# Supplementary Material for: Can Language Models Teach? Teacher Explanations Improve Student Performance via Theory of Mind

**Anonymous Author(s)**
Affiliation
Address
email

## 1 Datasets and Prompts

We experiment with the following three reasoning datasets: (1) StrategyQA [1], a set of open-domain questions where the required reasoning steps are implicit in the question, (2) GSM8k [2], which includes multi-step math reasoning problems, and (3) CommonsenseQA [3], a multiple-choice QA task focusing on commonsense reasoning. We use the reasoning steps in StrategyQA and GSM8k as the multi-step rationales, and for CommonsenseQA, we rely on the ECQA dataset [4], which is annotated with commonsense facts supporting the correct option and refuting the incorrect options. All datasets are licensed under the MIT license. Fig. 1 shows the student prompts for the three tasks of StrategyQA, CommonsenseQA, and GSM8k. Fig. 2 shows the pre- and post-intervention student simulation prompts for the teacher model.

## 2 Compute and Reproducibility

We conduct experiments either on A100 Google Cloud instances or on internal A6000 GPU servers. The LLMs (Flan-T5 and LLama) and the datasets used in our studies are publicly available. For reproducibility, we are making our code available as part of the supplementary material.

## 3 RQ1: Additional Results

**Results with Other Models and Datasets.** In Table 1, we report the accuracy obtained by different students and teachers (based on Flan-T5 models) on the StrategyQA task. We draw similar conclusions as Flan-T5 with other LLMs, specifically LLama-7B and LLama-13B models on the StrategyQA dataset (Figure 3(a), Table 2). Unlike Flan models, LLama-7B and LLama-13B do not exhibit significant differences in accuracy at no intervention (0%) but the trends still align – increasing for weaker students and decreasing for stronger students. Our conclusions generalize across datasets too. Figure 3(b) and Table 3 present the results on CommonsenseQA with Flan-T5 models. CommonsenseQA is an easier task and Flan-T5 models obtain accuracies of 85% and 92% when generating their own explanations. While Flan-T5-Large still benefits from human explanations, the larger model does not, perhaps because it already starts at a high 92% accuracy. Finally, in Figure 3(c) and Table 4, we present the results on GSM8k with LLama models. Note that in GSM8k, a student has access to partial explanations from the teacher, but even then we observe that these prove to be useful prompts for the student to complete their chain-of-thought, leading to up to 8-9% increase in accuracy with human teachers and 3% with model teachers.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

**StrategyQA**

**Q:** Are more people today related to Genghis Khan than Julius Caesar?
**A:** Julius Caesar had three children. Genghis Khan had sixteen children. Modern geneticists have determined that out of every 200 men today has DNA that can be traced to Genghis Khan. So the answer is yes

**Q:** {test_question}
**A:**

**CommonsenseQA**

**Q:** What might a person see at the scene of a brutal killing?
**Answer Choices:**
**Choice 1:** bloody mess
**Choice 2:** pleasure
**Choice 3:** being imprisoned
**Choice 4:** feeling of guilt
**Choice 5:** cake
**A:** Bloody mess is covered or stained with blood. A person might see a bloody mess at the scene of a brutal killing. Pleasure is about what a person sees at the scene of a brutal killing and one cannot be happy to see such brutality. You can't see someone in jail at the brutal killing scene. Feeling of guilt doesn't come as the killing is brutal or merciless. Cake is baseless and weird to think as it is a brutal killing scene and not a bakery. So the correct choice is 1

**Q:** {test_question}
**Answer Choices:**
**Choice 1:** {option_1}
**Choice 2:** {option_2}
**Choice 3:** {option_3}
**Choice 4:** {option_4}
**Choice 5:** {option_5}
**A:**

**GSM8k**

**Q:** Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?
**A:** Natalia sold 48/2 = 24 clips in May. Natalia sold 48+24 = 72 clips altogether in April and May. So the answer is 72

**Q:** {test_question}
**A:**

**Figure 1:** Examples of student prompts for different tasks with one demonstration.

| | | Intervention Budget | | | | | |
|---|---|---|---|---|---|---|---|
| Student | Teacher | 0% | 20% | 40% | 60% | 80% | 100% |
| Flan-T5-Large | Human | 58.51±2.00 | 63.75±0.43 | 66.95±2.19 | 73.94±2.77 | 78.02±2.40 | 81.95±1.65 |
| Flan-T5-XL | Human | 68.12±2.62 | 72.05±2.62 | 75.98±2.31 | 80.20±1.65 | 84.13±1.00 | 87.77±0.70 |
| Flan-T5-Large | Flan-T5-XL | 58.51±2.00 | 60.52±1.63 | 59.78±1.85 | 61.48±2.02 | 62.35±2.13 | 62.96±2.47 |
| Flan-T5-XL | Flan-T5-Large | 68.12±2.62 | 67.68±2.72 | 65.64±3.39 | 64.04±3.63 | 62.88±1.15 | 61.86±0.66 |

**Table 1:** RQ1 – Comparison of accuracy obtained with random intervention by different Teacher Models on different Student Models at different intervention budgets on StrategyQA.

**Results with Cross-family Student and Teacher.** We observe that larger teacher LLMs can teach smaller student LLMs, even when they are of different model families. In Table 5, we report the results with Flan-T5 and LLama models as students and teachers.

**Pre-Intervention Student Simulation**

Simulate an AI model's answer for the given question.

**Q:** Will the Albany in Georgia reach a hundred thousand occupants before the one in New York?
**AI Predicted Answer:** Albany, Georgia is a city in the U.S. state of Georgia. Albany, Georgia has a population of 59,080. The population of New York is 365,040. So the answer is no

**Q:** {question}
**AI Predicted Answer:**

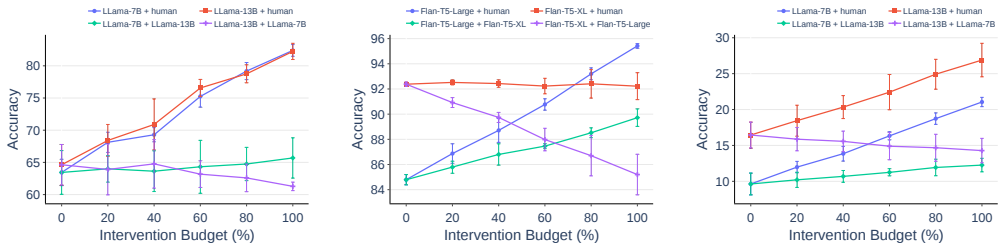
**Post-Intervention Student Simulation**

Simulate an AI model's answer for the given question.

**Q:** Will the Albany in Georgia reach a hundred thousand occupants before the one in New York?
**AI Predicted Answer:** Albany, Georgia is a city in the U.S. state of Georgia. Albany, Georgia has a population of 59,058. The Albany in New York has a population of 328,058. So the answer is no

**Q:** {question}
**AI Predicted Answer:** {teacher_explanation} So the answer is

**Figure 2:** Examples of StrategyQA prompts for the mental model of a teacher simulating student predictions pre-intervention and post-intervention. Pre-intervention: The demonstrations use student explanations and student predictions and at test time, the teacher simulates both. Post-intervention: The demonstrations use teacher explanations and student predictions and at test time, the teacher uses the teacher explanation to simulate the student prediction.




(a) (b) (c)

**Figure 3:** RQ1: Comparison of accuracy obtained with random intervention by different Teacher Models on different Student Models at different intervention budgets. (a) LLama models on StrategyQA. (b) Flan-T5 models on CommonsenseQA. (c) LLama models on GSM8k. A+B in the legends denote A is the student while B is the teacher.

| | | Intervention Budget | | | | | |
|---|---|---|---|---|---|---|---|
| Student | Teacher | 0% | 20% | 40% | 60% | 80% | 100% |
| LLama-7B | Human | 63.44±3.38 | 68.12±1.57 | 69.28±1.09 | 75.25±1.65 | 79.18±1.33 | 82.38±0.90 |
| LLama-13B | Human | 64.62±3.14 | 68.41±2.48 | 70.88±3.96 | 76.56±1.33 | 78.74±1.40 | 82.24±1.26 |
| LLama-7B | LLama-13B | 63.44±3.38 | 64.00±4.05 | 63.63±3.15 | 64.31±4.10 | 64.75±4.59 | 65.68±4.11 |
| LLama-13B | LLama-7B | 64.62±3.14 | 63.90±3.96 | 64.77±3.76 | 63.17±2.06 | 62.59±2.15 | 61.28±0.66 |

**Table 2:** RQ1 – Comparison of random intervention by different Teacher Models on different Student Models at different intervention budgets on StrategyQA.

| Student | Teacher | Intervention Budget | | | | | |
|---|---|---|---|---|---|---|---|
| | | 0% | 20% | 40% | 60% | 80% | 100% |
| Flan-T5-Large | Human | 84.78±0.41 | 86.86±0.76 | 88.70±0.94 | 90.77±0.45 | 93.20±0.47 | 95.42±0.17 |
| Flan-T5-XL | Human | 92.38±0.16 | 92.52±0.20 | 92.43±0.28 | 92.23±0.61 | 92.41±1.12 | 92.21±1.06 |
| Flan-T5-Large | Flan-T5-XL | 84.78±0.41 | 85.79±0.48 | 86.79±0.84 | 87.46±0.20 | 88.52±0.39 | 89.72±0.68 |
| Flan-T5-XL | Flan-T5-Large | 92.38±0.16 | 90.92±0.39 | 89.74±0.39 | 87.98±0.89 | 86.70±1.60 | 85.19±1.62 |

**Table 3:** RQ1 – Comparison of random intervention by different Teacher Models on different Student Models at different intervention budgets on CommonsenseQA.

| Student | Teacher | Intervention Budget | | | | | |
|---|---|---|---|---|---|---|---|
| | | 0% | 20% | 40% | 60% | 80% | 100% |
| LLama-7B | Human | 9.62±1.53 | 11.97±0.80 | 13.84±1.02 | 16.32±0.57 | 18.72±0.78 | 21.05±0.65 |
| LLama-13B | Human | 16.45±1.80 | 18.44±2.16 | 20.34±1.60 | 22.41±2.46 | 24.91±2.07 | 26.88±2.34 |
| LLama-7B | LLama-13B | 9.62±1.53 | 10.20±1.06 | 10.68±0.82 | 11.24±0.50 | 11.92±1.15 | 12.25±0.94 |
| LLama-13B | LLama-7B | 16.45±1.80 | 15.87±1.62 | 15.56±1.44 | 14.88±1.89 | 14.68±1.88 | 14.27±1.70 |

**Table 4:** RQ1 – Comparison of random intervention by different Teacher Models on different Student Models at different intervention budgets on GSM8k.

| Student | Teacher | Intervention Budget | | | | | |
|---|---|---|---|---|---|---|---|
| | | 0% | 20% | 40% | 60% | 80% | 100% |
| Flan-T5-Large | LLama-7B | 58.51±2.00 | 61.71±1.40 | 61.57±3.88 | 59.77±1.53 | 61.28±1.96 | 62.88±1.53 |
| LLama-7B | Flan-T5-Large | 63.44±3.38 | 61.43±2.46 | 62.00±3.70 | 60.26±1.85 | 62.00±1.85 | 60.69±1.23 |

**Table 5:** RQ1 – Comparison of random intervention on StrategyQA where the student and the teacher of different model families.

| Intervention Function | Intervention Budget | | | | | |
|---|---|---|---|---|---|---|
| | 0% | 20% | 40% | 60% | 80% | 100% |
| Random | 58.51±2.00 | 60.40±1.76 | 61.13±2.65 | 60.98±1.09 | 64.33±4.54 | 62.96±2.47 |
| Teacher Conf ↑ | 58.51±2.00 | 58.66±2.40 | 60.11±2.90 | 57.35±3.30 | 61.42±3.91 | 62.96±2.47 |
| Expected Student Conf (Pre) ↓ | 58.51±2.00 | 64.19±2.00 | 66.66±0.25 | 66.81±1.57 | 65.35±2.40 | 62.96±2.47 |
| Expected Student Conf (Post) ↑ | 58.51±2.00 | 64.77±1.76 | 68.26±0.66 | 69.71±2.01 | 68.26±2.63 | 62.96±2.47 |
| Expected Utility ↑ | 58.51±2.00 | 67.83±1.53 | 71.32±1.33 | 71.17±1.15 | 69.86±2.43 | 62.96±2.47 |
| True Student Conf (Pre) ↓ | 58.51±2.00 | 68.26±1.65 | 80.20±1.26 | 74.38±2.84 | 68.55±3.88 | 62.96±2.47 |
| True Student Conf (Post) ↑ | 58.51±2.00 | 65.64±1.40 | 72.63±1.09 | 80.05±0.90 | 72.19±4.39 | 62.96±2.47 |
| True Utility ↑ | 58.51±2.00 | 76.56±0.50 | 80.78±1.15 | 81.51±1.76 | 78.60±3.29 | 62.96±2.47 |

**Table 6:** RQ2 – Comparison of different Intervention Functions with a smaller student (Flan-T5-Large) and a larger teacher (Flan-T5-XL) on StrategyQA. The teacher assumes access to gold labels. ↑ denotes that the samples are ranked in decreasing order of the function (higher is better), while ↓ denotes that the samples in increasing order of the function (lower is better).

| | Intervention Budget | | | | | |
|---|---|---|---|---|---|---|
| | 0% | 20% | 40% | 60% | 80% | 100% |
| Random | 68.12±2.62 | 67.68±2.72 | 65.64±3.39 | 64.04±3.63 | 62.88±1.15 | 61.86±0.66 |
| Expected Student Conf (Pre) ↓ | 68.12±2.62 | 66.22±2.24 | 66.95±1.53 | 65.35±1.00 | 62.73±0.66 | 61.86±0.66 |
| Expected Student Conf (Post) ↑ | 68.12±2.62 | 70.59±3.27 | 71.76±3.63 | 72.48±2.86 | 69.86±2.62 | 61.86±0.66 |
| Expected Utility ↑ | 68.12±2.62 | 70.88±3.27 | 71.90±2.84 | 72.63±2.24 | 68.99±1.15 | 61.86±0.66 |
| True Student Conf (Pre) ↓ | 68.12±2.62 | 74.23±3.73 | 76.27±1.40 | 68.55±1.00 | 64.04±0.90 | 61.86±0.66 |
| True Student Conf (Post) ↓ | 68.12±2.62 | 70.16±3.27 | 73.94±1.76 | 80.05±1.65 | 71.32±1.09 | 61.86±0.66 |
| True Utility ↑ | 68.12±2.62 | 79.91±2.00 | 80.93±2.06 | 80.64±2.24 | 78.16±2.00 | 61.86±0.66 |

**Table 7:** RQ2 – Comparison of different Intervention Functions with a smaller teacher (Flan-T5-Large) and a larger student (Flan-T5-XL) on StrategyQA. The teacher assumes access to gold labels.

| | Intervention Budget | | | | | |
|---|---|---|---|---|---|---|
| | 0% | 20% | 40% | 60% | 80% | 100% |
| Random | 58.51±2.00 | 60.40±1.76 | 61.13±2.65 | 60.98±1.09 | 64.33±4.54 | 62.96±2.47 |
| Least Conf ↓ | 58.51±2.00 | 61.13±0.75 | 62.44±1.74 | 65.06±1.15 | 63.46±2.97 | 62.96±2.47 |
| Expected Student Conf (Pre) ↓ | 58.51±2.00 | 62.59±1.00 | 61.86±0.90 | 62.29±1.33 | 65.50±3.14 | 62.96±2.47 |
| Expected Student Conf (Post) ↑ | 58.51±2.00 | 61.86±1.96 | 62.88±1.74 | 61.71±3.39 | 60.11±4.62 | 62.96±2.47 |
| Expected Utility ↑ | 58.51±2.00 | 62.29±0.50 | 62.44±1.50 | 62.44±3.88 | 62.95±2.78 | 62.96±2.47 |

**Table 8:** RQ2 – Comparison of different Intervention Functions with a smaller student (Flan-T5-Large) and a larger teacher (Flan-T5-XL) when the teacher does not have access to gold labels.

| | Intervention Budget | | | | | |
|---|---|---|---|---|---|---|
| | 0% | 20% | 40% | 60% | 80% | 100% |
| Random | 63.44±3.38 | 64.00±4.05 | 63.63±3.15 | 64.31±4.10 | 64.75±4.59 | 65.68±4.11 |
| Expected Utility ↑ | 64.48±2.06 | 67.24±2.62 | 68.85±3.27 | 69.14±2.40 | 69.72±3.30 | 65.68±4.11 |
| True Student Conf (Pre) ↓ | 64.48±2.06 | 69.86±3.41 | 74.23±3.80 | 70.16±3.96 | 67.39±4.15 | 65.68±4.11 |
| True Student Conf (Post) ↑ | 64.48±2.06 | 65.79±2.06 | 68.85±2.19 | 73.21±3.30 | 70.45±4.48 | 65.68±4.11 |
| True Utility ↑ | 64.48±2.06 | 74.09±3.93 | 75.54±4.72 | 74.96±4.24 | 73.07±4.19 | 65.68±4.11 |

**Table 9:** RQ2 – Comparison of different Intervention Functions with a smaller student (LLama-7B) and a larger teacher (LLama-13B) when the teacher has access to gold labels.

| | Intervention Budget | | | | | |
|---|---|---|---|---|---|---|
| | 0% | 20% | 40% | 60% | 80% | 100% |
| Random | 84.79±0.41 | 85.79±0.48 | 86.79±0.84 | 87.46±0.20 | 88.52±0.39 | 89.72±0.68 |
| Expected Student Conf (Pre) ↓ | 84.79±0.41 | 84.57±0.69 | 86.35±0.73 | 87.99±0.87 | 89.51±0.82 | 89.72±0.68 |
| Expected Student Conf (Post) ↑ | 84.79±0.41 | 86.66±0.37 | 88.69±0.19 | 90.76±0.06 | 92.43±0.61 | 89.72±0.68 |
| Expected Utility ↑ | 84.79±0.41 | 87.34±1.09 | 89.33±0.55 | 90.27±0.40 | 91.30±0.22 | 89.72±0.68 |
| True Student Conf (Pre) ↓ | 84.79±0.41 | 92.03±0.19 | 91.70±0.04 | 91.03±0.34 | 90.27±0.41 | 89.72±0.68 |
| True Student Conf (Post) ↓ | 84.79±0.41 | 87.40±0.39 | 89.59±0.53 | 92.31±0.09 | 94.98±1.57 | 89.72±0.68 |
| True Utility ↑ | 84.79±0.41 | 92.87±0.18 | 93.99±0.02 | 94.65±0.13 | 95.57±0.24 | 89.72±0.68 |

**Table 10:** RQ2 – Comparison of different Intervention Functions with a smaller student (Flan-T5-Large) and a larger teacher (Flan-T5-XL) on CommonsenseQA. The teacher has access to gold labels.

| | Intervention Budget | | | | | |
|---|---|---|---|---|---|---|
| | 0% | 20% | 40% | 60% | 80% | 100% |
| Random | 9.62±1.53 | 10.20±1.06 | 10.68±0.82 | 11.24±0.50 | 11.92±1.15 | 12.25±0.94 |
| Expected Student Conf (Pre) ↓ | 9.62±1.53 | 11.11±1.44 | 11.37±1.17 | 11.56±1.34 | 12.40±1.01 | 12.25±0.94 |
| Expected Student Conf (Post) ↑ | 9.62±1.53 | 12.80±1.28 | 12.91±0.58 | 13.10±0.10 | 12.72±2.14 | 12.25±0.94 |
| Expected Utility ↑ | 9.62±1.53 | 13.68±1.87 | 14.06±1.44 | 13.99±0.80 | 13.68±0.58 | 12.25±0.94 |

**Table 11:** RQ2 – Comparison of different Intervention Functions with a smaller student (LLama-7B) and a larger teacher (LLama-13B) on GSM8k. The teacher has access to gold labels.

| | Intervention Budget | | | | | |
|---|---|---|---|---|---|---|
| | 0% | 20% | 40% | 60% | 80% | 100% |
| Unpersonalized-Rationales | 58.51±2.00 | 66.52±2.97 | 69.14±1.76 | 70.16±1.09 | 67.97±0.50 | 60.40±0.50 |
| Unpersonalized-CoT | 58.51±2.00 | 67.83±1.53 | 71.32±1.33 | 71.17±1.15 | 69.86±2.43 | 62.96±2.47 |
| Theory of Mind | 58.51±2.00 | 69.28±1.26 | 71.61±1.15 | 72.63±1.33 | 68.55±1.90 | 62.73±2.80 |
| Human Explanations | 58.51±2.00 | 72.34±0.90 | 77.72±0.75 | 81.51±1.09 | 82.09±0.87 | 81.36±0.66 |

**Table 12:** RQ3 – Comparison of Theory of Mind motivated teacher explanations with unpersonalized explanations on the student accuracy for StrategyQA with Flan-T5-Large as the student model and Flan-T5-XL as the teacher model.

# 4 RQ2: Additional Results

**Results with Flan Models.** This section provides detailed accuracy tables for RQ2. Table 6 compares different Intervention Functions on StrategyQA with Flan-T5-Large as the student and Flan-T5-Large as the teacher. Table 7 compares the same with a smaller teacher (Flan-T5-Large) and a larger student (Flan-T5-XL). In Table 8, we compare the accuracy on StrategyQA when the teacher (Flan-T5-XL) does not have access to gold labels.

**Results with Other Models and Different Datasets.** Table 9 compares Intervention Functions with LLama models (LLama-7B as the student and LLama-13B as the teacher) on StrategyQA. Table 10 compares different Intervention Functions on the CommonsenseQA dataset with Flan-T5-Large as the student and Flan-T5-XL as the teacher. Table 11 reports results on the GSM8k dataset with LLama-7B as the student and LLama-13B as the teacher.

# 5 RQ3: Additional Results

Table 12 shows RQ3 results on StrategyQA with Flan-T5-Large as the student and Flan-T5-XL as the teacher.

# Broader Impacts

Chain-of-Thought rationales have empowered almost all recent developments in complex reasoning tasks. We hope that our findings can help improve the understanding and evaluation of these rationales, as a way to also understand the behavior of LLMs and make them more interpretable. We do not foresee specific ethical risks arising from this work that do not already apply to the general use of Large Language Models, such as the potential to generate harmful or toxic content [5].

# Limitations

While teacher LLMs generate better explanations by building a Theory of Mind, the human explanations are unpersonalized i.e., collected without any particular student in mind. In spite of that, we observe that intervention with human explanations proves to be helpful in most cases. It remains to be seen whether human explanations that are directed toward improving a particular student model can lead to further improvements. Next, we make a simplifying assumption that the communication cost is uniform across all samples. Non-uniform costs (e.g., measured based on the number of tokens or reasoning steps) such that longer explanations incur larger costs is an interesting direction for future work. We also note that while both student and teacher generate explanations with the goal of improving student predictions, the predictions may still be unfaithful to the reasoning steps.

# References


[1] Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021.

[2] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[3] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 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, 2019.

[4] Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. Explanations for commonsenseqa: New dataset and models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the*

*11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3050–3065, 2021.

[5] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.