Aligning Large Language Models via Chain-of-Thought Reasoning

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 investigate the alignment of reasoning abilities from larger to smaller Language Models. In particular, using Instruction-tuning-CoT approach, that is, an Instruction-tuning empowered towards CoT-Demonstrations, we analyze the impact on the the downstream abilities. Hence, we instruct a smaller Language Model using outputs generated by more robust models belonging to the same family or not, and we analyze the impact and divergencies. Results obtained on four question-answering benchmarks show that SMLs can be instructed to reason via CoT-Demonstration produced by LLMs.

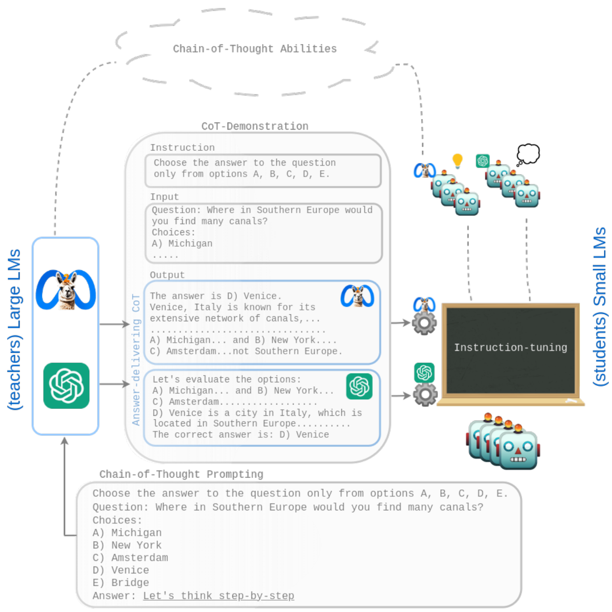

Figure 1: In Instruction-tuning-CoT, the Students models instruct themselves using CoT-Demonstrations, which are Demonstrations-delivering CoT reasoning generated by the Teachers models. We elicit a Large Language Model to answer questions through Chain-of-Thought reasoning mechanism. Then, we use the CoT-Demonstrations to instruct a Small Language Model to reason as a Large Language Model would.

1 Introduction

Chain-of-Thought (CoT) prompting elicits Large Language Models (LLMs) to break down a reasoning task towards a sequence of intermediate steps (Wei et al., 2022). Previous works have demonstrated that in LLMs with at least several billions of parameters, such as GPTs family (OpenAI, 2023) or PaLM (Chowdhery et al., 2022), CoTs enables the delivery of multi-step, controlled reasoning, achieving results across commonsense (Bubeck et al., 2023), symbolic and mathematical reasoning datasets (Gaur and Saunshi, 2023; Liu et al., 2023).

The size of LLMs, however, poses an adoption barrier for numerous users. In order to facilitate accessibility, derived scaled-down models from the same family but with reduced size have been introduced, such as Llama-2-7b and -13b as the corresponding 'Smaller Language Models (SLMs)' associated with Llama-2-70b (Touvron et al., 2023), both of them having less than half billion of parameters. Although these SLMs 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)). In fact, these SLMs produce illogical answers when prompted under the CoT framework.

In this paper, we propose a method to enable

CoT reasoning over SLMs by introducing two novel mechanisms. The first is the concept of in-family alignment for teacher-student Instruction-tuning (i.e. prioritising models from the same family instead from different families). In particular, we investigate the alignment of Chain-of-Thought reasoning abilities through the support of CoT-Demonstrations "taught" by LLMs teachers to SLMs students (see Figure 1), contrasting within in-family and and out-family settings. As concerning (Magister et al., 2023; Ho et al., 2023a; Shridhar et al., 2023) we introduce the Instruction-tuning approach through which, respect to (Li et al., 2023a), we instruct student models with CoT-Demonstrations produced by in-family and out-family teachers.

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?

RQ2) What is the effect of Demonstrations delivered with the Chain-of-Thought reasoning process?

RQ3) How much do Demonstrations produced by an in-family teacher impact a student alignment?

To answer these questions, we select Llama-2-7b and Llama-2-13b (Touvron et al., 2023) as students and Llama-2-70b and GPT-3.5 as in-family and out-family teachers. Then, 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 CoT-Demonstrations (see Figure 1) used to instruct Llama-2-7 and -13. We discern the CoT-Demonstrations between Demonstrations-delivering CoT and Demonstrations-misleading CoT stems from Answers-delivering CoT (correct CoT prediction) and Misleading CoT (wrong CoT predictions). Furthermore, to have a term of comparison, we produce the base Demonstrations formed the same way as the previous ones without CoT prompting. Figure 12 shows the terminology used in this work, and Figure 13 summarizes the overall pipeline.

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, students instructed with CoT-Demonstrations outperformed those instructed without CoT. Finally, students instructed with Demonstrations-delivering CoT provided by the in-family teachers outperformed those instructed with out-family ones and achieved the best performances.

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 Instruction-tuning via CoT-Demonstrations aligns the reasoning abilities of SLMs with the ones of LLMs. In fact, models instructed through CoT-Demonstrations that contain outputs generated via CoT prompting outperform models trained with Demonstrations.

iii) Finally, in-family alignment with Instruction-tuning via Demonstrations-delivering CoT formed by Answers-delivering CoT outperforms out-family alignment. Llama-2-7 and Llama-2-13 instructed with Answer-delivering-CoT Demonstrations produced by Llama-2-70 outperform the students instructed by teacher GPT-3.5 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 two steps, as shown in Figure 1[1]. In the first part, there is an annotation phase where the Large Language Models (LLMs) systematically prompt generate outputs (Section 2.1). These outputs will be the core of Demonstrations used during the Instruction-tuning phase from the smaller Language Models, presented in Section 2.2.

2.1 Teacher Model

Many state-of-the-art LLMs are available that differ in the number of parameters and training modes. However, our research questions focuse on Instruction-tuning and family-alignment of reasoning abilities. Therefore, we concentrated on robust models with different versions of the same family 4.

As a robust LLMs, we selected Llama-2-70b (Touvron et al., 2023), and in terms of comparison,

[1]Figure 13 shows the overall pipeline.

GPT-3.5[2] (OpenAI, 2023). Meanwhile, Llama-2-70b because, as introduced before, there are several smaller versions (presented in Section 2.2) despite the reduced number of parameters, they obtain remarkable results. In particular, we use the "chat" version of the LLM called Llama-2-70-chat. We selected this version because, as reported by Touvron et al. (2023), it 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 chose an out-family model to observe the impact of the Llama-2-70bs abilities. We select GPT-3.5 because it generates high-quality data either with and without the CoT prompting approach, as shown by Fu et al. (2023).

Although our focus is on CoT abilities, in order to conduct a comprehensive study, we proposed two different input-prompts, both in a zero-shot scenario. The first input-prompt is a classic standard prompt, consisting of the question and its choices as follows:

```
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:
```

The second input-prompt is like the first, but behind "`Answer:`" we insert "**`Let's think step by step`**" as proposed by Wei et al. (2022), that is shown in Table 6.

Following the annotation process performed by LLMs, the annotations have been used to construct the Demonstrations.

2.2 Student Model

Several SLMs have been fine-tuned either for instruction-following (Taori et al., 2023) or reinforcement learning with human feedback (Ouyang et al., 2022). However, whatever the techniques, the smaller Language Models[3] do not seem able to reproduce the step-by-step reasoning abilities.

However, recent work proposes techniques of knowledge distillation (Li et al., 2023a), skill refinement (Huang et al., 2022), and enriched fine-tuning (Magister et al., 2023) to teach SLMs to produce

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

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

`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 model)

```
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 CoT-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.

reasoned answers. Following this idea, we analyze the impact of Instruction-tuning from Demonstrations provided by in-family and out-family 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, conducted as in Alpaca (Taori et al., 2023). This approach concerns the tuning of Demonstrations, which consist of an instruction which, 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.

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 the approach presented in Section 3.2. All code is available in the supplementary material, to be released if accepted.

3.1 Data

With the successful growth of the LLMs, several question-answering benchmarks with multiple-choice questions have been proposed to build solid assessments of the models' abilities. In this paper, we selected four benchmarks that deal with topics around reasoning:

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 Commonsense 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 a series of everyday situations with a pair of typical and atypical solutions.

Social Commonsense 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 need open-source and reproducible data to conduct a detailed evaluation of the student models. In Table 10, we report the quantitative information, global, and splitting ratios, and in Table 9, we show one example for each benchmark. The data are fully accessible and open-source, as described in Table 11.

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 one-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, as proposed in (Taori et al., 2023). Hence, the SLMs are instructed on the Demonstrations that contain the answers generated by the teachers, as explained in Section 2.2. Table 1 shows a CoT-Demonstration that is Demonstration that contains the Instruction, the Input, and, as Output, the Answer-delivering CoT that is an output generated by GPT-3.5 CoT-prompted.

3.2.1 Models Setup

We conduct Instruction-tuning phase using QLoRA proposed by Dettmers et al. (2023). This approach allows instruction-tuning (and, more generally, fine-tuning) to be conducted while reducing memory usage. In particular, Dettmers et al. (2023) propose several techniques for tuning models with many parameters on GPUs with limited resources while preserving 16-bit tuning performance.

We follow the training approach proposed in Alpaca (Taori et al., 2023). Our models are trained for one epoch 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.

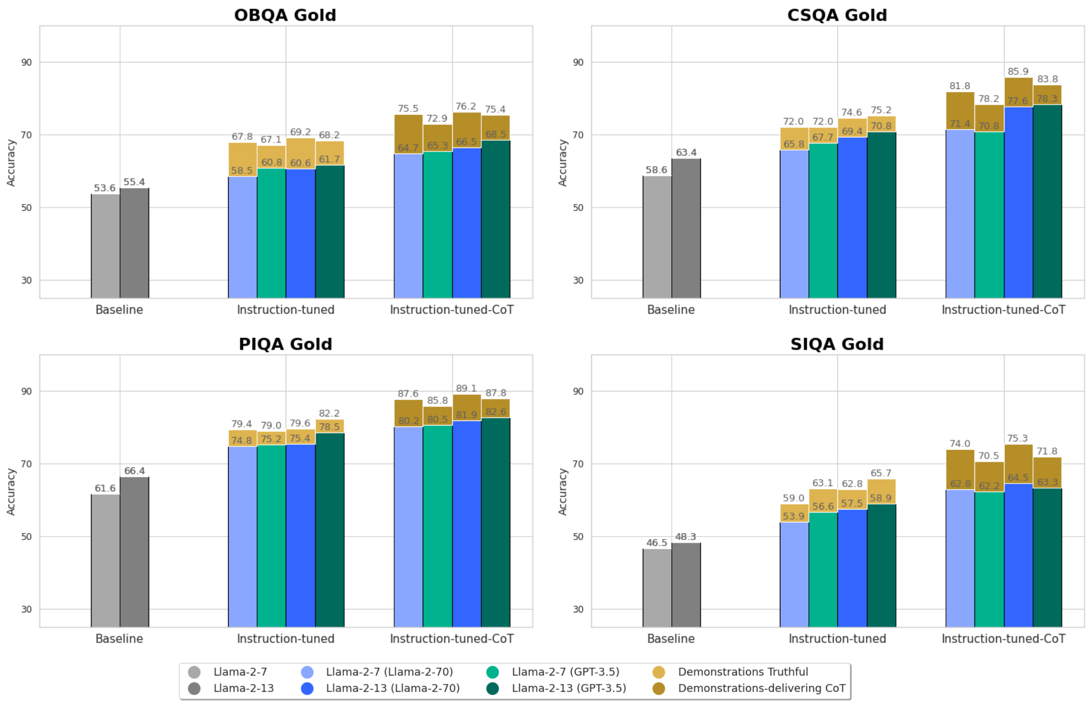

Figure 2: Accuracies (%) on benchmarks (Section 3.1) before Instruction-tuning (i.e., Baselines) and behind on Demonstrations (i.e., Instruction-tuned) and CoT-Demonstrations (i.e., Instruction-tuned-CoT). In addition, Instruction-tuning phases only on Demonstrations-delivering CoT and Demonstrations Truthful, specifically, demonstrations with Answers-delivering CoT and Answer Truthful (correct predictions), provided by teachers models without Misleading ones.

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. The evaluation in the first case is performed with a function taking the maximum value and in the second case with a string matching. The second method is widely used in recent evaluations because it applies to models such as GPT-x (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. Therefore, we performed a string matching between the generated outputs and the target choice.

4 Results & Discussion

Language Models that were unable to reason can be elicited to do it through the knowledge of teacher models. These conclusions can be observed in Figure 2, where there are the downstream accuracies without the Instruction-tuning phase (see the Baseline) and the Instruction-tuning phase on Demonstrations. In fact, 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 other models (see the Instruction-tuned-CoT in Figure 2).

However, although CoT-Demonstrations produced better students, the complete alignments between students and teachers are realized via Demonstrations-delivering CoT, as discussed in Section 4.2. In particular, the "Demonstrations-delivering CoT" and "Demonstrations Truthful" bars in Figure 2 show that student models instructed via Demonstrations-delivering CoT outperformed students instructed via CoT-Demonstrations, which contained Demonstrations Misleading CoT.

Finally, students instructed with Demonstrations-delivering CoT produced by in-family teachers always outperformed students instructed with Demonstrations-delivering CoT produced by out-family teachers. In Figure 2, 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 4.2.

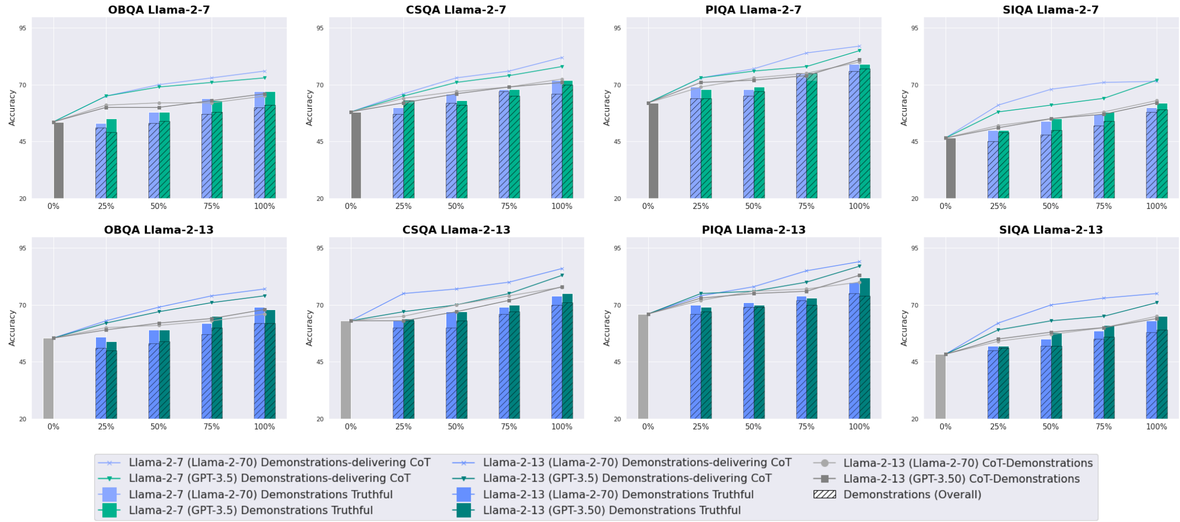

Figure 3: Accricacies (%) on the test set of benchmarks. The Instruction-tuning performed on different splits (see Appendix A for major details) of Demonstrations and CoT-Demonstrations (correct and not correct predictions), Demonstrations Truthful, and Demonstrations-delivering CoT (correct predictions).

4.1 CoT-abilities of Small Language Models

Chain-of-Thought (CoT) prompts are not always delivering downstream performance improvements. In fact, SLMs, i.e., with fewer parameters, have not outperformed when prompted 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) in a zero-shot scenario. 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 ↓).

However, the same behaviour 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 10. 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	**55.3**	49.5⇓	**57.6**	55.2↓
CSQA	**59.2**	50.6⇓	**64.3**	60.8↓
SIQA	47.5	*45.3*	49.3	47.6
PIQA	63.5	*63.8*	69.5	71.2

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 Impact

Instruction-tuning led by Large Language Models (teachers models), able to reason, conduct the Smaller Language Models (students models) to do the same. This can be seen in the experiments in Figure 2. The student models behind Instruction-tuning on Demonstrations produced by teacher models outperformed the baselines in the four proposed benchmarks. Moreover, the students models instructed with CoT-Demonstrations, defined as Instruction-tuned-CoT in Figure 2, achieve the best results in terms of accuracy.

While performances are conspicuous improvements overall, they have sensible variations. The teacher models have different characteristics, as shown in Figure 4. GPT-3.5 is trained on 175B of parameters and Llama-2-70 by analog name on 70B of parameters. They consequently achieve different performances in the proposed benchmarks.

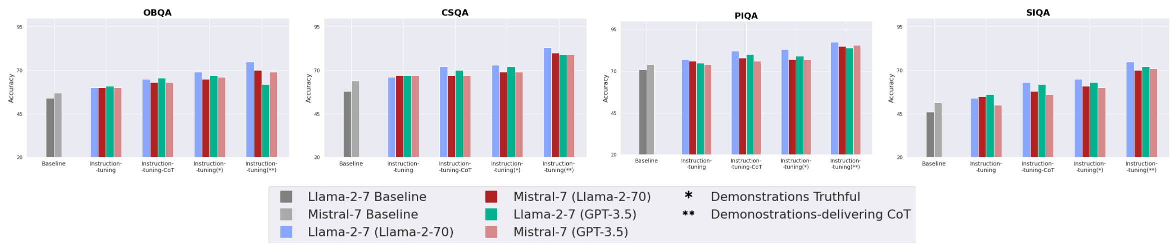

Figure 4: Performances of Llama-2-7 and Mistral-7 Instruction-tuned using the same setup proposed in the previous experiments.

Table 3 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 3), this bias does not affect the models instructed on overall Demonstrations (correct and incorrect). The Llama-2-7 and -13 that have GPT-3.5 as teacher outperform the Llama-2-7 and -13 that have Llama-2-70 as teacher only on OpenBookQA; see OBQA in Figure 2. As far as CommonSenseQA and PIQA are concerned, there is a balancing that is not present in SIQA, where the students of Llama-2-70 outperform the others. Therefore, to study the influence of the quality of Demonstrations, we conducted detailed analyses in Section 4.3.

4.3 Demonstrations-delivering CoT vs Misleading CoT

Demonstrations without Misleading ones yield better students. In addition, the Demonstrations-delivering CoT led to a family-alignment of students' reasoning abilities (Llama-2-7 and -13) with teacher Llama-2-70. In Figure 2, the models instructed on Demonstrations Truthful and Demonstrations-delivering CoT outperformed those instructed on overall Demonstrations and overall CoT-Demonstrations. In particular, the Demonstrations-delivering CoT produced by the in-family teacher outperforms those produced by the out-family teacher. As specified in Figure 12, with the terms "Demonstrations Truthful" and "Demonstrations-delivering CoT", we indicate all correct answers produced by the teacher models.

Hence, in detail, we reproduced the experimental setup proposed in Section 3.2.1. However, unlike previous experiments for Demonstrations and CoT-Demonstrations, we performed Instuction-tuning only for Demonstrations-delivering CoT and Demonstrations Truthful. From the results, these second ones better impact the students models. Furthermore, the subset of Demonstrations used is smaller than the number of total Demonstrations because Misleading instances were discerned. Thus, the students models used fewer instances to perform the tuning.

However, Instruction-tuned students seem to perform better on fewer but distilled Demonstrations. Even more, the Demonstrations-delivering CoT enabled the family-alignment of reasoning abilities. Therefore, in order to observe the true impact of these Demonstrations versus Demonstrations with equal amounts of training instances in Section 4.4, we perform a deep study using different sets.

4.4 The Role of Demonstrations-delivering CoT

Instruction-tuning via Demonstrations-delivering CoT still aligns students' reasoning abilities with those of family teachers, even as instruction decreases. In fact, from Figure 3, we can observe that the performances obtained by students instructed with Demonstrations Truthful (shown with bars) and Demonstrations-delivering CoT (shown with lines) outperform students instructed with overall Demonstrations. Moreover, the Demonstrations-delivering CoT consistently outperforms the Demonstrations Truthful. (technical details about splitting in Appendix A) In conclusion, as also stated in Section 4.3, the Demonstrations-delivering CoT of teacher Llama-2-70 are more productive as all students outperformed the students of teacher GPT-3.5. As they increase, students instructed via in-family teachers increasingly outperform other students.

Finally, to validate our hypothesis of family-alignment, we introduced Mistral-7b (Jiang et al., 2023), a new SLMs that, with 7 billion parameters, outperforms the Llama-2-13 version on several benchmarks as shown by Jiang et al. (2023).

In particular, we reproduced the experiments introduced in Section 4.3 using the different kinds of Demonstrations presented in the previous section. In Figure 4, 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.

5 Related Work

5.1 Chain-of-Thought Prompting

Large Language Models (LLMs) with billions of parameters demonstrate in-context learning and few-shot learning abilities (Brown et al., 2020; Wei et al., 2022; Min et al., 2022) to guide LLMs to generate desired task responses, marking the advent of the prompting era.

These new approaches have surpassed the age of the intermediate steps in algorithmic or structured reasoning Roy and Roth (2015); Ling et al. (2017). Nevertheless, early works challenged the efficacy of few-shot techniques for empowering the prompting phase and downstream performances. In particular, Wang et al. (2022) refined the original idea of Chain-of-Thought (CoT) (Wang et al., 2022) by considering various reasoning paths, while Wang et al. (2023) explored different prompts. Although prompt engineering appears to be the right way to improve performance, many works have used self-generated CoTs to self-improve reasoning ability (Zelikman et al., 2022; Huang et al., 2022; Golovneva et al., 2022).

5.2 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), targets (Talmor et al., 2020) or inputs (Narang 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 Li et al. (2023b); Magister et al. (2023); Shridhar et al. (2023); Ho et al. (2023a). 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., 2023b; Ho et al., 2023a). Li et al. (2023b) claims that massive demonstrations significantly improve performance over the single-sample approach Shridhar et al. (2023).

5.3 Large Language Models as a Teacher

Several papers have been published simultaneously, including those by Magister et al. (2023); Huang et al. (2022), and Ho et al. (2023b) that prove the effect of fine-tuning to transfer the ability to produce Chain-of-Thought (CoT) reasoning from larger to smaller models. Using further fine-tuning, Huang et al. (2022) and Ho et al. (2023b) exploit the known CoT abilities of GPTs (OpenAI, 2023) while Magister et al. (2023) introduces PaLM (Chowdhery et al., 2022) as a teacher. Table 8 resumes these contributions.

Our work goes beyond in the following ways: 1) 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. 2) We investigate which teacher model delivers the most appropriate demonstrations for a student model. In particular, we study the alignment performance between in-family and out-family models on four question-answering benchmarks. 3) Hence, we offer an analysis identifying crucial factors aligning reasoning abilities between teachers and students.

6 Conclusion

In this paper, we analyzed the alignment of reasoning abilities between teachers models, Large Language Models (LLMs), and students models, Small Language Models (SLMs). In particular, we propose the Instruction-tuning-CoT, an instruction tuning via Chain-of-Thought (CoT) demonstrations based on explanations delivered by LLMs CoT prompted. Specifically, we align a set of SLMs using the explanations provided by LLMs that belong to the same family, in-family or out-family. Our results showed the impact of the Instruction-tuning-CoT method both with out-family teachers and particularly with in-family teachers. These results highlight our approach's feasibility in harnessing the multi-step reasoning abilities of LLMs for smaller models designed to pave the way for more efficient and scalable applications.

Limitations

In this paper, we analyzed the impact of Answers delivered by Large Language Models using them as Demonstrations to reinforce the skills of Small Language Models. Although we proposed an extensive study there are several limitations:

- only English-language both in Chain-of-Thought (CoT) methods and tasks evaluation is considered. Although estimating these effects in other languages is interesting, our work only focused on experiments in English. In future works, we intend to take care of this aspect.

- analysis of benchmarks relating to common sense knowledge of social and physical interactions. However, we would like to extend our analyses using more extensive and comprehensive benchmarks such as GSM8K (Cobbe et al., 2021) and MMLU (Hendrycks et al., 2021) in future developments.

- dependence on Large Language Models, 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. Consequently, it is not easy to analyze the differences in pre-training data between models, but observing the outputs in natural language is possible.

In conclusion, 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 have overconfidence in the model based on the CoT. We observed many cases where the CoT examined promising, but ultimately, the models had misleading effects.

Ethical 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

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.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168.

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.

Olga Golovneva, Pan Wei, Khadige Abboud, Charith Peris, Lizhen Tan, and Haiyang Yu. 2022. Task-driven augmented data evaluation. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 18–25, Abu Dhabi, United Arab Emirates (Hybrid). 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.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023a. 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.

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

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023a. 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.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023b. Making language models better reasoners with step-aware verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5315–5333, Toronto, Canada. Association for Computational Linguistics.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, 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.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

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.

Subhro Roy and Dan Roth. 2015. Solving general arithmetic word problems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1743–1752, Lisbon, Portugal. Association for Computational Linguistics.

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.

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.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022. Rationale-augmented ensembles in language 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.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning.

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.

A Experimental Details

In order to observe the impact of the demonstrations (CoT, non-CoT, truthful or Misleading), we produced a series of experiments by systematically decreasing the Instruction-tuning data. In particular, from the total number of demonstrations, we chose three sub-sets with 75%, 50%, and 25%. In detail, the Instruction phases on the number of equal Demonstrations are performed by taking about 3000 examples in splitting 100%, 2250 in splitting 50%, 1500 in splitting 50%, and 750 in splitting 25%. We chose the value 3000 because it is the smallest number of CoT-Gold Demonstrations available. For the total Demonstrations, we selected random samples; instead, for the CoT-Gold and Gold, we selected all the Demonstrations available.

B Accuracy of LLMs on different Benchhmark

Benchmarks	Llama-2-70		GPT-3.5	
	Baseline	CoT	Baseline	CoT
Training				
OBQA	64.6	65.4	66.2	**74.6**
CSQA	70.8	73.4	79.3	**84.8**
SIQA	65.4	67.5	67.6	**70.3**
PIQA	82.3	**85.6**	80.5	84.3
Testing				
OBQA	62.8	64.8	66.7	**73.8**
CSQA	72.4	74.3	80.2	**83.7**
SIQA	64.2	66.9	66.9	**71.3**
PIQA	80.6	84.8	81.6	**85.7**

Table 3: 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).

C Model Sizes

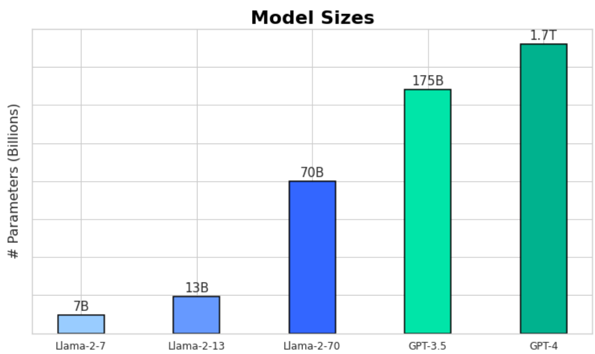

Table 4: Number of parameters of proposed Language Models, B is for Billions and T is for Trillions

D Prompting Approaches

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.

E Models

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

Table 7: List and specific versions of the models proposed in this work, which can be found on `huggingface.co`. For each model we used all the default configurations proposed in the repositories.

Work	**Method**	**Teachers**	**Students**
(Magister et al., 2023)	Fine-tuning	PaLM GPT-3.5	T5-small, -medium T5-large, -xxl
(Li et al., 2023a)	Fine-tuning	GPT-3 175B	OPT-1.3b
(Shridhar et al., 2023)	Fine-tuning	GPT-3 175B	GPT-2
(Ho et al., 2023a)	Fine-tuning	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 8: Summary of methods, teacher and student models of previous work.

F Description of proposed Benchmark

Dataset	Example
Open Book Question Answering (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.
Common Sense Question Answering (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.
Physical Interaction Question Answering (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.
Social Interaction Question Answering (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 9: 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 10: 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 11: In this table, we list the versions of the benchmark proposed in this work, which can be found on huggingface.co.

G Conceptual Map of Names

CoT-Demonstrations
- Demonstrations delivering CoT (as Output: Answer-delivering CoT)
- Misleading CoT (as Output: Answer-misleading CoT)

Demonstrations
- Demonstrations Truthful (as Output: Answer Trthful)
- Misleading (as Output: Answer Misleading)

Table 12: Different types of Demonstrations used in our work. The Demonstrations are composed by: **`Instruction`**, **`Input`** and **`Output`** (see Table 1). Based on the target of the output, there are different types of Demonstrations.

H Overall Pipeline

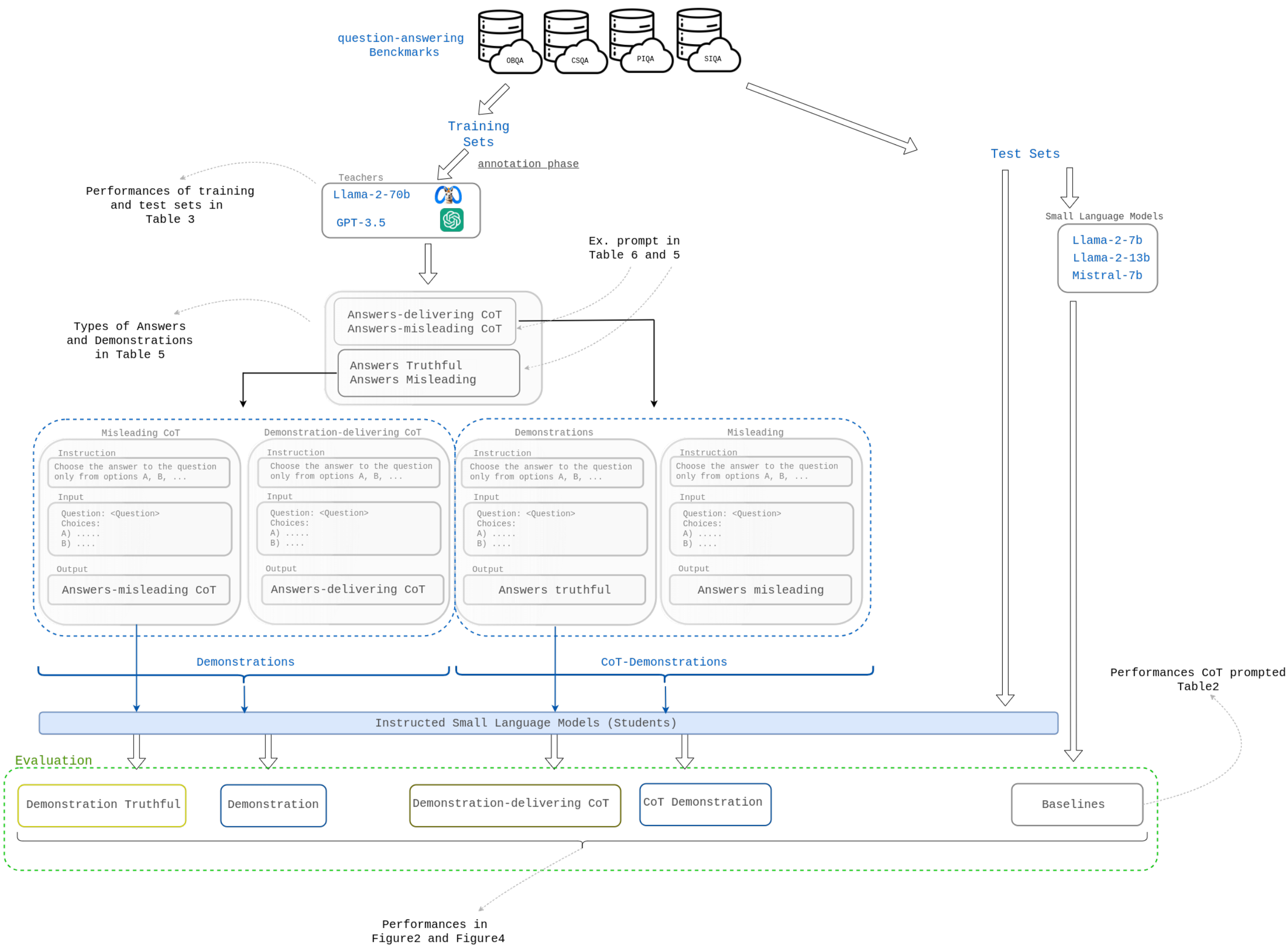

Table 13: Our Experimental Pipeline with a descriptions of data splitting, tables, and results generated.