu Shen.	1814
1759	inference . In <i>Proceedings of the 2018 Conference on</i>	2024. Unveiling in-context learning: A coordinate	1815
1760	<i>Empirical Methods in Natural Language Processing</i> ,	system to understand its working mechanism. <i>arXiv</i>	1816
1761	pages 93–104, Brussels, Belgium. Association for	<i>preprint arXiv:2407.17011</i> .	1817
1762	Computational Linguistics.		
		Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein,	1818
1763	Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali	and Sameer Singh. 2021. Calibrate before use:	1819
1764	Farhadi, and Yejin Choi. 2019. HellaSwag: Can a	Improving few-shot performance of language models .	1820
1765	machine really finish your sentence? In <i>Proceedings</i>	In <i>Proceedings of the 38th International Conference</i>	1821
1766	<i>of the 57th Annual Meeting of the Association</i>	<i>on Machine Learning</i> , volume 139 of <i>Proceedings</i>	1822
1767	<i>for Computational Linguistics</i> , pages 4791–4800,	<i>of Machine Learning Research</i> , pages 12697–12706.	1823
		PMLR.	1824

- Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. Can we edit factual knowledge by in-context learning? *arXiv preprint arXiv:2305.12740*.
- Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language usin. *ArXiv preprint, abs/1709.00103*.
- Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. “going on a vacation” takes longer than “going for a walk”: A study of temporal commonsense understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3363–3369, Hong Kong, China. Association for Computational Linguistics.
- Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large 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 θ_T , student model θ_S , number of subsets N , weight of ranking loss λ

- 1: **for** mini-batch \mathcal{B} in \mathcal{D}_{ICL} **do**
- 2: CALCULATE the KL divergence loss \mathcal{L}^{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 θ_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 Details of Splitting Groups by Similarity

We use Sentence-BERT (Reimers and Gurevych, 2019) to obtain contextual representations of the examples and employ cosine similarity to measure the similarity between these representations. Based on the similarity to the test example, we categorize all demonstrations into two groups, G_{sim} and G_{dissim} , ensuring an approximately equal split between the two groups (*i.e.*, a 1:1 ratio).

A.2 Algorithm

The learning process of BiAlign is illustrated in Algorithm 1.

A.3 Details of Source Tasks

We report the full list of source tasks used in our work in Table 22. All tasks are taken from CrossFit (Ye et al., 2021).

A.4 Details of Target Tasks

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

	KATE	MMR	IDS
Selection _{Vanilla}	18.1	17.4	19.2
Selection _{BiAlign}	20.2	19.3	20.8

Table 8: Integration of BiAlign with ICL demonstration selection methods.

	Output-Align	BiAlign
Llama 3-8B	62.5	65.3

Table 9: Integration of BiAlign with reverse KL divergence.

- **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 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.5 Implementation Details

Our methods are implemented with the PyTorch and Transformers library (Wolf et al., 2020). Model

	1.5B	3B	7B	14B
Output-Align	35.2	35.9	36.2	36.7
BiAlign	36.9	38.0	38.8	40.1

Table 10: Results for different teacher model sizes with a fixed 0.5B student (Qwen-2.5).

	0.5B	1.5B	3B	7B
Output-Align	36.7	51.4	59.6	70.9
BiAlign	40.1	54.3	62.7	73.4

Table 11: Results for different student model sizes with a fixed 14B teacher (Qwen-2.5).

	Vanilla	Output-Align	BiAlign
BBH	31.4	33.8	45.3
GSM8K	24.7	28.4	38.6
LogiQA	29.1	32.3	44.7

Table 12: Average preference score consistency (%) comparison between different methods.

	Output-Align	BiAlign
BBH	40.2	43.3

Table 13: Performance on BBH for models trained on MMLU.

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 λ 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.

A.6 Combination with ICL Demonstration Selection Methods

BiAlign is complementary to ICL demonstration selection methods and can be seamlessly integrated with them to further improve ICL performance. To validate this, we investigate three demonstration selection methods: KATE (Liu et al., 2022), MMR (Ye et al., 2023), and IDS (Qin et al., 2024). For each method, we evaluate the following two variants: selecting demonstrations and performing ICL using the vanilla model (Selection_{Vanilla}), and selecting demonstrations and performing ICL using the model after BiAlign (Selection_{BiAlign}). We conduct experiments on GSM8K and report the results in Table 8, demonstrating that BiAlign consistently boosts performance across all three selection methods. Furthermore, BiAlign (18.3) surpasses both KATE (18.1) and MMR (17.4), highlighting its superiority over several ICL demonstration selection approaches.

A.7 Combination with Reverse KL Divergence

Gu et al. (2024) reveals that reverse KL divergence is more suitable for knowledge distillation in generative LLMs, as it helps prevent the student model from overestimating low-probability regions of the teacher’s distribution. Building on this insight, we investigate the integration of BiAlign

with reverse KL divergence. Specifically, we replace the forward KL divergence in both Output-Align and BiAlign with reverse KL divergence and conduct experiments using Llama 3-70B as the teacher model and Llama 3-8B as the student model. As shown in Table 9, BiAlign continues to significantly outperform Output-Align with reverse KL divergence, further demonstrating its effectiveness.

A.8 Scalability to More Model Scales

We investigate the scalability of BiAlign across different model scales using Qwen-2.5 (Yang et al., 2024). Specifically, we conduct experiments on five model scales: 0.5B, 1.5B, 3B, 7B, and 14B. Our evaluation follows two settings: (i) *Varying teacher model sizes*: We fix the student model at 0.5B and experiment with teacher models ranging from 1.5B to 14B. (ii) *Varying student model sizes*: We fix the teacher model at 14B and test student models ranging from 0.5B to 7B. The results for both settings are presented in Table 10 and 11, respectively. We can see that:

- BiAlign consistently outperforms Output-Align with different sizes of teacher models.
- BiAlign *benefits more from increasing the size of the teacher model* compared to Output-Align.
- BiAlign is both applicable and robust across student models of different sizes.

A.9 Average Preference Score Consistency

We report the average preference score consistency (%) comparison between different methods on the other three datasets (BBH, GSM8K and LogiQA) in Table 12. From the results, we can see that BiAlign consistently outperforms Output-Align across all datasets.

	Default	Variant
BiAlign	31.0	30.3

Table 14: Comparison between different subset sampling methods.

Method	Subset number			
	3	4	5	6
BiAlign	30.7	31.0	30.8	31.1

Table 15: Average performance (%) of BiAlign with different numbers of subsets N .

A.10 Further Justification of Input Preference Alignment

We outline the justification for input preference alignment from the following perspectives.

(i) *Impact of ICL Demonstrations on Model Performance.* ICL demonstrations have been shown to have a significant impact on the performance of ICL (Liu et al., 2022; Qin et al., 2024). Selecting different sets of demonstration examples can yield performance ranging from almost random to better than state-of-the-art fine-tuned models, indicating that the model has different preferences for different inputs. 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 model needs to learn “what to output” as well as “which input examples should be preferred”**. This 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. To this end, we introduce input preference alignment to align the student and teacher models’ preferences for different demonstrations.

(ii) *Explanatory Mechanisms of ICL.* Another perspective supporting input preference alignment stems from the way LLMs process and prioritize information during ICL. Kossen et al. (2024) discover that LLMs do not treat all available information equally; instead, they exhibit a natural tendency to prioritize information closer to the query. This selective attention mechanism suggests that LLMs inherently favor contextually relevant details over more distant or less relevant ones.

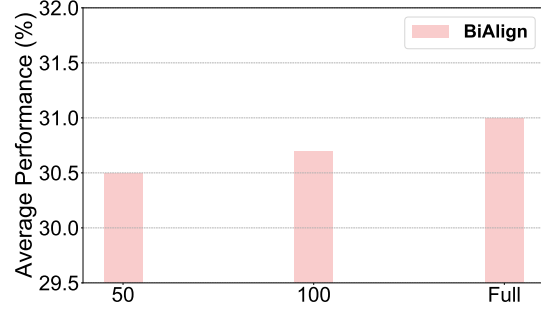


Figure 5: Average performance (%) of BiAlign with different numbers of source tasks.

Method	Type	
	Whole Sequence	Label Only
BiAlign	31.0	30.8

Table 16: Average performance (%) of BiAlign using different types of KL divergence calculation methods.

Building on this insight, our proposed input preference alignment ensures that **the student model learns to replicate the teacher model’s information prioritization strategy**. By aligning the student’s input selection process with that of the teacher, we make the learning process more effective. This joint alignment ultimately enables the student model to utilize information in a manner consistent with the teacher model’s intrinsic preferences, thereby improving its overall ICL performance.

A.11 More Cross-Task Generalization Experiments

To further verify the cross-task generalization ability of BiAlign, we train the model on MMLU and evaluate it on BBH. Specifically, we use Llama 2-7B as the student model and Llama 2-70B as the teacher model. The results reported in Table 13 highlight the superiority of BiAlign over Output-Align.

A.12 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 14. We can observe a noticeable performance drop, highlighting the crucial role of incorporating example similarity in the sampling process.

Method	25%	50%	100%
Output-Align	29.1	29.3	29.4
BiAlign	30.3	30.8	31.0

Table 17: Comparison between BiAlign and Output-Align at different proportions of training steps.

A.13 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 15. 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.14 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, 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.15 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. (2024). In Table 16, 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.16 Different Proportions of Training Steps

Table 17 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

Method	λ				
	0.2	0.5	1.0	2.0	5.0
BiAlign	30.8	31.2	31.0	30.9	29.9

Table 18: Average performance (%) of BiAlign with different λ for the ranking loss $\mathcal{L}^{\text{rank}}$.

Method	Pair number			
	3	4	5	All (6)
BiAlign	30.2	30.8	30.7	31.0

Table 19: Average results (%) of BiAlign with different numbers of contrastive pairs.

steps.

A.17 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.18 Ranking Loss Weights

To further investigate the influence of the ranking loss $\mathcal{L}^{\text{rank}}$ (Equation 3), we conduct experiments with different weights λ and report the results in Table 18. All variants except the variant with $\lambda = 5.0$ (too large) outperform baseline approaches by a large margin, which demonstrates the superiority of $\mathcal{L}^{\text{rank}}$.

A.19 Contrastive Pair Selection

While we use all $C(N, 2)$ ($N = 4$ is the number of subsets) pairs of positive and negative subsets for input preference alignment, we also investigate the effect of contrastive pair selection. Specifically, we conduct controlled experiments on $\{3, 4, 5\}$ randomly selected contrastive pairs and report the average results in Table 19. The best performance is achieved using all pairs, justifying our selection strategy.

A.20 Case Study

We select GSM8K as a representative task and show several examples of output in Table 20.

Compared with Vallina, BiAlign is able to generate more precise and fine-grained reasoning paths, *e.g.*, BiAlign can successfully understand the meaning of ‘*pair*’ and generate the rationale ‘*10 pairs of socks is $10 * 2 = 20$ socks*’ while Vallina fails to do so.

A.21 Example in LogiQA

We show the reasoning path required to solve a sample in LogiQA in Table 21, which is quite fine-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 20: 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 21: 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	(Pouran Ben Veyseh 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)
samsum	(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)
crowds_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 22: List of all source tasks.