Improving Small and Large Language Models Alignment on Chain-of-Thought Reasoning using Curriculum Learning

Anonymous ACL submission

Abstract

Chain-of-Thought (CoT) prompting empowers the reasoning abilities of Large Language Models (LLMs), eliciting them to solve complex reasoning tasks step-by-step. However, these capabilities appear only in models with billions of parameters, which represent a barrier to entry for many users who are forced to operate on a smaller model scale, i.e., Small Language Models (SLMs). Although many companies are releasing LLMs of the same family with a reduced number of parameters, these models sometimes produce misleading answers and are unable to deliver CoT reasoning. In this paper, we propose a method to enable CoT reasoning over SLMs by introducing two novel mechanisms. First, we propose aligning CoT abilities via Instruction-tuning with the support of CoT Demonstrations "taught" by LLMs teacher to SLMs students. Second, we use Curriculum Learning, a pedagogically motivated learning method that empowers the Instruction-tuning phase. Hence, we analyze the impact on the downstream abilities of four question-answering benchmarks. The results show that SMLs can be instructed to reason via Demonstration produced by LLMs. We move a step further in research: conceiving SLMs as human learners, we expose them to a CL teaching-based approach, obtaining better results on downstream performances.

1 Introduction

Chain-of-Thought (CoT) prompting enables Large Language Models (LLMs) to deliver multi-step, controlled reasoning (Kojima et al., 2023; Wei et al., 2022), achieving outstanding results in commonsense (Bubeck et al., 2023), symbolic (Gaur and Saunshi, 2023), and mathematical (Liu et al., 2023) reasoning tasks. LLMs achieve all these results with at least several billions of parameters, such as GPTs family (OpenAI, 2023), PaLM (Chowdhery et al., 2022), Llama-2-70b (Touvron et al., 2023) and Mistral (MistralAI, 2023).

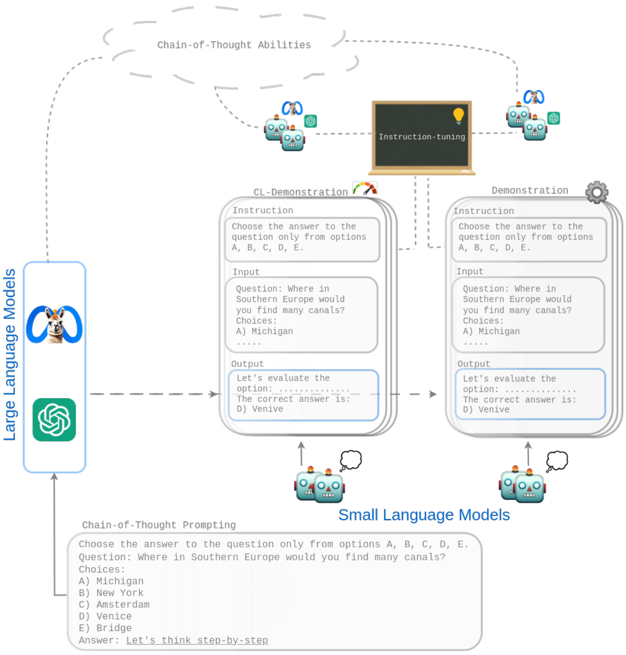
Chain-of-Thought Abilities
Instruction-tuning
CL-Demonstration
Demonstration
Instruction
Choose the answer to the question only from options A, B, C, D, E.
Input
Question: Where in Southern Europe would you find many canals? Choices: A) Michigan .....
Output
Let's evaluate the option: .... The correct answer is: D) Venive
Instruction
Choose the answer to the question only from options A, B, C, D, E.
Input
Question: Where in Southern Europe would you find many canals? Choices: A) Michigan .....
Output
Let's evaluate the option: .... The correct answer is: D) Venive
Large Language Models
Small Language Models
Chain-of-Thought Prompting
Choose the answer to the question only from options A, B, C, D, E.
Question: Where in Southern Europe would you find many canals?
Choices:
A) Michigan
B) New York
C) Amsterdam
D) Venice
E) Bridge
Answer: Let's think step-by-step

Figure 1: In Instruction-tuning, the smaller models instruct themselves using the reasoning generated by the larger models. In a zero-shot scenario, we elicit a larger model to answer complex questions through Chain-of-Thought reasoning. Moreover, we evaluate the reasoning chain using Curriculum Learning metrics to facilitate the instruction phase and expose the Demonstrations in a meaningful way.

Small Language Models (SLMs) seem to use Chain-of-Thought (CoT) prompting in a less effective way. Although these models are highly functional across different tasks, the CoT prompting mechanism only proved beneficial for models at a certain threshold scale (e.g., with more than 60B parameters (Wei et al., 2023)). SLMs are crucial in fostering research since these are smaller versions of LLMs that are often open-source and accessible to most researchers, e.g., Llama-2-7b and Llama-2-13b (Touvron et al., 2023). Yet, these SLMs produce illogical answers when prompted under the CoT framework.

In this work, we propose an approach to align the reasoning abilities of the SMLs (students) with the LLMs (teachers) via Instruction-tuning-CoT, that is, an instruction tuning over CoT Demonstrations delivered from larger models (see Figure 1). With respect to the fundation teacher-student approach (Magister et al., 2023; Ho et al., 2023a; Li et al., 2023), we move a step further by introducing the Instruction-tuning via CoT, and, with respect to (Ranaldi and Freitas, 2024; Paul et al., 2024), we improve the strategies to expose the student to examples in a reasonable, pedagogically-motivated order using Curriculum Learning (Bengio et al., 2009). Hence, starting from the idea that humans acquire first elemental concepts and then, gradually, more complex ones, Bengio et al. (2009) proposed Curriculum Learning (CL) and demonstrated its benefits in several tasks. We adopt this idea to reorder the Instruction-tuning Demonstrations in a meaningful way. In particular, we evaluate the reasoning chains that are answers delivered by teachers via CoT prompting to elicit student learning.

This leads to the target research questions, which are the focus of this paper:

RQ1) How does Instruction-tuning via Demonstrations impact the reasoning abilities of students models? And what is the effect of Demonstrations delivered with the Chain-of-Thought reasoning process?

RQ2) How important is reasoning chain valuation to facilitate the presentation of demonstrations during Instruction-tuning?

To answer these questions, we select Llama-2-7b, Llama-2-13b (Touvron et al., 2023) as students and Llama-2-70b, and GPT-3.5 as teachers. Hence, we conduct an extensive analysis using four question-answering benchmarks. We use Llama-2-70 and GPT-3.5 to deliver Answers at the core of the Demonstrations (see Fig. 1) to instruct Llama-2-7 and -13. Furthermore, in order to expose the students to Demonstrations delivered by teachers we evaluate the complexity of the reasoning chains present in CoT Answers. Hence, we propose a metric based on informativeness comprehensibility used as a pivot in the Instruction-tuning phase.

Behind a comprehensive analysis, we show that the Instruction-tuning approach on Demonstrations instructs students, and they outperform baseline SLMs in all proposed benchmarks. Moreover, the students exposed to the Demonstrations via the CL approach outperformed students instructed via non-CL.

Our findings can be summarized as follows:

i) The Instruction-tuning of SLM students via Demonstrations delivered by an LLM teacher outperformed the baselines in terms of downstream performance. The SLMs instructed via Demonstrations consistently outperformed the baselines defined by non-tuned SLMs on the four proposed question-answering benchmarks.

ii) The CL-based Instruction-tuning approach outperforms standard Instruction-tuning. Llama-2-7 and Llama-2-13, instructed via the CL method, outperform the instructed models without CL.

iii) Finally, the CL method favors the alignment of CoT abilities within the family. In fact, Llama-2-7 and Llama-2-13 exposed to CL Demonstrations produced by Llama-2-70 outperform students instructed by GPT-3.5 teachers in other SMLs as well.

2 Method

In order to align the reasoning abilities of smaller Language Models using further knowledge generated by larger Language Models, we propose three steps, as shown in Figure 1. In the first part, there is an annotation phase where the Large Language Models (LLMs) systematically prompt generate outputs (Section 2.1). The outputs will be the core of Demonstrations used during the Instruction-tuning phase from the smaller Language Models, presented in Section 2.2. However, the Curriculum Learning approach is behind the Instruction-tuning phase, where the Demonstrations are reorganized following our measure introduced in Section 2.3.

2.1 Teacher Model

As teacher model, we selected the largest Llama version (Touvron et al., 2023), that is, Llama-2-70b, and in terms of comparison, GPT-3.5[1] (OpenAI, 2023). We selected GPT-3.5 because it generates high-quality data with and without the CoT prompting approach, as shown in (Fu et al., 2023). Meanwhile, Llama-2-70b because it has smaller versions that can be used as students of the same family (presented in Section 2.2), and these smaller versions obtain remarkable results despite the reduced number of parameters. In particular, as teacher model, we used the "chat" version of the LLM called Llama-2-70-chat. We selected this version because, as reported by Touvron et al. (2023), it

[1] We use *GPT-3.5-turbo*, however in the rest of work we will use only GPT-3.5

is optimized for dialogue use cases and provides better demonstrations. In the rest of the paper, we will call this model Llama-2-70.

Hence, we proposed the following input-prompt in a zero-shot scenario:

```
Choose the answer to the question only from
options A, B, C, D.
Question: <Question>
Choices:
A) <Option1>
B) <Option2>
C) <Option3>
D) <Option4>
Answer: Let's think step by step
```

Input prompts have a generic structure, but behind "Answer:" we insert the formula "**Let's think step by step**" as done in (Kojima et al., 2023; Wei et al., 2022), that is shown in Table 6.

Following the annotation process performed by LLMs, the answers generated by teachers models that are the annotations have been used to construct the Demonstrations (see Table 1).

2.2 Student Model

Several SLMs have been fine-tuned for instruction-following (Taori et al., 2023) and reinforcement learning with human feedback (Ouyang et al., 2022). However, whatever the techniques, the smaller Language Models[2] do not seem able to reproduce the step-by-step reasoning abilities.

Recent work proposes techniques of knowledge distillation (Li et al., 2023), skill-refinement (Huang et al., 2022), and enriched fine-tuning (Magister et al., 2023) to teach SLMs to produce reasoned answers. Following this idea, we analyzed the impact of Instruction-tuning from Demonstrations provided by different teachers. Hence, we chose two SLMs students, Llama-2-7b and Llama-2-13b (Touvron et al., 2023). In both cases, we used "chat" versions, i.e., Llama-2-7-chat and Llama-2-13-chat (versions detailed in Table 7), which we will refer to in the paper as Llama-2-7 and -13.

Student models are evaluated before and after the Instruction-tuning phases. In particular, we propose the standard approach where the students are instructed via Demonstrations delivered arbitrarily and in the Curriculum Learning manner (introduced in Section 2.3). The structure of Instruction-tuning stems from Alpaca (Taori et al., 2023), which uses Demonstrations consisting of an instruction that, in our case, is fixed, i.e., `Choose the answer to the question only from options A, B, C, D.`, an input which is the question, and an expected output which, in our case, are the output generated by the LLMs teachers. Table 1 shows an example of input. Additional details about the Instruction-tuning steps are provided in Section 3.2.1.

Instruction

```
Choose the answer to the question only from
options A, B, C, D.
```

Input

```
Question: Which form can be made into a
triangle?
Choices:
A) ice
B) steam
C) water
D) salt water
```

Output (that is the Answer produced by teacher)

```
Answer:
A) ice: Ice is a solid form of water. In solid
form, substances can be molded or shaped into
various forms, including triangles. If you
had a block of ice, you could carve or mold
it into a triangular shape.
B) steam: Steam is the gaseous form of
water. Being a gas, it doesn't have a fixed
shape. You can't shape steam into a triangle
without containing it within some form of
solid boundary, and even then, the steam
itself isn't triangular - the container is.
C) water: Water is a liquid that takes the
shape of its container. You can pour water
into a triangular container, and while the
volume inside would be triangular, the water
itself isn't inherently triangular.
D) salt water: This is essentially the same
as option C, but with salt dissolved in it.
Being a liquid, it takes the shape of its
container.
Given these considerations, the only
substance among the options that can be shaped
into a triangle by itself is A) Ice.
```

Table 1: Example of Demonstration. The structure is composed by: **Instruction**, **Input** and **Output**. In this case, the last part is the **Answer** produced by GPT-3.5 CoT prompted.

2.3 Curriculum Learning

Aligning the teacher-student reasoning abilities via Demonstrations delivered by multi-step answers provided by teachers CoT prompted is a promising technique. However, there are some aspects that need to be clarified: what constitutes an answer containing a good reasoning chain and how to evaluate it to optimize the Instruction-tuning phase. Following the Curriculum Learning (CL) where

[2] We consider Smaller models with less than 60B of parameters based on (Wei et al., 2022).

training algorithms can achieve better results when training data are presented according to the model's current skills (Bengio et al., 2009). We propose a method for evaluating the reasoning chain present in the CoT Answers (that represent the outputs of CoT Demonstration) using two fundamental properties: (1) comprehensibility, that is, the comprehensibility of a text according to metrics proposed by Talburt (1986), and (2) informativeness, that is, every step of the chain provides new information that is useful and informative for deriving the generated answer. We apply this metric to the CoT Answers provided by the teachers; then, we reorder the demonstrations according to our measure.

Informativeness To quantify the effectiveness of each step contributing novel information beneficial for deriving the final Answer A, we propose an assessment based on the Entropy and Information Gain (IG). The Entropy, represented by $H(S)$, evaluates the unexpected within a given sequence S, where $S_i \in A$. The entropy is given by:

$$H(S) = -\sum_{w \in S} p(w) \log_2 p(w) \qquad (1)$$

where $p(w)$ denotes the probability of token w occurring in the sequence. Hence, we compute the IG between a previous S_{prev} and a current sequence S_i as:

$$IG(S_{prev}, S_i) = H(S_{prev} + S_i) - H(S_{prev}) \quad (2)$$

This metric quantifies how much new information the current step adds relative to the cumulative content previously considered. To obtain a comprehensive measure, we calculate the average IG across the different steps as follows:

$$d_I(A_i) = \frac{1}{N} \sum_{i=1}^{N} IG(S_{prev}, S_i) \qquad (3)$$

where N represents the total number of steps in the Answer or the sequences S_i. We calculate this value for each answer A_i and obtain the maximum $d_{I_{max}}$ and the minimum $d_{I_{min}}$ scores. Finally, we normalize these values:

$$\hat{d}_I(A_i) = \frac{d_I(A_i) - d_{I_{min}}}{d_{I_{max}} - d_{I_{min}}}, \forall i \in [0, |D|]. \quad (4)$$

where $|D|$ are all answers to a specific benchmark.

Comprehensibility Typical factors for measuring comprehensibility are Speed of perception, Perceivability in peripheral vision, Reflex blink technique, Eye movements, Cognitively motivated features, and Word difficulty. However, it is not always possible to capture all these features.

Hence, we used the Flesch-Kincaid metric (Talburt, 1986). This metric is used to assess the comprehensibility of a text. It is based on the length of sentences and words within a text and provides a score that indicates the text's difficulty level. The lower the score, the easier it is to read and comprehend the text. The formula for calculating the Flesch-Kincaid Grade Level score is as follows:

$$\begin{aligned} d_C(A_i) = 0.39\frac{Avg(d_L(A_i))}{100} + \\ 11.8\frac{Avg(d_L(w_i))}{100} - 15.59 \end{aligned} \qquad (5)$$

where $Avg(d_L(A_i))$ average answer length is the number of words in a sentence divided by the number of sentences, and $Avg(d_L(w_i)$ is the average word length, i.e. is the number of syllables per word divided by the number of words. The value 0.39 is used to scale the effect of the average sentence length to compare it to the effect of the average word length, weighted by 11.8. The final score is then adjusted by subtracting the value of 15.59, which adjusts the score scale to match the grading levels used in education more closely. We calculate this value for each Answer A_i and obtain the maximum $d_{C_{max}}$ and the minimum $d_{C_{min}}$ scores. Finally, we normalize these values:

$$\hat{d}_C(A_i) = \frac{d_C(A_i) - d_{C_{min}}}{d_{C_{max}} - d_{C_{min}}}, \forall i \in [0, |D|]. \quad (6)$$

Constructing the CL-Demonstration We gather the annotations (answers) delivered by the CoT-prompted teachers (as explained in Section 2.1), and we estimate the informativeness $\hat{d}_I(A_i)$ and complexity $\hat{d}_C(A_i)$ for each answer $A_i, \forall i \in |D|$.

Then we merge the two values in:

$$d_{IC}(A_i) = \hat{d}_I(A_i) + \hat{d}_C(A_i) \qquad (7)$$

We use $d_{IC}(A_i)$ as a pivot value to reorder the Answers provided by the teachers. The Answers (which form the output of the Demonstrations) will be delivered in the Instruction-tuning phase to the students in ascending order with respect to the value $d_{IC}(A_i)$. These heuristics are very lightweight: using only 16GB of memory, we can process up to 20k Responses per second to produce the informativeness and comprehensibility metrics.

3 Experimental Setup

In order to make the experiments comparable with state-of-the-art models, we use four benchmarks (introduced in Section 3.1) that are generally used to assess the abilities of Large Language Models (LLMs). Moreover, to conduct the Instruction-tuning phase on the Small Language Models (SMLs), we use two approaches: the first one is presented in Section 3.2, which we call Instruction-tuning on Demonstrations; the second is based on the Curriculum Learning (CL) approach where the students are exposed to CL-Demonstrations that are Demonstrations reordered in a CL way, as exemplified in Section 2.3. All code is available in the supplementary material, to be released if accepted.

3.1 Data

General Commonsense Reasoning We evaluate the models' ability to perform general reasoning on the CommonSenseQA (Talmor et al., 2019) (CSQA) and OpenBookQA (Mihaylov et al., 2018) (OBQA). CommonSenseQA is one of the best-known datasets of answers to multiple-choice questions dealing with different types of general commonsense knowledge. OpenBookQA is a resource that contains questions requiring multi-step reasoning, common knowledge, and rich text comprehension. It is inspired by high school-level open-book exams in physics and biology, aiming to assess human comprehension and application of foundational concepts.

Physical Interaction Reasoning We evaluate the models' ability to perform physical reasoning on the Interaction Question Answering (PIQA) (Bisk et al., 2019). It is a resource consisting of everyday situations with typical and atypical solutions.

Social Interaction Reasoning We evaluate the models' ability to perform social reasoning on the Social Interaction Question Answering (SIQA) (Sap et al., 2019). It is a benchmark focusing on reasoning about people's actions and social implications. The actions in Social IQa cover various social situations and candidates for plausible and not plausible answers.

Splitting Details Since a test split for all benchmarks is not always available open-source, we adopt the following strategy: we use 4000 examples with equally distributed target classes as training data and the validation versions found on huggingface as test data. We performed this split because we needed to observe the impact of the responses provided by the teacher models on different benchmarks. The same is true for validation since we needed open-source and reproducible data to conduct a detailed evaluation of the student models. In Table 9, we report the quantitative information, global, and splitting ratios, and in Table 8, we show one example for each benchmark. The data are fully accessible and open-source, as described in Table 10.

3.2 Teaching to Reason

We selected Llama-2-70 and GPT-3.5 as the teachers (introduced in Section 2.1). Consequently, the LLMs are prompted in the zero-shot scenarios, as shown in Table 5 and Table 6.

We selected Llama-2-7 and Llama-2-13 (Touvron et al., 2023) as student models (as described in Section 2.2). Therefore, the students models are Instruction-tuned via Demonstrations, as introduced in Section 3.2, and via CL-Demonstrations, as explained in Section 2.3. Table 1 shows a Demonstration containing the Instruction, Input, and, as Output, the Answer-delivering CoT, an output generated by GPT-3.5 CoT-prompted.

3.2.1 Models Setup

We conduct the Instruction-tuning phases using QLoRA proposed by Dettmers et al. (2023). This approach allows tuning to be conducted while reducing memory usage and preserving the performances. We follow the training approach proposed in Alpaca (Taori et al., 2023). Our models are trained for three epochs and set the learning rate as 0.00002 with 0.001 weight decay. We use the cosine learning rate scheduler with a warmup ratio of 0.03. We conducted our experiments on a workstation equipped with two Nvidia RTX A6000 with 48GB of VRAM.

3.3 Evaluation

The most commonly used evaluation methods for question-answering tasks are language-model probing, in which the option with the highest probability is chosen (Brown et al., 2020), and multiple-choice probing, in which the models are asked to answer.

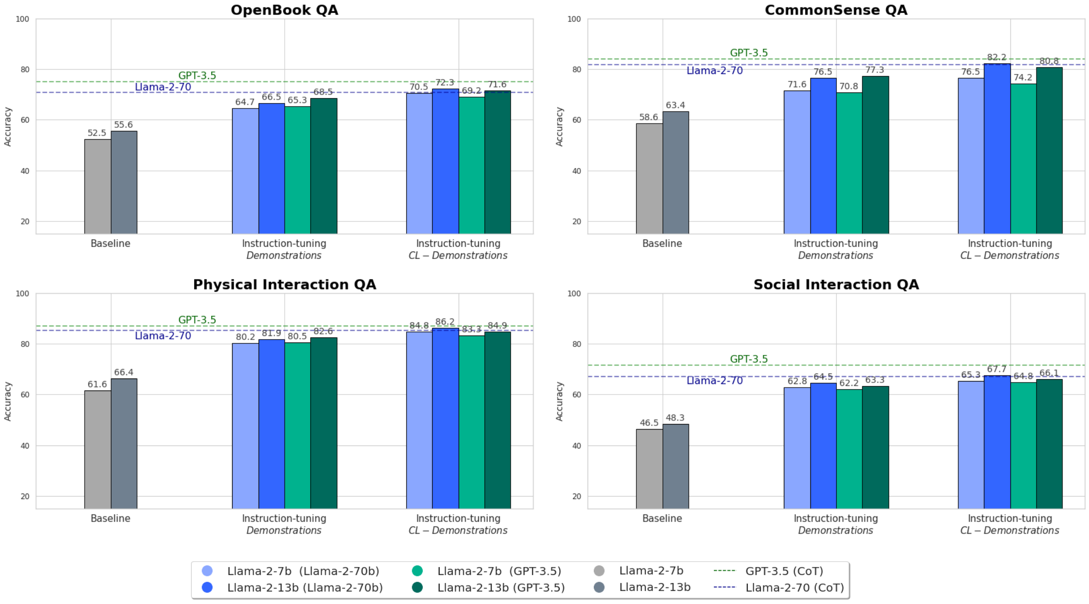

Figure 2: Accuracies (%) on benchmarks (Section 3.1) before Instruction-tuning (i.e., Baselines) and after on Demonstrations and CL-Demonstrations. Moreover, there are the teachers' performances also shown in Table 4

The evaluation is performed with a function taking the maximum value and, in the second case, with string matching. The second method is widely used in recent evaluations because it applies to models such as GPT-3.5 and GPT-4 (OpenAI, 2023) where probability values cannot be accessed. In our experiments, we chose the latter to have a comparable and scalable pipeline. Hence, we performed a string matching between the generated outputs and the targets.

4 Results & Discussion

Language Models that do not get it can be elicited to do it through the knowledge of teacher models. These conclusions can be observed in Figure 2, where we reported the downstream performances without the Instruction-tuning phase (see the Baseline) and the Instruction-tuning on Demonstrations. As discussed in Section 4.1, Small Language Models (SLMs) CoT prompted obtained weak results. In contrast, models that are instructed via Chain-of-Thought (CoT) Demonstrations, i.e., Demonstrations produced by CoT-prompted Large Language Models (LLMs), outperform non-instructed models (Section 4.2). However, although Demonstrations produced better students, the complete alignments between students and teachers are realized with the Curriculum Learning approach, as discussed in Section 4.3. In particular, the students instructed via the CL approach (Instruction-tuning CL-Demonstration in Figure 2) outperformed the students instructed via standard Instruction-tuning.

Finally, the CL approach delivers the teacher-student family-alignment. In Figure 2 (horizontal lines), it is possible to observe the phenomenon of family-alignment between Llama-2-70 and Llama-2-7 and -13 in more detail in Section 4.4.

4.1 CoT-abilities of Small Language Models

Chain-of-Thought (CoT) prompts do not consistently deliver downstream performance improvements. SLMs, i.e., with fewer parameters, have not benefitted the prompting with the CoT mechanism. In particular, we evaluated performance on four question-answering benchmarks, described in Section 3.1, using two versions of Llama-2-chat (7b-13b billion). Proposing a classical prompt (which we call "Baseline") and a CoT prompt (Table 5 and Table 6), we obtained the performances in Table 2.

The results confirm what Wei et al. (2022) have claimed about the limitations of the emergent CoT prompting abilities that are not observable in SLMs. Moreover, using CoT prompting leads to model confusion with subsequent degradation of downstream results. It is possible to observe these phenomena in OpenBookQA (OBQA) and CommonSenseQA (CSQA) (down arrows in Table 2). In particular, there is a marked deterioration in Llama-2-7 (see ⇓), which has half the parameters of Llama-2-13 (see ↓).

The same behavior was not observed for Physical- and Social-Interaction Question Answering (PIQA) and (SIQA). In fact, not considering the nature of benchmarks, unlike the others, they are always question-answering multiple-choice-questions but have fewer possible choices, as shown in Table 9. In this regard, we hypothesize that the most controllable scenarios, where chain reasoning is limited to fewer options, are reasonable by SLMs elicited with CoT prompts.

Benchmarks	**Llama-2-7**		**Llama-2-13**	
	Baseline	**CoT**	**Baseline**	**CoT**
OBQA	**52.5**	49.5⇓	**57.6**	55.6↓
CSQA	**58.6**	50.6⇓	**63.4**	60.8↓
SIQA	46.5	*45.3*	48.3	47.6
PIQA	61.6	*63.8*	66.4	*68.2*

Table 2: Accuracies of Llama-2-7 and Llama-2-13, both without further tuning, on testing data with the standard prompt (Baseline) (see Table 5) and CoT prompt (CoT) (see Table 6).

4.2 The Instruction-tuning Method

Instruction-tuning led by Large Language Models (teachers models), able to reason elicit the Smaller Language Models (students models) to do the same. This is shown in Figure 2. The student models behind Instruction-tuning on Demonstrations produced by teacher models outperformed the baselines in the four proposed benchmarks. While performances are conspicuous improvements overall, they have sensible variations. The teacher models have different characteristics. GPT-3.5 is trained on 175B and Llama-2-70 on 70B of parameters. They consequently achieve different performances in the proposed benchmarks. Table 4 shows the performances in the zero-shot scenario (CoT prompting and not) on the data used to conduct the Instruction-tuning phase and on the same test set used to evaluate the proposed models.

Although the performances on the "training set" are different (see the CoT performances of GPT-3.5 and the same for Llama-2-70 in Table 4), this bias does not affect the students. The Llama-2-7 and -13 with GPT-3.5 as teacher outperform the Llama-2-7 and -13 with Llama-2-70 as teacher only on OBQA. As far as CSQA and PIQA are concerned, there is a balance that is not present in SIQA, where the students of Llama-2-70 outperform the others.

However, in the Instruction-tuning method, instruction is conducted using Demonstrations (composed of Answers provided by teachers) that are delivered arbitrarily. Therefore, we propose to study both the intrinsic complexity of the answers and their impact on the students' exposure. In particular, we propose a CL-based instruction approach where demonstrations are delivered to students in a meaningful order (Section 4.3).

4.3 The Impact of Curriculum Learning

Instruction-tuning via Curriculum Learning Demonstrations elicits the reasoning abilities of students. The students gradually exposed to increasingly meaningful Demonstrations (CL-Demonstrations) learn better than those exposed to arbitrary Demonstrations. This is shown in Figure 2 (bars Instruction-tuned CL-Demonstrations), where Llama-2-7 and -13 consistently outperformed the other models.

The benchmarks where the most significant effects can be observed are CSQA and OBQA, with an increase in average accuracy scores of 6 and 5 points, respectively. The same effects are less evident in PIQA and SIQA. One possible reason for this phenomenon might be tied again to the nature of the benchmarks, as hypothesized in Section 4.1. To analyze this phenomenon, we studied the components of the complexity measure proposed in Section 4.5.

4.4 The role of CL in family-alignment

Instruction-tuning via CL-Demonstrations still aligns students' reasoning abilities with family teachers, even as instruction decreases. In fact, from Figure 2, we can observe that the performances of students instructed via CL-Demonstrations delivered by teachers from the same family outperform the others.

Moreover, to validate our hypothesis of family-alignment, we introduced Mistral-7b (MistralAI, 2023), a new SLMs with 7 billion parameters that outperforms the Llama-2-13 version on several benchmarks, as shown by MistralAI (2023). In particular, we reproduced the experiments introduced in Section 4.2. In Figure 3, it can be seen that Llama-2-7 instructed on different types of Demonstrations delivered by Llama-2-70 almost consistently outperforms Mistral-7b. These results confirm that Demonstrations derived from in-family teachers have a more significant impact on student models than the others.

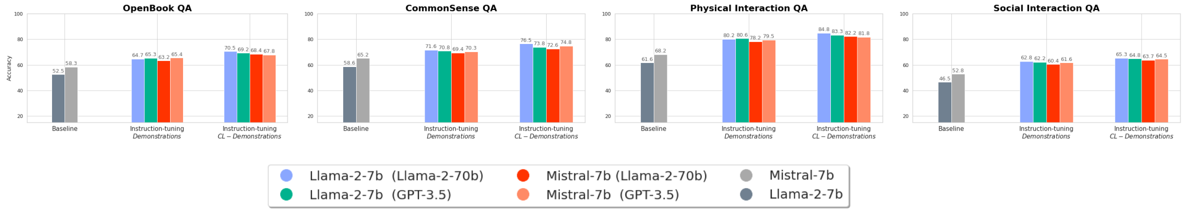

Figure 3: Accuracies of Llama-2-7 and Mistral-7 Instruction-tuned using setup proposed Section 3.

4.5 Ablation Study

The informativeness and complexity exposed to students in a meaningful order instruct better students. We conducted an Ablation study to estimate the impact of our evaluation measures proposed in Section 2.3. Hence, we reproduced the same configurations proposed in Section 4.2, but removed one of the two components (informativeness and complexity) presented in Section 2.3. The results in Table 3 show that students instructed on CL-Demonstrations ordered by comprehensibility and informativeness consistently outperform students instructed via Demonstrations. The results show that students instructed on the Demonstrations sorted by informativeness are more productive in QA tasks with more choices. In comparison, complexity proved helpful in cases where the number of choices is minor. Phenomenon manifested in CSQA and OBQA with 5 and 4 choices and PIQA and SIQA with 2 and 3 choices, respectively (see Appendix, Tables 8 and 9).

5 Related Work

5.1 Learning from Explanation

Current methods for conditioning models on task instructions and provided explanations for individual data points replace the ancient intermediate structures (Hase and Bansal, 2022) that used rationales (Zhang et al., 2016) or inputs (Narang et al., 2020; Talmor et al., 2020) to learn the models. Reasoning via the CoT builds upon prior efforts wherein explanations are viewed as intermediary constructs produced during inference (Rajani et al., 2019). Our research stems from the studies of Shridhar et al. (2023); Ho et al. (2023b). In particular, we adopt the idea of an LLM teacher and a second LLM, sometimes smaller, that assumes a student's position (Magister et al., 2023). Learning uses teacher-generated explanations, demonstrating prompt CoT downstream (Li et al., 2023; Ho et al., 2023b). Li et al. (2023) claims that massive demonstrations significantly improve performance over the single-sample approach Shridhar et al. (2023).

5.2 Large Language Models as a Teacher

Several papers have been published simultaneously, including those by Ranaldi and Freitas (2024); Paul et al. (2024), and Saha et al. (2023) that prove the effect of transferring ability to produce CoT reasoning from larger to smaller models. Table 11 resumes all main points of these contributions.

Our work goes beyond the following ways: i) We propose a method for aligning CoT abilities via Instruction-tuning through Demonstrations produced by answers generated by GPT-3.5 and Llama-2-70. ii) We study how to provide Demonstrations to students by proposing a measure for evaluating the Answers provided by teachers. In particular, we analyze the alignment performance between in-family and out-family models. iii) Hence, we propose an approach for improving the alignment of reasoning abilities between teachers and students by employing our evaluations to expose the students meaningfully.

6 Conclusion

In this paper, we propose a method to enable CoT reasoning over SLMs by introducing two novel mechanisms. First, we propose aligning CoT abilities via Instruction-tuning with the support of CoT Demonstrations delivered by LLMs teacher. Second, we use the Curriculum Learning approach to empower the Instruction-tuning phase. Hence, we analyze the impact on the downstream abilities of four question-answering benchmarks. Our results show that SMLs can be instructed to reason via Demonstration produced by LLMs. We move a step further in research: conceiving SLMs as human learners, we expose them to a CL teaching-based approach, obtaining better results on downstream performances.

Limitations

In our contribution, we analyzed the impact of Answers delivered by Large Language Models, using them as Demonstrations to empower the abilities of Small Language Models. Although we proposed an extensive study, there are several limitations. Firstly, only English-language methods, both in Chain-of-Thought (CoT) methods and task evaluation, are considered. In future works, we will investigate this aspect, starting from Cross-lingual alignment approaches.

Secondly, dependence on LLMs, which are closed-source products or not, but sometimes the training sets are unknown. Although the characteristics of the corpora are reported in the system reports, these are only processable by some researchers. Analyzing the differences in pre-training data between models is difficult.

Finally, learning from and with Demonstrations carries some specific risks associated with automation. Although a model may generalize its predictions using a seemingly consistent series of natural language steps, even if the prediction is ultimately correct, there is no guarantee that the predicted output comes from a process represented by the generalization. A user might be overconfident in the model based on the CoT. Hence, in the future, we will investigate refinement approaches based on RLHF and DPO to improve the generalization abilities of Student models.

Ethic Statement

Although this research enhances the reasoning abilities of smaller Language Models, they still need to be sufficiently robust for sensitive contexts such as education. The primary ethical concerns arise from the text generation process; both the "teacher" and "student" models might produce misleading answers. The content is largely influenced by the input data, which, in our case, are standard benchmarking tasks peer-reviewed within the NLP domain. We intend to release our code; however, like many generative models, ours can be exposed to hallucinations.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In Proceedings of the 26th annual international conference on machine learning, pages 41–48.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Yao Fu, Litu Ou, Mingyu Chen, Yuhao Wan, Hao Peng, and Tushar Khot. 2023. Chain-of-thought hub: A continuous effort to measure large language models' reasoning performance.

Vedant Gaur and Nikunj Saunshi. 2023. Reasoning in large language models through symbolic math word problems. In Findings of the Association for Computational Linguistics: ACL 2023, pages 5889–5903, Toronto, Canada. Association for Computational Linguistics.

Peter Hase and Mohit Bansal. 2022. When can models learn from explanations? a formal framework for understanding the roles of explanation data. In Proceedings of the First Workshop on Learning with Natural Language Supervision, pages 29–39, Dublin, Ireland. Association for Computational Linguistics.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023a. Large language models are reasoning teachers.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023b. Large language models are reasoning teachers. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14852–14882, Toronto, Canada. Association for Computational Linguistics.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023. Symbolic chain-of-thought distillation: Small models can also “think” step-by-step. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2665–2679, Toronto, Canada. Association for Computational Linguistics.

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. Evaluating the logical reasoning ability of chatgpt and gpt-4.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching small language models to reason. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 1773–1781, Toronto, Canada. Association for Computational Linguistics.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering.

MistralAI. 2023. Mistral-7b-instruct.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. 2024. Refiner: Reasoning feedback on intermediate representations.

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.

Leonardo Ranaldi and Andre Freitas. 2024. Aligning large and small language models via chain-of-thought reasoning. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1812–1827, St. Julian’s, Malta. Association for Computational Linguistics.

Swarnadeep Saha, Peter Hase, and Mohit Bansal. 2023. Can language models teach weaker agents? teacher explanations improve students via personalization.

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.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2023. Distilling reasoning capabilities into smaller language models. In Findings of the Association for Computational Linguistics: ACL 2023, pages 7059–7073, Toronto, Canada. Association for Computational Linguistics.

John Talburt. 1986. The flesch index: An easily programmable readability analysis algorithm. New York, NY, USA. Association for Computing Machinery.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Alon Talmor, Oyvind Tafjord, Peter Clark, Yoav Goldberg, and Jonathan Berant. 2020. Leap-of-thought: Teaching pre-trained models to systematically reason over implicit knowledge.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

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. 2022. Emergent abilities of large language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Ye Zhang, Iain Marshall, and Byron C. Wallace. 2016. Rationale-augmented convolutional neural networks for text classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 795–804, Austin, Texas. Association for Computational Linguistics.

Appendix A

	Benchmarks			
Students	**OBQA**	**CSQA**	**PIQA**	**SIQA**
Llama-2-7 (Llama-2-70)				
Arbitrary Teaching	64.7	71.6	80.2	62.8
Teaching via IC	**70.5**	76.5	**84.8**	65.3
Teaching via I	70.2⇑	**77.2**⇑	81.2	61.8⇓
Teaching via C	66.4	69.7⇓	84.3⇑	**66.2**
Llama-2-7 (GPT-3.5)				
Arbitrary Teaching	65.3	70.8	80.5	62.2
Teaching via IC	**69.2**	**74.2**	83.3	**64.8**
Teaching via I	68.5⇓	73.7⇑	79.6⇓	63.8
Teaching via C	66.3	69.8	**83.9**⇑	65.7⇑
Llama-2-13 (Llama-2-70)				
Arbitrary Teaching	66.5	76.5	81.9	64.5
Teaching via IC	72.3	**82.2**	**86.2**	67.7
Teaching via I	**73.4**⇑	81.9⇑	80.7⇓	63.8
Teaching via C	67.9	76.6	84.3⇑	**70.3**
Llama-2-7 (GPT-3.5)				
Arbitrary Teaching	68.5	77.3	82.6	63.3
Teaching via IC	**71.6**	80.5	**84.9**	**66.1**
Teaching via I	70.8⇑	**81.7**	81.9	62.7
Teaching via C	68.2⇓	78.5	82.3	65.9⇑

Table 3: Ablation study on our Instruction-tuning CL-Demonstrations approach.

Benchmarks	**Llama-2-70**		**GPT-3.5**	
	Baseline	**CoT**	**Baseline**	**CoT**
Training				
OBQA	65.6	71.3	66.2	**75.4**
CSQA	74.2	79.6	79.3	**84.8**
SIQA	65.4	67.5	67.6	**70.3**
PIQA	82.6	**85.8**	83.5	85.3
Testing				
OBQA	65.9	70.8	67.8	**74.6**
CSQA	73.4	81.8	80.2	**83.7**
SIQA	64.2	66.9	66.9	**71.3**
PIQA	82.6	85.6	84.3	**85.8**

Table 4: Accuracy (%) of Llama-2-70 and GPT-3.5 (teachers) on training and testing data with CoT prompt (CoT) and with the standard prompt (Baseline).

Appendix B

Zero-Shot
Choose the answer to the question only from options A, B, C, D. Question: Which animal gives birth to live young? A) Shark B) Turtle C) Giraffe D) Spider Answer:

Table 5: Example of Zero-Shot prompting.

Zero-Shot Chain-of-Thought
Choose the answer to the question only from options A, B, C, D. Question: Which animal gives birth to live young? A) Shark B) Turtle C) Giraffe D) Spider Answer: **Let's think step by step**

Table 6: Example of Zero-Shot Chain-of-Thought prompting.

Appendix C

Model	**Version**
Llama-2-7-chat	meta-llama/Llama-2-7b
Llama-2-13-chat	meta-llama/Llama-2-13b
Llama-2-70-chat	meta-llama/Llama-2-70b
Mistral-7-instruct	mistralai/Mistral-7B-Instruct-v0.1
GPT-3.5-turbo	OpenAI API

Table 7: In this table, we list the versions of the models proposed in this work, which can be found on huggingface.co. We used all the default configurations proposed in the repositories for each model.

Appendix D

Dataset	Example
OBQA (Mihaylov et al., 2018)	*When birds migrate south for the winter, they do it because* **A) they are genetically called to.** B) their children ask them to. C) it is important to their happiness. D) they decide to each.
CSQA (Talmor et al., 2019)	*Aside from water and nourishment what does your dog need?* A) bone. B) charm. C) petted. **D) lots of attention.** E) walked.
PIQA (Bisk et al., 2019)	*How do you attach toilet paper to a glass jar?* **A) Press a piece of double-sided tape to the glass jar and then press the toilet paper onto the tape.** B) Spread mayonnaise all over the jar with your palms and then roll the jar in toilet paper.
SIQA (Sap et al., 2019)	*Taylor gave help to a friend who was having trouble keeping up with their bills. What will their friend want to do next?* A) Help the friend find a higher paying job. **B) Thank Taylor for the generosity.** C) pay some of their late employees.

Table 8: Examples of the benchmarks used in this paper.

	OBQA	CSQA	PIQA	SIQA
classes	4	5	2	3
Training # examples for each class	1000	800	2000	1330
Test # examples for each class	125* (± 8)	235* (± 11)	924* (± 18)	640* (± 19)

Table 9: Characteristics Training and Test set of benchmarks proposed in Section 3.1. The * indicates that the number of examples are not perfect balanced, but the difference from the average is marginal.

Name	Repository
CSQA (Talmor et al., 2019)	`huggingface.co/datasets/commonsense_qa`
OBQA (Mihaylov et al., 2018)	`huggingface.co/datasets/openbookqa`
PIQA (Bisk et al., 2019)	`huggingface.co/datasets/piqa`
SIQA (Sap et al., 2019)	`huggingface.co/datasets/social_i_qa`

Table 10: In this table, we list the versions of the benchmark proposed in this work, which can be found on huggingface.co.

Work	Method	Teachers	Students
(Magister et al., 2023)	SFT	PaLM GPT-3.5	T5-small, -medium T5-large, -xxl
(Li et al., 2023)	SFT	GPT-3 175B	OPT-1.3b
(Shridhar et al., 2023)	SFT	GPT-3 175B	GPT-2
(Ho et al., 2023b)	SFT	InstructGPT (text-davinci-002)	GPT-3 (ada,babbage,curie)
Ours	**Instruction-tuning**	**Llama-2-70b** GPT-3.5 (turbo)	**Llama-2-7b, -13b** **Mistral-7b**

Table 11: Summary of methods, teacher and student models of previous work, we indicate Supervised Fine-tuning as (SFT) employed in most previous work.