Aligning Large and Small Language Models via Chain-of-Thought Reasoning

Leonardo Ranaldi [(†)], **André Freitas**[(†,*)]
[(†)] Idiap Research Institute, Switzerland
[(*)]Department of Computer Science, University of Manchester, UK
`[firstname].[lastname]@idiap.ch`

Abstract

Chain-of-Thought (CoT) prompting empowers the reasoning abilities of Large Language Models (LLMs), eliciting them to solve complex reasoning tasks in a step-wise manner. However, these abilities appear only in models with billions of parameters, which represent an entry barrier for many users who are constrained to operate on a smaller model scale, i.e., Small Language Models (SLMs). Although many companies are releasing LLMs of the same family with fewer parameters, these models tend not to preserve all the reasoning capabilities of the original models, including CoT reasoning.

In this paper, we propose a method for aligning and transferring reasoning abilities between larger to smaller Language Models. By using an Instruction-tuning-CoT method, an Instruction-tuning designed around CoT-Demonstrations, we enable the SLMs to generate multi-step controlled reasoned answers when elicited with the CoT mechanism. Hence, we instruct a smaller Language Model using outputs generated by more robust models belonging to the same family or not, evaluating the impact across different types of models. Results obtained on question-answering and mathematical reasoning benchmarks show that LMs instructed via the Instruction-tuning CoT method produced by LLMs outperform baselines within both in-domain and out-domain scenarios.

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 within the GPT(OpenAI, 2023) or PaLM (Chowdhery et al., 2022) families, CoTs enable the delivery of multi-step, controlled reasoning, improving results across commonsense (Bubeck et al., 2023), symbolic and mathematical reasoning datasets (Gaur and Saunshi, 2023; Liu et al., 2023).

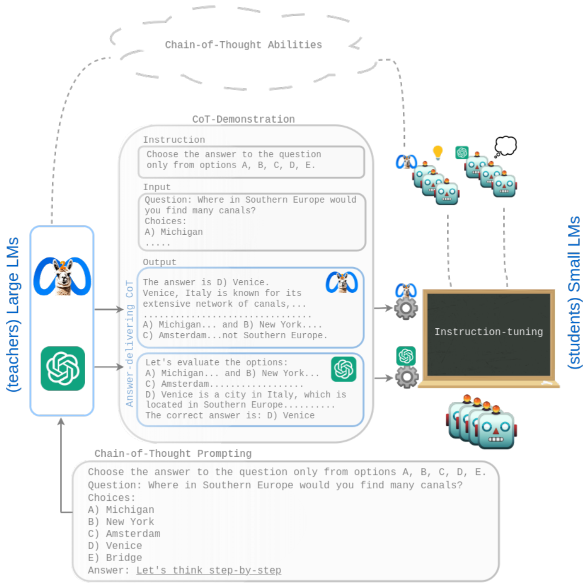

Figure 1: In Instruction-tuning-CoT, students models use CoT-Demonstrations delivered by teacher models. We investigate different properties between the teacher-student models, including the impact of in/out family alignment and the impact of different demonstration styles within the teacher-student alignment.

The size of LLMs, however, presents an adoption barrier for certain users and specific use case scenarios. 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). Although these SLMs are highly functional across different tasks, the CoT prompting mechanism only proved to be consistently operational for models at a certain scale (e.g., with more than 60B parameters (Wei et al., 2023)). These SLMs produce illogical answers when prompted under the CoT framework.

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1812–1827
March 17-22, 2024

In this paper, we propose a method to enable CoT reasoning over SLMs (named student models) by performing Instruction-tuning via demonstrations delivered by LLMs (teacher models). Moreover, we introduce the concept of in-family alignment for teacher-student Instruction-tuning. Hence, we investigate the induction and alignment of Chain-of-Thought reasoning abilities through the support of CoT-Demonstrations "taught" by LLMs teachers to SLMs students (see Figure 1), contrasting between in-family and out-family settings. Complementing the foundation work of (Magister et al., 2023; Shridhar et al., 2023) we introduce the Instruction-tuning CoT approach (i.e., a task-oriented specialization of Supervised Fine-Tuning) through which 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 the student models' performances?

To answer these questions, we selected Llama-2-7b and Llama-2-13b (Touvron et al., 2023) as students and Llama-2-70b and GPT-3.5 as, respectively, in-family and out-family teachers. Then, we conduct an extensive analysis using different types of benchmarks, from arithmetic reasoning to commonsense tasks. Experimentally, we contrast Llama-2-70 and GPT-3.5 as teacher models to deliver CoT-Demonstrations and answers (see Figure 1) which are 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 associated 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 5 shows the terminology used in this work.

We show that the Instruction-tuning approach on Demonstrations instructs student models, and they consistently outperform baseline SLMs in all proposed benchmarks. Finally, students instructed with Demonstrations-delivering CoT provided by the in-family teachers outperformed those instructed by out-family and achieved the best performances.

Our findings can be summarized as follows:

- The Instruction-tuning that is a task-oriented Supervised Fine Tuning (SFT) of SLM students via Demonstrations delivered by an LLM teacher outperformed the non-tuned SLMs (baselines) in terms of downstream performance.

- The Instruction-tuning via CoT-Demonstrations aligns the reasoning abilities of SLMs and LLMs. Models instructed through CoT-Demonstrations that contain outputs generated via CoT prompting outperform models instructed with Demonstrations. In particular, students instructed via CoT-Demonstrations outperform the others both in in-domain and out-domain settings.

- Finally, in-family alignment with Instruction-tuning via Demonstrations-delivering CoTs outperforms out-family alignments.

2 Method

In order to align the reasoning abilities of smaller Language Models using the step-wise reasoning knowledge generated by larger Language Models, we propose a two-phase alignment approach. In the first part, there is an automated 'annotation phase' where the Large Language Models (LLMs) systematically prompt generate outputs (Section 2.1). These Demonstration outputs will be used during the second phase which will perform the Instruction-tuning from the smaller Language Models (Section 2.2).

2.1 Teacher Model

Many state-of-the-art LLMs differ in the number of parameters and training settings. Therefore, we concentrated on larger, widely investigated models with different versions of the same family. As a robust LLMs, we selected Llama-2-70b (Touvron et al., 2023), and GPT-3.5 (OpenAI, 2023). The two models are used to contrast in/out family teacher-student alignment. Another part of the model selection criteria involved the availability of corresponding smaller models within the same family.

We proposed two different prompt types, both in a zero-shot scenario. The first input-prompt is a classic standard prompt, consisting of the question and the associated choices (in case of multiple-choice tasks) as shown in Table 1.

```
Choose the answer to the question only
from options A, B, C, [...].
Question: <Question>
Choices:
A) <Option1>
B) <Option2>
C) <Option3>
....
Answer:
```

Table 1: Example of prompt for multiple-choices question-answering benchmarks.

The second prompt adds the canonical CoT prompt component, where the suffix "**`Let's think step by step`**" is appended (Table 7), in Appendix B. The output provided by the LLM is then used to build an 'annotated set' of Demonstrations.

2.2 Student Model

Although SLMs have been fine-tuned with different techniques, the smaller Language Models[1] do not seem able to reproduce the step-by-step reasoning abilities. Previous work use techniques of knowledge distillation (Li et al., 2023a), curriculum learning (Ranaldi et al., 2023c), skill refinement (Huang et al., 2022), and enriched fine-tuning (Magister et al., 2023) to teach SLMs to deliver multi-step reasoned answers. Following this idea, we propose the Instruction-tuning method using Demonstrations provided by in-family and out-family teachers. Hence, we chose two students SLMs, Llama-2-7b and Llama-2-13b (Touvron et al., 2023), their corresponding "chat" versions (detailed in Table 9), which we will refer to in the paper as Llama-2-7 and -13.

Student models are evaluated before and after the Instruction-tuning. A Demonstration consists of an overall **`Instruction`**: `"Choose the answer to the question only from options A, B,..."` for question answering tasks and `"Answer the following mathematical question with a numerical solution."` for math word problem tasks, an **`Input`** which is the question, and an expected output which, in our case, is the **`Output`** generated by the teachers LLMs. Table 2 shows an example of input. Additional details about the Instruction-tuning steps are provided in Section 3.2.1.

[1]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 2: Example of CoT-Demonstration produced by the GPT-3.5 answer on an Open Book Question Answering benchmark instance. The structure is composed by: **`Instruction`**, **`Input`** and **`Output`**.

3 Experimental Setup

In order to make the experiments comparable with state-of-the-art models, we use two math word problems and four closed-ended question answering benchmarks (introduced in Section 3.1) that are generally used to assess the inference abilities of Large Language Models (LLMs). Moreover, to conduct the Instruction-tuning phase on the Small Language Models (SLMs), we use the approach presented in Section 3.2. The experimental pipeline and the supporting model is available at `github.com/lranaldii/Aligning_LLMs`.

3.1 Tasks & Datasets

In this paper, we selected different benchmarks that focus on reasoning tasks:

Commonsense Task We adopt two benchmarks to evaluate commonsense reasoning: CommonSenseQA (Talmor et al., 2019) (CSQA) and OpenBookQA (Mihaylov et al., 2018) (OBQA) are two multi-choice commonsense question-answering tasks.

Physical & Social Interaction Task We adopt two benchmarks to evaluate the reasoning ability in the context of everyday situations, aiming to establish the most reasonable solution: Interaction Question Answering (PIQA) (Bisk et al., 2019) and Social Interaction Question Answering (SIQA) (Sap et al., 2019), which emphasises people's actions and social implications.

Mathematical Task Finally, we use two math word problem benchmarks to evaluate the models with regard to mathematical reasoning. MultiArith (Roy and Roth, 2015) covers a set of multi-step arithmetic reasoning tasks, while GSM8k (Cobbe et al., 2021) covers a set of primary school-level mathematical problems.

Datasets Since the test split is not prescribed for all the benchmarks, we adopt the following strategy: for SIQA, PIQA, CSQA, and OBQA, we use 4000 examples with equally distributed target classes as training data and the validation versions found on huggingface as test data, while for GSM8K and MultiArith we use the full huggingface datasets. In Table 13, we report the descriptive statistics and splitting ratios, while in Table 12, we report one example for each benchmark. The supporting datasets are publicly accessible as described in Table 14.

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 a zero-shot scenario, as shown in Table 7.

We selected Llama-2-7 and Llama-2-13 (Touvron et al., 2023) as student models, which are fine-tuned using the Instruction-tuning approach, as proposed in (Taori et al., 2023). Finally, we evaluate the performance with evaluation pipelines detailed in Section 3.3. Hence, the SLMs are instructed on the Demonstrations that contain the answers generated by the teachers, as explained in Section 2.2. Table 2 shows an example CoT-Demonstration which contains the Instruction, the Input, and, as Output, the Answer-delivering CoT (in this case generated by a CoT-prompted GPT-3.5).

3.2.1 Models Setup

We conduct the Instruction-tuning phase using QLoRA Dettmers et al. (2023). This approach allows instruction-tuning (and, more generally, fine-tuning) to be performed 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 four epochs and set the learning rate as 0.00002 with a 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 four 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 selected (Brown et al., 2020), and multiple-choice probing, in which the models are asked to commit to an answer. The evaluation in the first case is performed with a function taking the argmax and, in the second case, with a direct string matching. The second method is more widely used in recent evaluations because it can be applied to models from the larger GPT family (OpenAI, 2023) where probability values are not readily accessible.

In our experiments, we chose the latter to have a comparable and scalable pipeline (Details provided in Appendix E.2). Finally, we performed string matching between the generated outputs and the target choice to evaluate the percentages of the correct answers.

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, which report the downstream accuracies without the Instruction-tuning phase (see the Baseline) and the Instruction-tuning phase on Demon-

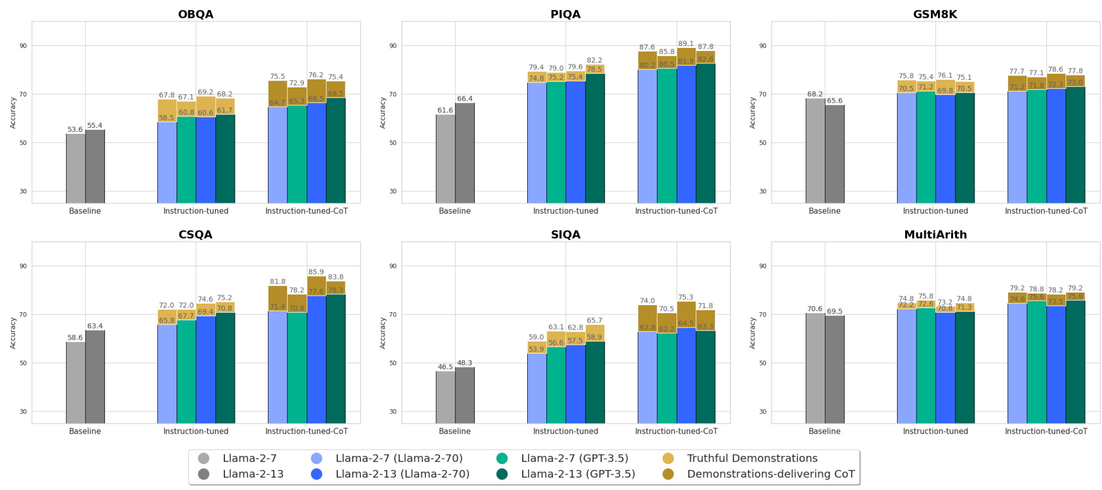

Figure 2: Accuracies (%) on benchmarks (Section 3.1) before Instruction-tuning (i.e., Baselines), on Demonstrations (i.e., Instruction-tuned) and CoT-Demonstrations (i.e., Instruction-tuned-CoT). In addition, Instruction-tuning phases only on Demonstrations-delivering CoT and Truthful Demonstrations,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

strations. 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, an improved alignment between students and teachers can be observed via the Demonstrations-delivering CoT mechanism, as discussed in Section 4.2. In particular, the "Demonstrations-delivering CoT" and "Truthful Demonstrations" 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. Additional details can be found in Section 4.2 and Section 4.5.

4.1 CoT-abilities of Small Language Models

Chain-of-thought (CoT) prompts do not always deliver downstream performance improvements. SLMs have not performance improvements when prompted with the CoT mechanism. In particular, we evaluated performance on four question-answering benchmarks, described in Section 3.1, using Llama-2-chat (7b-13b billion) in a zero-shot scenario. Proposing a classical prompt (Baseline) and a CoT prompt, we obtained the performances in Table 3.

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. Using CoT prompting leads to model confusion with the degradation of downstream results. It is possible to observe these phenomena in OpenBookQA (OBQA) and CommonSenseQA (CSQA) (Table 3). In particular, there is a marked deterioration in Llama-2-7 (see ⇓), which has half the parameters of Llama-2-13 (see ↓). This behaviour is not observable in PIQA and SIQA, which have tasks consisting of fewer answer choices. In this setting, this is likely to be explained by a possible lower inference complexity induced by the smaller answer sets (as shown in Table 13).

4.2 The Instruction-tuning Impact

Instruction-tuning supported by Large Language Models (teachers models) was able to guide the Smaller Language Models (students models) to deliver a step-wise reasoning. This can be observed in the experimental outcomes of Figure 2. The student models based on Instruction-tuning on Demonstrations produced by teacher models outperformed the baselines in the four proposed benchmarks.

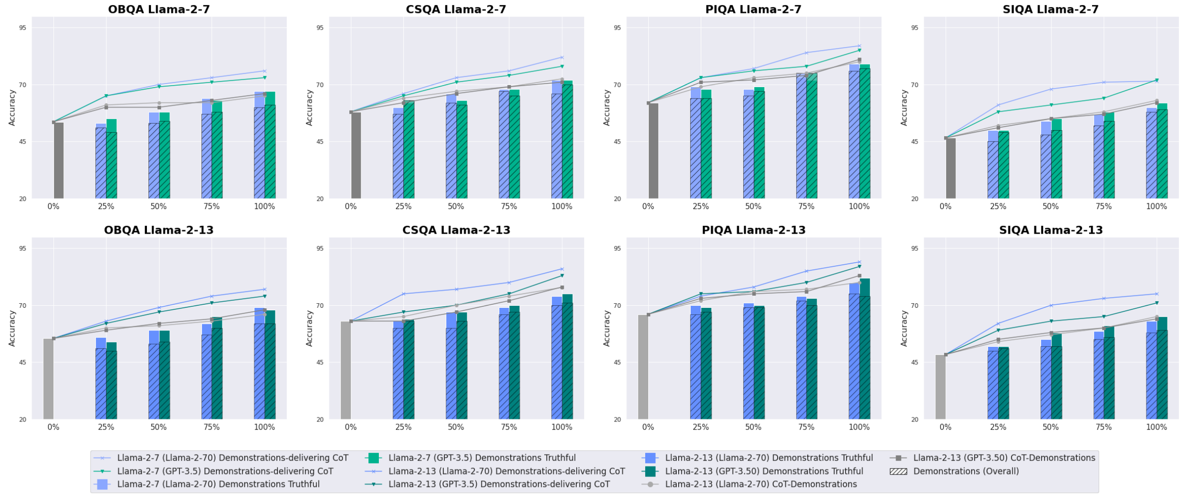

Figure 3: Accuracies (%) on the test set of benchmarks. Instruction-tuning performed on different splits (see Appendix E.1 for additional details) of Demonstrations and CoT-Demonstrations (correct and not correct predictions), Truthful Demonstrations, and Demonstrations-delivering CoT (correct predictions).

Task	**Llama-2-7**		**Llama-2-13**	
	Baseline	**CoT**	**Baseline**	**CoT**
OBQA	**53.6**±.2	49.5±.3⇓	**55.4**±.2	54.2±.3↓
CSQA	**58.6**±.3	50.6±.1⇓	**63.4**±.2	60.8±.2↓
SIQA	46.5±.2	*45.3*±.3	48.3±.4	46.9±.3
PIQA	61.6±.2	*63.8*±.2	66.4±.1	71.2±.3
GSM8K	68.2±.3	**71.3**±.3	65.6±.4	70.5±.1
MultiArith	69.5±.2	**72.6**±.3	67.2±.2	70.8±.4

Table 3: Accuracies of Llama-2-7 and Llama-2-13, both without further tuning, on testing data with the standard prompt (Baseline) (see Table 6) and CoT prompt (CoT) (see Table 7).

Moreover, the students models instructed with CoT-Demonstrations, defined as Instruction-tuned-CoT in Figure 2, achieved best accuracy.

While there are performance improvements across the board, this analysis can be nuanced by looking into the specific characteristics of the reference models, for example, in terms of parameters GPT-3.5 (175B parameters) versus Llama-2-70 (70B). This is reflected in performance differences within the proposed benchmarks. Table 11 shows the performance (with and without CoT prompting) on the data used to conduct the Instruction-tuning phase and on the same test set used to evaluate the proposed models.

Although the performance delivered on the "training set" is different across different models (see the CoT performances of GPT-3.5 and the same for Llama-2-70 in Table 11), 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 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. 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

Instruction-tuning through consistent Demonstrations performs better than that done on Demonstrations with misaligned answers. 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 Truthful Demonstrations 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 5, with the terms "Demonstrations Truthfu" and "Demonstrations-delivering CoT", we indicate all correct answers produced by the teacher models.

Using the basic experimental setup proposed in Section 3.2.1 we performed Instuction-tuning only for Demonstrations-delivering CoTs and Demonstrations Truthful. From the results, the latter mechanism further improves the performance of the students models. Furthermore, the subset of Demon-

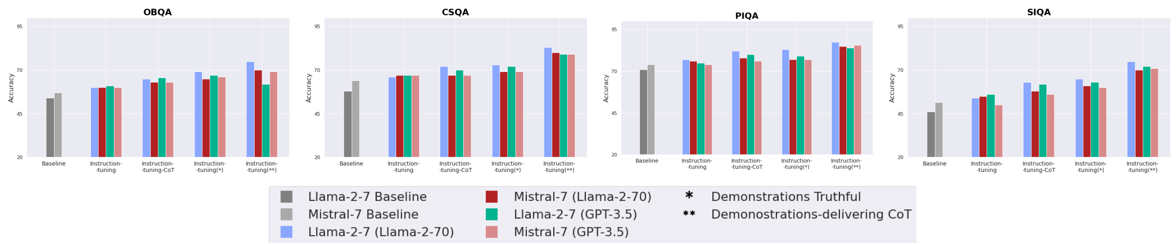

Figure 4: Performance of Llama-2-7 and Mistral-7 Instruction-tuned using the same setup proposed in the previous experiments.

strations used is smaller than the number of total Demonstrations because Misleading instances were removed. Thus, the students models used comparatively fewer instances.

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 further analysis 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. From Figure 3, we can observe that the performance 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 E.1) 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, student models instructed via in-family teachers increasingly outperform other student model types.

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 Llama-2-13 on several benchmarks as shown by Jiang et al. (2023). In particular, we reproduced the experiments introduced in Section 4.3 using the different types of Demonstrations presented in the previous section. In Figure 4, it can be observed that Llama-2-7 instructed on different types of Demonstrations delivered by Llama-2-70 outperforms Mistral-7b in most cases. These results confirm that Demonstrations derived from in-family teachers have a more significant impact on student models than the others.

4.5 In-Domain and Out-Domain

Instruction-tuning through CoT-Demonstrations enables student models to cover both in-domain and out-of-domain tasks. Figure 4 shows the results of student Llama-2-7 and Figure 15 of Llama-2-13. In both cases, it can be observed that the instructed models always outperform the baselines. However, as expected, models instructed on in-domain scenarios (e.g., two QA tasks with different seeds) achieve significantly better results when contrasted to models instructed on out-domain scenarios (e.g., instruction via QA demonstrations and tests on mathematical problems).

Finally, as shown in Table 16, it is possible to observe that the performance obtained by the instructed models consistently surpasses the baselines on evaluation benchmarks. This shows that (i) the instruction-tuning process does not degrade baseline performance, and (ii) the instructed models outperform the uninstructed even on tasks they were not trained, showing they have learned generalization abilities.

5 Related Work

5.1 Chain-of-Thought Prompting

Large Language Models with billions of parameters demonstrate in-context learning and few-shot learning properties (Brown et al., 2020; Wei et al., 2022) to guide LLMs to generate desired task responses, marking the transition towards a prevalent prompting-based paradigm. Zero and few-shot prompting methods, in particular in complex reasoning settings, have been extended and refined to accommodate the multi-step nature of different

Trained on	Teacher	Evaluated on					
		OBQA	CSQA	PIQA	SIQA	GMS8K	MultiArith
Baseline	-	53.6±.2	58.6±.4	61.6±.1	46.5±.3	68.2±.5	69.5±.2
Baseline CoT	-	49.5±.4	50.6±.3	63.8±.1	71.3±.5	71.3±.2	72.6±.4
OBQA	GPT-3.5	72.9±.3	65.3±.2	74.6±.5	64.3±.2	67.6±.4	68.6±.3
	Llama-2-70	**75.5**±.4	76.2±.5	75.1±.2	65.2±.4	68.2±.2	69.2±.4
CSQA	GPT-3.5	68.5±.2	78.2±.5	82.2±.1	65.3±.3	65.9±.4	68.3±.2
	Llama-2-70	67.8±.3	**81.8**±.4	81.9±.1	66.2±.5	66.1±.2	67.5±.3
PIQA	GPT-3.5	63.6±.4	64.3±.5	85.8±.2	56.8±.1	61.2±.3	64.4±.2
	Llama-2-70	64.3±.1	65.2±.2	**87.6**±.3	57.2±.4	60.7±.5	65.3±.4
SIQA	GPT-3.5	65.2±.2	63.8±.1	79.4±.3	70.5±.4	63.2±.5	66.9±.4
	Llama-2-70	65.6±.5	64.1±.4	80.3±.2	**74.0**±.1	62.4±.2	66.3±.3
GSM8K	GPT-3.5	55.6±.3	56.2±.4	60.3±.1	50.7±.2	77.1±.5	78.4±.4
	Llama-2-70	55.8±.5	55.9±.2	59.6±.3	52.3±.2	**77.7**±.2	77.9±.3
MultiArith	GPT-3.5	55.7±.2	57.6±.5	60.5±.3	50.6±.1	75.9±.4	78.8±.2
	Llama-2-70	55.4±.4	57.8±.1	59.9±.2	51.6±.5	76.2±.3	**79.2**±.5

Table 4: Evaluation of Llama-2-7 instructed on CoT-Demonstrations using different test sets. We evaluate in-domain (QA vs QA) and out-domain (QA vs math-word problem) benchmarks. "Baseline" refers to the non-instructed model. Results colored in green indicate the in-domain benchmark, blue the out-domain benchmark, and orange the same benchmark on which perform the evaluation phase.

tasks. Wang et al. (2022) refined the original idea of Chain-of-Thought (CoT) (Wang et al., 2022) by considering different reasoning paths, while Wang et al. (2023) explored different prompting properties. Emerging methods include self-generated CoTs (Ranaldi and Zanzotto, 2023; Zelikman et al., 2022; Huang et al., 2022).

5.2 Learning from Explanations

Contemporary methods include the conditioning of models on specific task instructions and provide explanations for individual data points to 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 CoT builds upon prior efforts wherein explanations are viewed as intermediary constructs produced during inference (Rajani et al., 2019).

Our research is based on the foundation built by Li et al. (2023b); Magister et al. (2023); Shridhar et al. (2023); Ho et al. (2023a). In particular, we adopt the Teacher-Student model configuration (in our case teacher LLMs and student SLMs) (Magister et al., 2023). Learning uses teacher-generated explanations, demonstrating the impact of CoT prompts on downstream tasks (Li et al., 2023b; Ho et al., 2023a). Li et al. (2023b) claiming that larger sets of demonstrations significantly improve performance over a single-sample approach Shridhar et al. (2023).

5.3 Large Language Models as a Teacher

Previous work, including Magister et al. (2023); Huang et al. (2022), and Ho et al. (2023b) focused on the analysis of the effect of fine-tuning as a mechanism to transfer the ability to produce Chain-of-Thought (CoT) reasoning from larger to smaller models, using both GPT-type (OpenAI, 2023) Huang et al. (2022); Ho et al. (2023b) and PaLM Magister et al. (2023) models. Table 10 summarizes these contributions.

This work extends these foundational contributions by investigating the particular CoT and model features that contribute to supporting CoT learning in the teacher-student model setting, including in/out family alignment and the analysis across different commonsense and mathematical reasoning benchmarks.

6 Conclusion

In this paper, we analyzed the alignment of stepwise CoT reasoning between teacher Large Language Models (LLMs) and student Small Language Models (SLMs). In particular, we propose Instruction-tuning-CoT, an instruction tuning via

Chain-of-Thought (CoT) demonstrations, based on explanations delivered by LLMs prompted with the CoT mechanism. We also contrast the impact of in-family and out-family alignment across teacher and student models. The results highlight the impact of teacher-student Instruction-tuning interventions as a mechanism to improve the step-wise reasoning properties of smaller language models.

Limitations

In our contribution, we analyzed the impact of Answers delivered by Large Language Models, using them as Demonstrations to improve the step-wise reasoning properties of Small Language Models. The first limitation is in relation to the target languages which is constrained to English. In future work, we will investigate this aspect starting from Cross-lingual alignment approaches (Ranaldi et al., 2023b).

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, but observing the outputs in natural language is possible (Ranaldi et al., 2023a; Ranaldi and Pucci, 2023). 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 correct, there is no guarantee that the predicted output comes from a consistent and faithful reasoning process. Future work includes improving the understanding of the specific CoT alignment mechanisms by using more granular interpretability mechanisms.

Ethics Statement

Although this research intervention was able to demonstrate an improvement in the reasoning abilities of Smaller Language Models, further investigation is required to understand the exact mechanisms that are in place with regard to the transference of step-wise CoT reasoning from larger to smaller models. This improved understanding is required to develop robust real-world applications in domains such as education, law and clinical reasoning.

Acknowledgements

This paper has been partially supported by the SINFONIA project funded by Innosuisse n. 104.170 IP-ICT. We would like to thank the reviewers for their constructive comments.

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.

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. 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.

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.

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.

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, Aria Nourbakhsh, Elena Sofia Ruzzetti, Arianna Patrizi, Dario Onorati, Michele Mastromattei, Francesca Fallucchi, and Fabio Massimo Zanzotto. 2023a. The dark side of the language: Pre-trained transformers in the DarkNet. In Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, pages 949–960, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Leonardo Ranaldi and Giulia Pucci. 2023. Knowing knowledge: Epistemological study of knowledge in transformers. Applied Sciences, 13(2).

Leonardo Ranaldi, Giulia Pucci, and Andre Freitas. 2023b. Empowering cross-lingual abilities of instruction-tuned large language models by translation-following demonstrations.

Leonardo Ranaldi, Giulia Pucci, and Fabio Massimo Zanzotto. 2023c. Modeling easiness for training transformers with curriculum learning. In Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, pages 937–948, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Leonardo Ranaldi and Fabio Massimo Zanzotto. 2023. Empowering multi-step reasoning across languages via tree-of-thoughts.

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 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 5: Different types of Demonstrations used in our work. The Demonstrations are composed by: **`Instruction`**, **`Input`** and **`Output`** (see Table 2). Based on the target of the output, there are different types of Demonstrations.

B Prompting Approaches

Prompt for task: OBQA, CSQA, PIQA, SIQA
Choose the answer to the question only from options A, B, C, [...]. **Question:** <Question> **Choices:** A) <Option1> B) <Option2> C) <Option3> .... **Answer:**

Prompt for task: GSM8k, MultiArith
Answer the following mathematical question with numerical solution. **Question:** <Question> **Answer:**

Table 6: Example of input-prompt for multiple-choices (left) and mathematical (right) question-answering benchmarks.

Prompt for task: OBQA, CSQA, PIQA, SIQA
Choose the answer to the question only from options A, B, C, [...]. **Question:** <Question> **Choices:** A) <Option1> B) <Option2> C) <Option3> .... **Answer: Let's think step by step**

Prompt for task: GSM8k, MultiArith
Answer the following mathematical question with numerical solution. **Question:** <Question> **Answer: Let's think step by step**

Table 7: Example **Zero-shot CoT** of input-prompt for multiple-choices (left) and mathematical (right) question-answering benchmarks.

C Examples Misleading Answers Llama-2-7b

Example for task: PIQA
Choose the answer to the question only from options A, B **Question:** How do you properly prepare a steak? **Choices:** A) Take the steak out of warm storage and let come to room temperature, generously add salt and pepper to both sides and let sit for 10 minutes. B) Take the steak out of cold storage and let come to room temperature, generously add salt and pepper to both sides and let sit for 10 minutes. **Answer: Let's think step by step**

Example for task: MultiArith
Answer the following mathematical question with numerical solution. **Question:** Mike invited 13 friends to a birthday party, but 7 couldn't come. If he wanted to buy enough cupcakes so each person could have exactly 4, how many should he buy? **Answer: Let's think step by step**

Table 8: Examples of two Zero-Shot Chain-of-Thought prompting from Physical Interaction Question Answering (left) and MultiArith (right). In the example on the left, the number of choices depends on the composition of the task.

D 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 9: 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)	SFT	PaLM GPT-3.5	T5-small, -medium T5-large, -xxl
(Li et al., 2023a)	SFT	GPT-3 175B	OPT-1.3b
(Shridhar et al., 2023)	SFT	GPT-3 175B	GPT-2
(Ho et al., 2023a)	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 10: Summary of methods, teacher and student models of previous work, we indicate Supervised Fine-tuning as (SFT) employed in most previous work.

E Experimental Details

E.1 Data Splitting

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.

E.2 Parameters

The annotation phase that the Teachers performed was done on the training set. The evaluation phase of both the basic models and the Students and the Teachers was done on the test splitting. The evaluation, described in Section 3.3, was done with question probing and string matching of the generated answers. More specifically:

Teachers We performed the annotation phase for each benchmark by delivering to `GPT-3.5-turbo` and `Llama-2-70-chat` the prompts structured as shown in Table 6 and Table 7 (customized for each benchmark). We set the temperatures to 0.7 for `GPT-3.5-turbo` and 0.1 for `Llama-2-70-chat` as recommended in technical reports. Moreover, we kept all the other parameters as default. All parameters are shown in our code.

Baseline & Students We evaluated the performance of the Small Language Models (`Llama-2-7-chat`, `Llama-2-13-chat`, `Mistral-7b`) by prompting them with the same format used for the Teachers. For both the baselines and the instructed models, we set the temperature to 0.1 and kept all the other parameters as default.

F Accuracy of LLMs on different Benchhmark

Benchmarks	**Llama-2-70**		**GPT-3.5**	
	Baseline	**CoT**	**Baseline**	**CoT**
Training				
OpenBook QA	65.6$_{\pm.3}$	71.3$_{\pm.1}$	66.2$_{\pm.2}$	**75.4**$_{\pm.4}$
CommonSesnse QA	74.2$_{\pm.1}$	79.6$_{\pm.3}$	79.3$_{\pm.4}$	**84.8**$_{\pm.1}$
Social Interaction QA	65.4$_{\pm.2}$	67.5$_{\pm.1}$	67.6$_{\pm.5}$	**70.3**$_{\pm.4}$
Physical Interaction QA	82.6$_{\pm.2}$	**85.8**$_{\pm.2\pm.3}$	83.5$_{\pm.3}$	85.3$_{\pm.2}$
GSM8K	74.6$_{\pm.1}$	77.2$_{\pm.2}$	83.2$_{\pm.2}$	**86.5**$_{\pm.2}$
MultiArith	88.6$_{\pm.1}$	90.8$_{\pm.3}$	94.9$_{\pm.4}$	**96.7**$_{\pm.2}$
Testing				
OpenBook QA	65.9$_{\pm.2}$	70.8$_{\pm.1}$	67.8$_{\pm.1}$	**74.6**$_{\pm.4}$
CommonSesnse QA	73.4$_{\pm.2}$	81.8$_{\pm.3}$	80.2$_{\pm.2}$	**83.7**$_{\pm.1}$
Social Interaction QA	64.2$_{\pm.2}$	66.9$_{\pm.4}$	66.9	**71.3**$_{\pm.3}$
Physical Interaction QA	82.6$_{\pm.3}$	85.6$_{\pm.5}$	84.3$_{\pm.2}$	**85.8**$_{\pm.5}$
GSM8K	75.2$_{\pm.5}$	77.8$_{\pm.5}$	82.8$_{\pm.2}$	**84.6**$_{\pm.4}$
MultiArith	89.2$_{\pm.3}$	92.3$_{\pm.2}$	95.6$_{\pm.2}$	**97.4**$_{\pm3}$

Table 11: 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).

G 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.
(GSM8K) (Cobbe et al., 2021)	Tina makes $18.00 an hour. If she works more than 8 hours per shift, she is eligible for overtime, which is paid by your wage + 1/2 your hourly hourly wage. If she works 10 hours every day for 5 days, how much money does she make?
(MultiArith) (Roy and Roth, 2015)	Chloe was playing a video game where she scores 9 points for each treasure she finds. If she found 6 treasures on the first level and 3 on the second, what would her score be?

Table 12: Examples of the benchmarks used in this paper.

	OBQA	CSQA	PIQA	SIQA	GSM8K	MultiArith
classes	4	5	2	3	-	-
Training # examples for each class	1000	800	2000	1330	4000	420
Test # examples for each class	125* (± 8)	235* (± 11)	924* (± 18)	640* (± 19)	1318	180

Table 13: 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. GMS8K e MultiArith are not closed-ended question answering; they only have a question and a numerical solution.

Name	Repository
CommonSenseQA (Talmor et al., 2019)	`huggingface.co/datasets/commonsense_qa`
OpenBookQA (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`
GSM8K (Cobbe et al., 2021)	`huggingface.co/datasets/gsm8k`
MultiArith (Roy and Roth, 2015)	`huggingface.co/datasets/ChilleD/MultiArith`

Table 14: In this table, we list the versions of the benchmark proposed in this work, which can be found on huggingface.co.

Trained on	Teacher	Evaluated on					
		OBQA	**CSQA**	**PIQA**	**SIQA**	**GMS8K**	**MultiArith**
Baseline	-	55.4±.2	63.4±.2	66.4±.1	48.3±.4	65.6±.4	67.2±.2
Baseline CoT	-	54.2±.3	60.8±.2	71.2±.3	46.9±.3	70.5±.1	70.8±.4
OBQA	GPT-3.5	75.4±.4	66.2±.3	75.3±.6	63.2±.3	67.8±.2	69.4±.2
	Llama-2-70	**76.2**±.2	77.3±.4	75.6±.1	66.3±.3	67.9±.3	70.1±.2
CSQA	GPT-3.5	69.4±.3	83.8±.4	82.6±.2	66.2±.4	66.3±.4	70.1±.2
	Llama-2-70	68.9±.3	**85.9**±.3	81.9±.1	66.2±.5	66.8±.1	68.6±.3
PIQA	GPT-3.5	64.1±.3	64.6±.5	87.8±.3	57.2±.1	60.9±.4	66.9±.1
	Llama-2-70	65.3±.3	65.8±.5	**89.1**±.4	58.4±.3	65.9±.2	66.3±.2
SIQA	GPT-3.5	66.4±.3	64.3±.2	80.2±.4	71.8±.3	64.9±.3	67.6±.4
	Llama-2-70	67.2±.3	64.9±.3	81.9±.3	**75.3**±.2	66.1±.3	66.8±.2
GSM8K	GPT-3.5	56.8±.3	58.6±.5	62.3±.2	51.8±.3	77.8±.4	77.9±.3
	Llama-2-70	57.8±.2	56.3±.3	60.3±.4	54.2±.3	**78.6**±.3	**79.2**±.2
MultiArith	GPT-3.5	56.7±.3	57.9±.3	60.5±.3	50.6±.1	75.9±.4	78.8±.2
	Llama-2-70	55.4±.4	59.8±.1	60.3±.2	52.8±.4	76.4±.4	78.2±.3

Table 15: Evaluation of Llama-2-13 instructed on CoT-Demonstrations using different test sets. We evaluate in-domain (QA vs QA) and out-domain (QA vs math-word problem) benchmarks. "Baseline" refers to the non-instructed model. Results colored in green indicate the in-domain benchmark, blue the out-domain benchmark, and orange the same benchmark on which perform the evaluation phase.

Trained on	Teacher	Evaluated on			
		BBH (Llama-2-7)	**BBH (Llama-2-13)**	**MMLU (Llama-2-7)**	**MMLU (Llama-2-13)**
Baseline	-	32.8±.3	39.4±.5	45.3±.2	55.2±.3
Baseline CoT	-	33.5±.2	38.2±.3	44.8±.2	56.3±.2
OBQA	GPT-3.5	34.3±.3	39.9±.4	45.2±.3	55.8±.2
	Llama-2-70	33.9±.3	40.7±.3	45.8±.3	54.9±.4
CSQA	GPT-3.5	34.2±.4	39.2±.3	45.9±.4	56.1±.3
	Llama-2-70	33.9±.2	40.2±.2	46.2±.4	55.3±.1
PIQA	GPT-3.5	33.2±.5	38.9±.3	44.8±.6	55.9±.2
	Llama-2-70	33.9±.2	39.2±.2	46.2±.3	55.3±.1
SIQA	GPT-3.5	32.9±.1	38.2±.4	45.2±.3	55.3±.1
	Llama-2-70	33.2±.3	39.5±.4	45.7±.2	54.9±.2
GSM8K	GPT-3.5	35.6±.2	39.3±.2	45.9±.2	56.1±.1
	Llama-2-70	35.8±.3	39.3±.4	46.3±.3	56.3±.2
MultiArith	GPT-3.5	34.7±.4	39.2±.1	45.7±.4	56.3±.2
	Llama-2-70	34.9±.2	40.3±.3	46.8±.1	55.9±.3

Table 16: Evaluation of Llama-2-7 instructed on CoT-Demonstrations delivered in different settings (on columns) on BBH and MMLU.