Distilling Step-by-Step!
Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes

ACL 2023

Peifeng Wang*[1] , Zhengyang Wang[2], Zheng Li[2],
Yifan Gao[2], Bing Yin[2], Xiang Ren[1]

[1]Department of Computer Science, University of Southern California

[2]Amazon.com Inc

Part 1.

Background

- **Large language models (LLMs)**
  - Models are challenging to deploy in real world applications
    - Computational requirements are far beyond affordable for most product teams, especially for applications that require low latency performance
  - Deploy smaller specialized models
    - Finetuning updates a pretrained smaller model using downstream human annotated data
    - Distillation trains the same smaller models with labels generated by a larger LLM
  - Limitation
    - Finetuning requires expensive human labels
    - Distillation requires large amounts of unlabeled data which can be hard to obtain

Jason Wei et al. Chain of thought prompting elicits reasoning in large language models. 2022.

Part 2.

Introduction

- **Distilling step-by step**
  - Reduce the amount of training data required for both finetuning and distillation
  - Perspective as a source of noisy labels to viewing LLMs as agents that can reason
    - LLMs can produce natural language rationales justifying their predicted labels
    - Utilize extracted rationales as richer information with both label prediction and rationale prediction tasks
  - Expriemental result
    - Achieve better performance with over 50% less training examples on average across datasets

Question

Jesse's room is 11 feet long and 15 feet wide. If she already has 16 square feet of carpet. How much more carpet does she need to cover the whole floor?

Intermediate rationales

Area = length × width.
Jesse's room has 11 × 15 square feet.

Relevant task knowledge

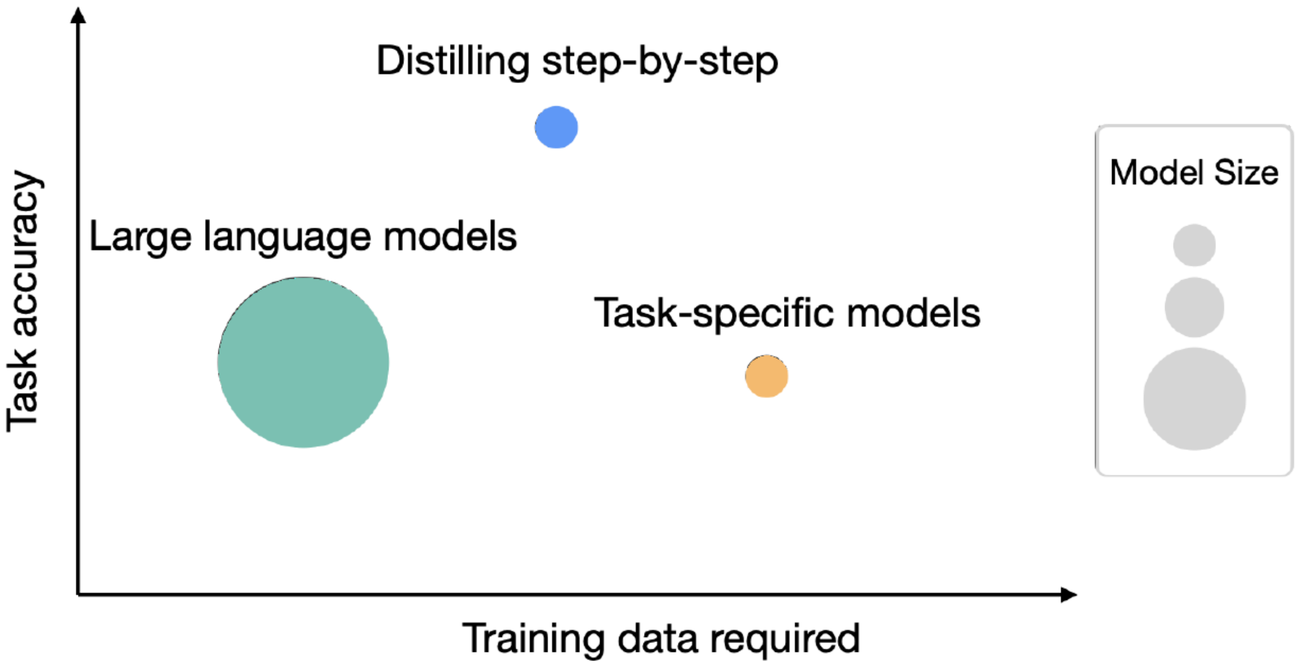
$Area = length \times width$

Final answer

$(11 \times 15) - 16$

Introduction

- **Distilling step-by step**
  - Expriemental result
    - Surpass the performance of 540B parameter LLMs using a 770M T5 model
    - This smaller model only uses 80% of a labeled dataset
    - Outperform 540B PaLM's performance with only a 11B T5 model



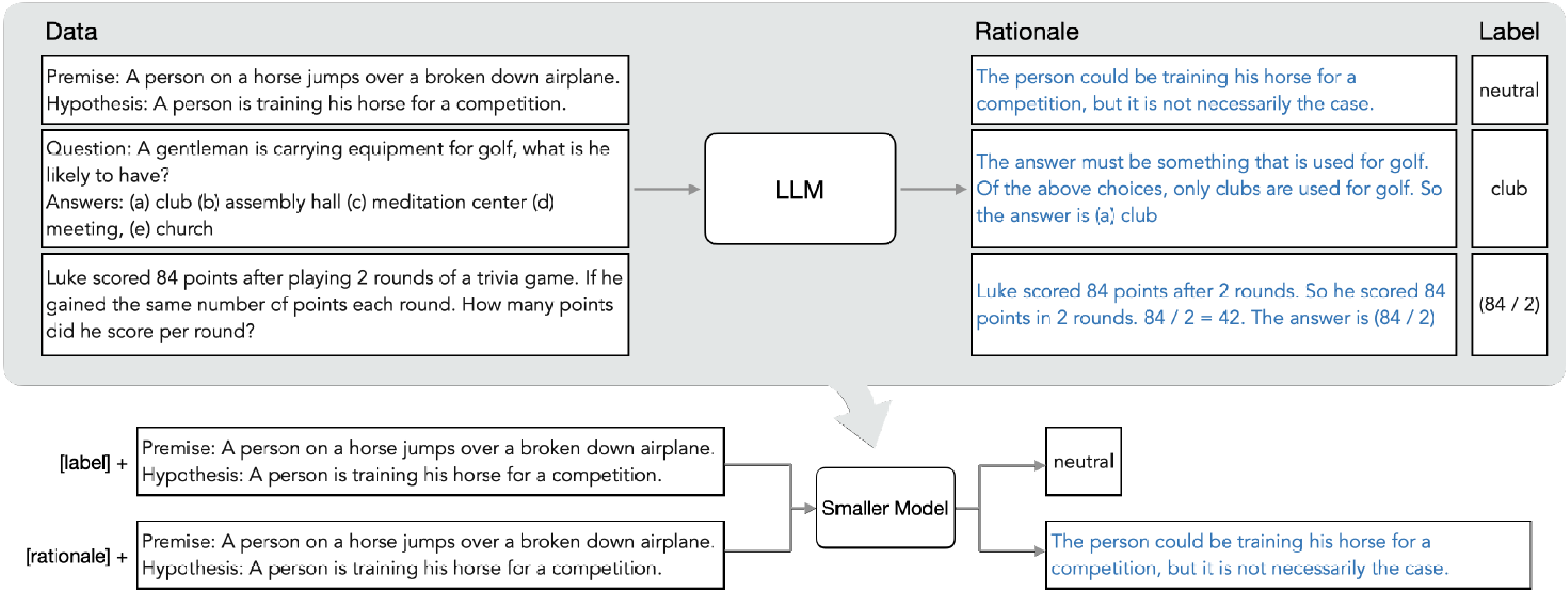
Model Architecture

- 500× less model parameters (up to 2000× smaller)
- 50% less training example (up to over 85% reduction)

Part 2.

Introduction

- **Overview**
  - Train small task-specific models within a multi-task learning framework



Data

Premise: A person on a horse jumps over a broken down airplane.
Hypothesis: A person is training his horse for a competition.

Question: A gentleman is carrying equipment for golf, what is he likely to have?
Answers: (a) club (b) assembly hall (c) meditation center (d) meeting, (e) church

Luke scored 84 points after playing 2 rounds of a trivia game. If he gained the same number of points each round. How many points did he score per round?

LLM

Rationale	Label
The person could be training his horse for a competition, but it is not necessarily the case.	neutral
The answer must be something that is used for golf. Of the above choices, only clubs are used for golf. So the answer is (a) club	club
Luke scored 84 points after 2 rounds. So he scored 84 points in 2 rounds. 84 / 2 = 42. The answer is (84 / 2)	(84 / 2)

[label] + Premise: A person on a horse jumps over a broken down airplane.
Hypothesis: A person is training his horse for a competition.

[rationale] + Premise: A person on a horse jumps over a broken down airplane.
Hypothesis: A person is training his horse for a competition.

Smaller Model

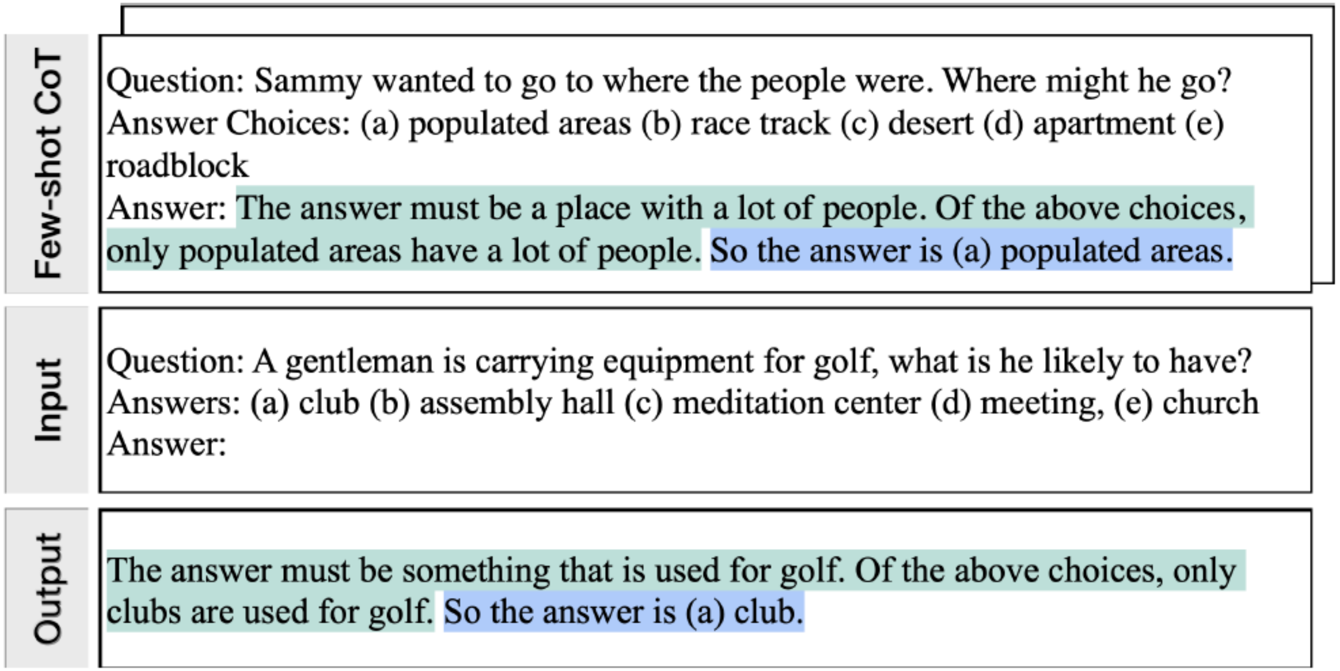
neutral

The person could be training his horse for a competition, but it is not necessarily the case.

Part 3.

Method

- **Extract rationales from LLMs**
  - Chain-of-thought (CoT) prompting (Wei et al., 2022) to elicit & extract rationales
  - Unlabeled dataset $x_i \in D$ Prompt template p
  - Append each input x_i to p to generate rationales and labels


Few-shot CoT
Input
Output

Question: Sammy wanted to go to where the people were. Where might he go?
Answer Choices: (a) populated areas (b) race track (c) desert (d) apartment (e) roadblock
Answer: The answer must be a place with a lot of people. Of the above choices, only populated areas have a lot of people. So the answer is (a) populated areas.

Question: A gentleman is carrying equipment for golf, what is he likely to have?
Answers: (a) club (b) assembly hall (c) meditation center (d) meeting, (e) church
Answer:

The answer must be something that is used for golf. Of the above choices, only clubs are used for golf. So the answer is (a) club.

Highlight: Example rationale (green) / Label (blue)

- Triplet $(x^{\mathrm{p}}, r^{\mathrm{p}}, y^{\mathrm{p}})$
- Example input x^{p}
- Its corresponding label y^{p}
- User-provided rationale r^{p}
  - Why x^{p} can be categorized as y^{p}
- Generate the rationale $\hat{r}_i$
- Generate the output $\hat{y}_i$

Part 3.

Method

- **Train smaller models with rationales**
  - Extend it to incorporate rationales into the training process
  - Label prediction loss
    - Learn to Generates pseudo noisy training labels $\hat{y}_i$
  - Rationale generation loss
    - Learn to generate the intermediate reasoning steps for the prediction

$$\boxed{\begin{aligned}
&\mathcal{L} = \mathcal{L}_{\text{label}} + \lambda \mathcal{L}_{\text{rationale}} \\
&\mathcal{L}_{\text{label}} = \frac{1}{N} \sum_{i=1}^{N} \ell(f(x_i), \hat{y}_i) \\
&\mathcal{L}_{\text{rationale}} = \frac{1}{N} \sum_{i=1}^{N} \ell(f(x_i), \hat{r}_i)
\end{aligned}}$$

Method

- **Standard finetuning and task distillation**
  - Common practice
    - Train a task-specific model is to finetune a pretrained model with supervised data $\hat{y}_i$
    - Task-specific distillation uses teachers to generates pseudo noisy training labels
  - Labels
    - Human-annotated labels for the standard finetuning case y_i
    - LLM-predicted labels for the model distillation case $\hat{y}_i$

- Smaller Model f
- Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$
- Cross-entropy loss ℓ
- Label prediction loss $\mathcal{L}_{\text{label}} = \frac{1}{N}\sum_{i=1}^{N} \ell(f(x_i), \hat{y}_i)$

Method

- **Multi-task learning with rationales**
  - To create a more explicit connection between x_i to $\hat{y}_i$
  - Use extracted rationales $\hat{r}_i$ as additional supervision
  - Learn to generate the intermediate reasoning steps for the prediction
  - Previous design
    - LLM is still necessary during deployment, limited its deployability
  - Distilling step-by-step
    - Rationale $\hat{r}_i$ is not required in the test time
    - Removes the need for an LLM at test-time

- Cross-entropy loss ℓ
- Previous approach $$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \ell(f(x_i, \hat{r}_i), \hat{y}_i)$$
- Rationale generation loss $$\mathcal{L}_{\text{rationale}} = \frac{1}{N}\sum_{i=1}^{N} \ell(f(x_i), \hat{r}_i)$$

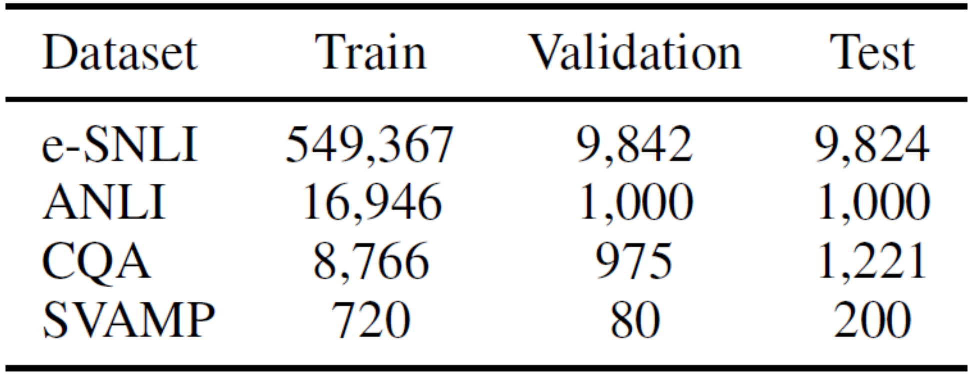
Part 4.

Experiment

Experimental Setup

- LLM(Teacher): 540B PaLM (Chowdhery et al., 2022)
- Task-specific models(Student): T5-Base (220M), T5-Large (770M), T5-XXL (11B)
  - Initialize the models with pretrained weights obtained from publicly available sources
- Dataset
  - Natural language inference : e-SNLI (Camburu et al., 2018), ANLI (Nie et al., 2020)
  - Commonsense question answering: CQA (Talmor et al., 2019; Rajani et al., 2019)
  - Arithmetic math word problems: SVAMP (Patel et al., 2021)

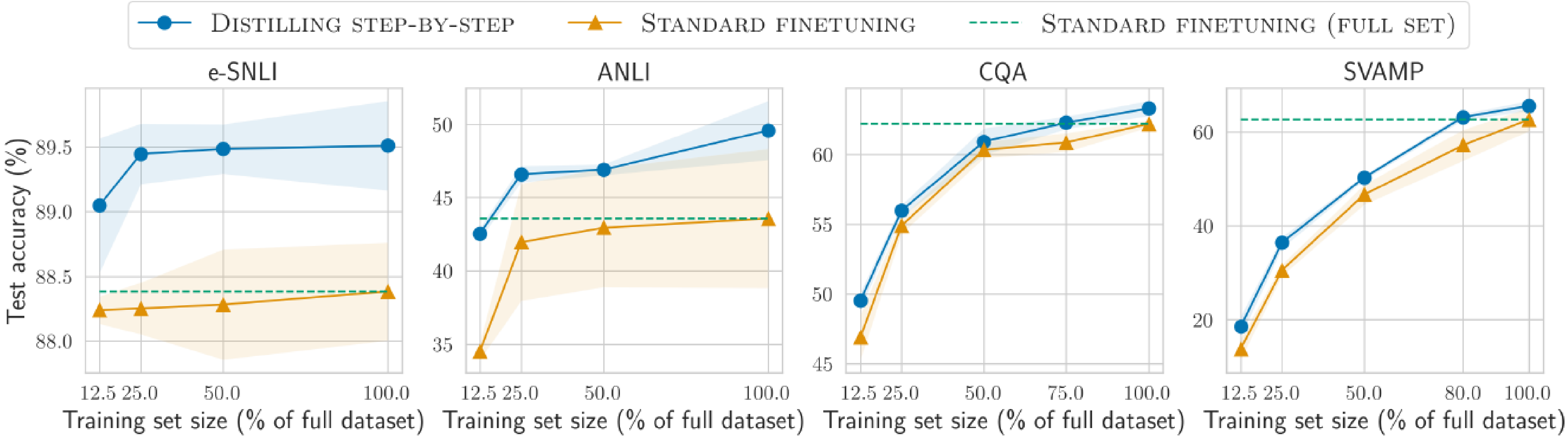
Dataset statistics



Dataset	Train	Validation	Test
e-SNLI	549,367	9,842	9,824
ANLI	16,946	1,000	1,000
CQA	8,766	975	1,221
SVAMP	720	80	200

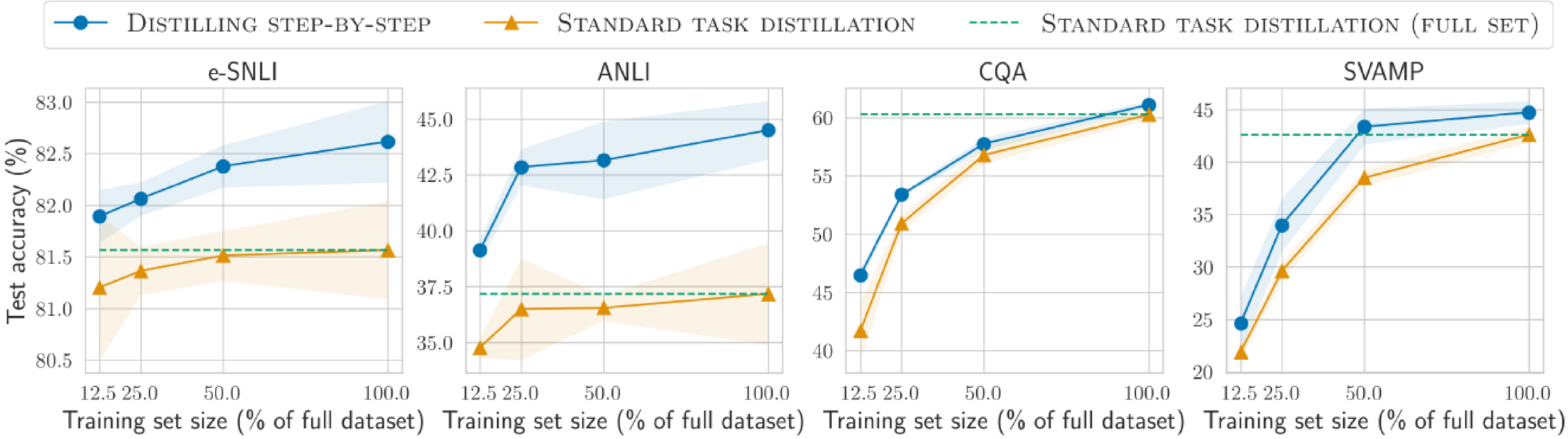
Experiment

- **Finetuning on varying sizes of human-labeled datasets**
  - Distilling step-by-step is able to outperform Standard finetuning
    - much less training examples (e.g., 12.5% of the full e-SNLI dataset)



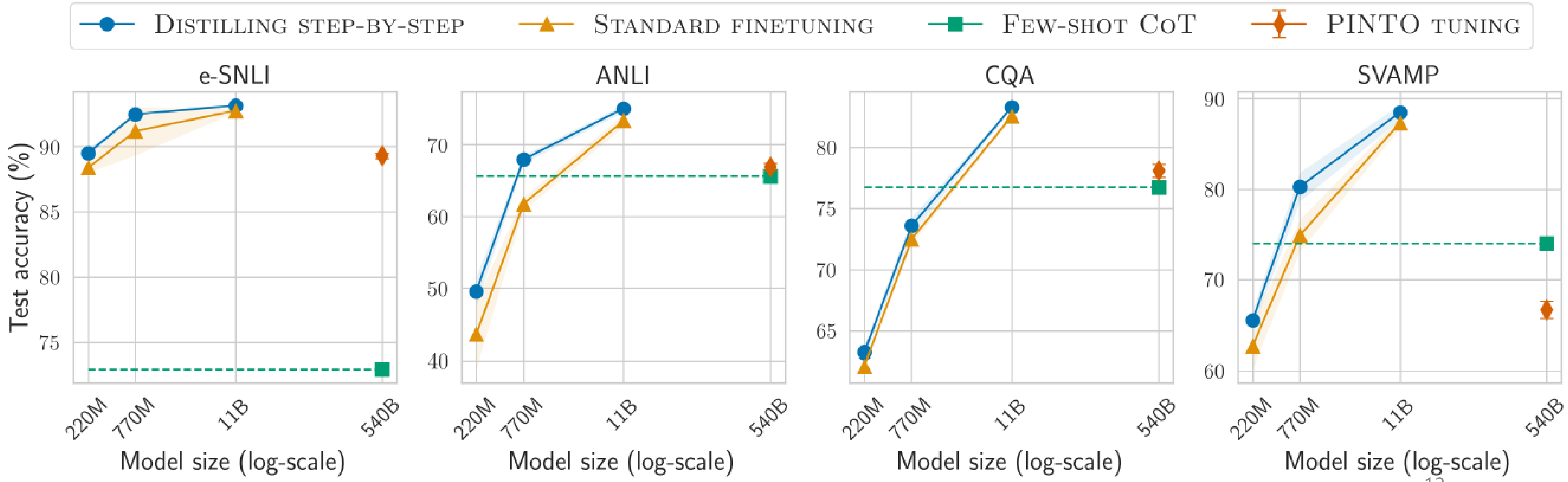
Experiment

- **Distillation on varying sizes of human-labeled datasets**
  - Distilling step-by-step is able to outperform Standard task distillation
    - much less training examples (e.g., 12.5% of the full ANLI dataset)



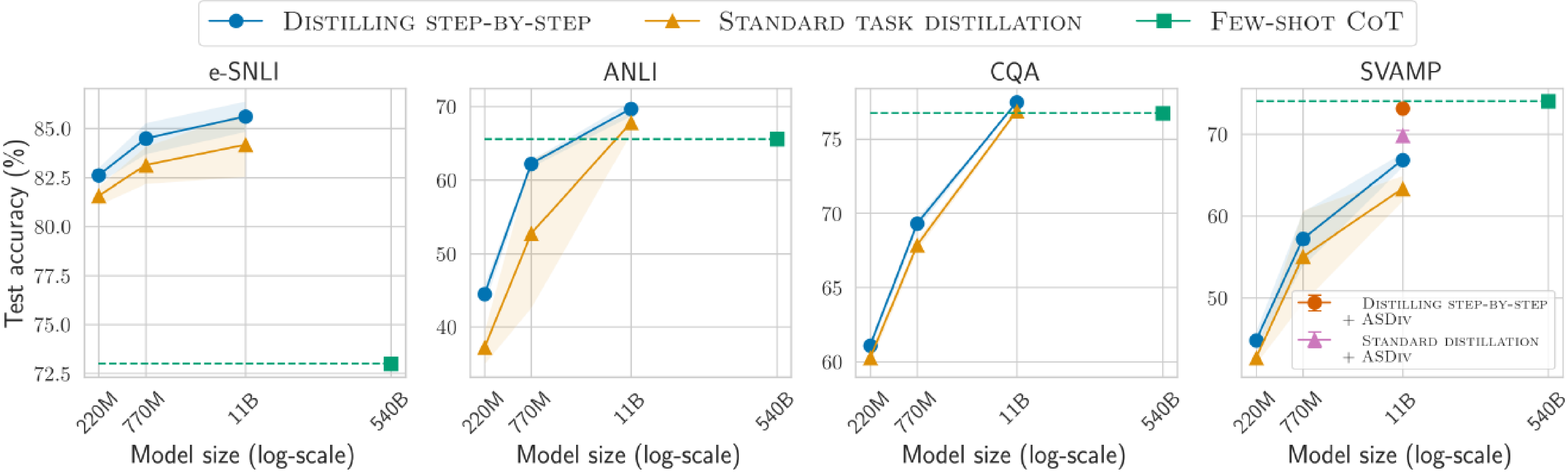
Experiment

- **Task-specific model sizes & Full human-labeled datasets**
  - Distilling step-by-step is able to outperform task-specific model
    - Always outperform Few-shot CoT and PINTO tuning on all 4 datasets considered considered, by using much smaller T5 models



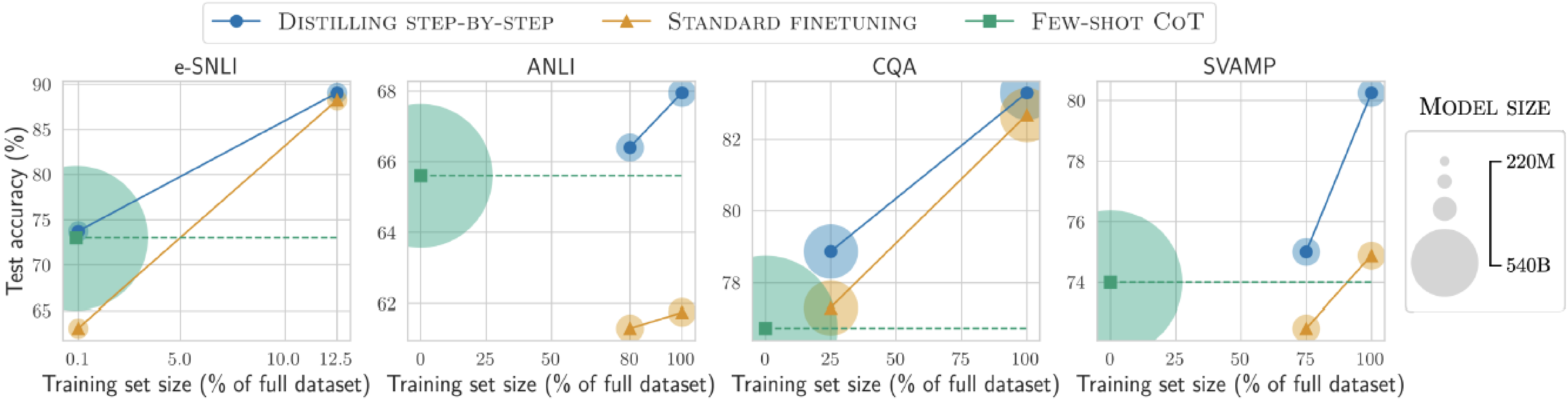
Experiment

- **Task-specific model sizes & Unlabeled datasets**
  - Distilling step-by-step is able to outperform task-specific model
  - Unlabeled data augmentation
    - Distilling step-by-step is able to much more efficiently exploit the value of the added examples



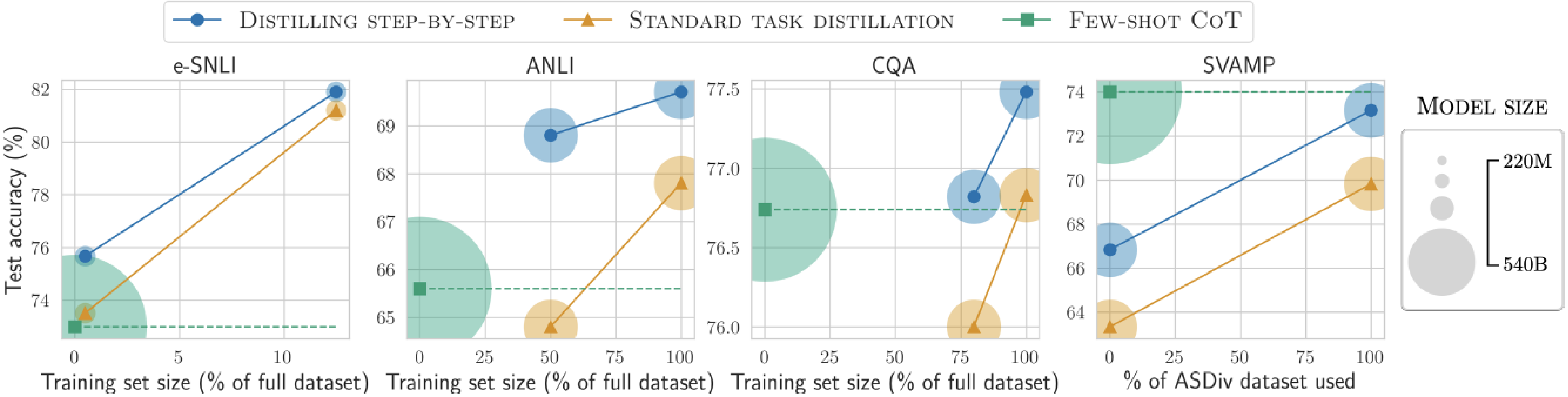
Experiment

- **Minimum task-specific model & Human-labeled datasets**
  - Outperform Few-shot CoT with smaller model as well as less data
  - Standard finetuning requires more data & larger model to match the performance



Experiment

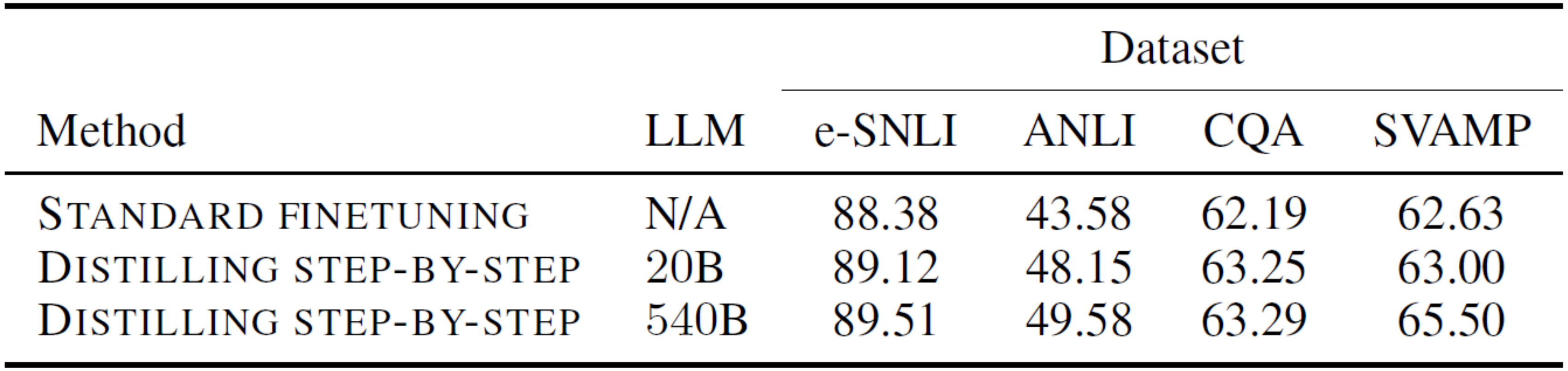
- **Minimum task-specific model & Unlabeled datasets**
  - Outperform Few-shot CoT with smaller model as well as less data
  - Standard finetuning requires more data & larger model to match the performance



Part 4.

Experiment

- **Ablation studies**
  - Distilling step-by-step works with different sizes of LLMs
    - Teacher: 20B GPT-NeoX / Student: 220M T5 models



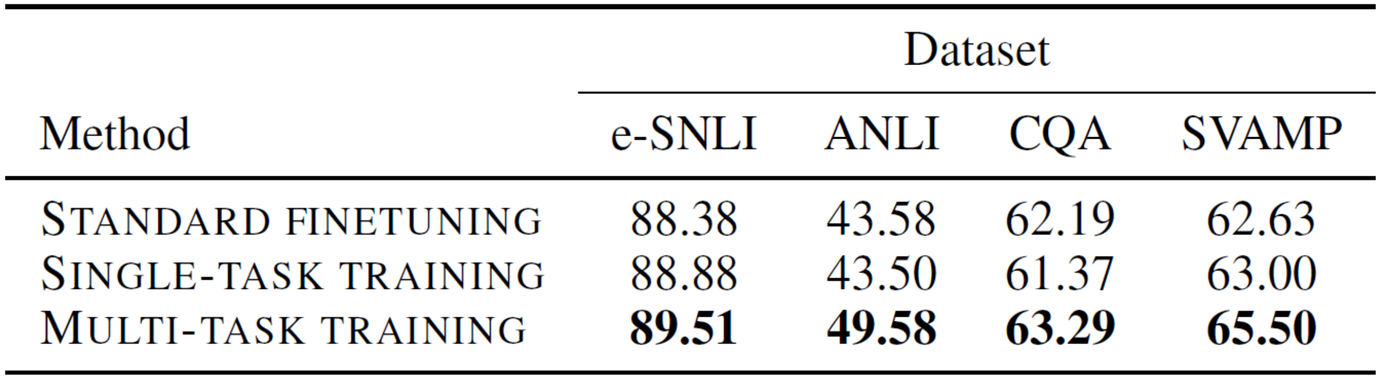
		Dataset			
Method	LLM	e-SNLI	ANLI	CQA	SVAMP
STANDARD FINETUNING	N/A	88.38	43.58	62.19	62.63
DISTILLING STEP-BY-STEP	20B	89.12	48.15	63.25	63.00
DISTILLING STEP-BY-STEP	540B	89.51	49.58	63.29	65.50

J R Landis, G G Koch. The measurement of observer agreement for categorical data. Biometrics. 1977.

Experiment

- **Ablation studies**
  - Teacher: 20B GPT-NeoX / Student: 220M T5 model
  - Single-task training
    - Simply treating rationale and label predictions as a single joint task may harm the model's performance on label prediction

Single-task training VS Multi-task training



	Dataset			
Method	e-SNLI	ANLI	CQA	SVAMP
STANDARD FINETUNING	88.38	43.58	62.19	62.63
SINGLE-TASK TRAINING	88.88	43.50	61.37	63.00
MULTI-TASK TRAINING	**89.51**	**49.58**	**63.29**	**65.50**

Single-task training objective

$$\mathcal{L}_{\text{single}} = \frac{1}{N}\sum_{i=1}^{N} \ell(f(x_i), [\hat{r}_i, \hat{y}_i])$$

- Rationale $\hat{r}_i$
- Label $\hat{y}_i$
- Single sequence $[\hat{r}_i, \hat{y}_i]$

J R Landis, G G Koch. The measurement of observer agreement for categorical data. Biometrics. 1977.

Conclusion

- **Distilling step-by-step**
  - Method
    - Extract rationales from LLMs as informative supervision in training small task-specific models
  - Experimental Result
    - Reduce the training dataset required to curate task-specific smaller models
    - Reduce the model size required to achieve, and even surpass, the original LLM's performance

Part 5.

Conclusion

- **Limitations**
  - Require users to produce a few example demonstrations (~ 10-shot for all tasks) in order to use the few-shot CoT (Wei et al., 2022) prompting mechanism
    - Using recent advances that suggest that rationales can be elicited without any userannotated demonstrations (Kojima et al., 2022)
  - Incur slight training-time computation overhead
    - At test time, our multi-task design avoids the computation overhead since it allows one to only predict labels without generating the rationales
  - Observe success using LLM rationales
    - Llms exhibit limited reasoning capability on more complex reasoning and planning tasks (Valmeekam et al., 2022)

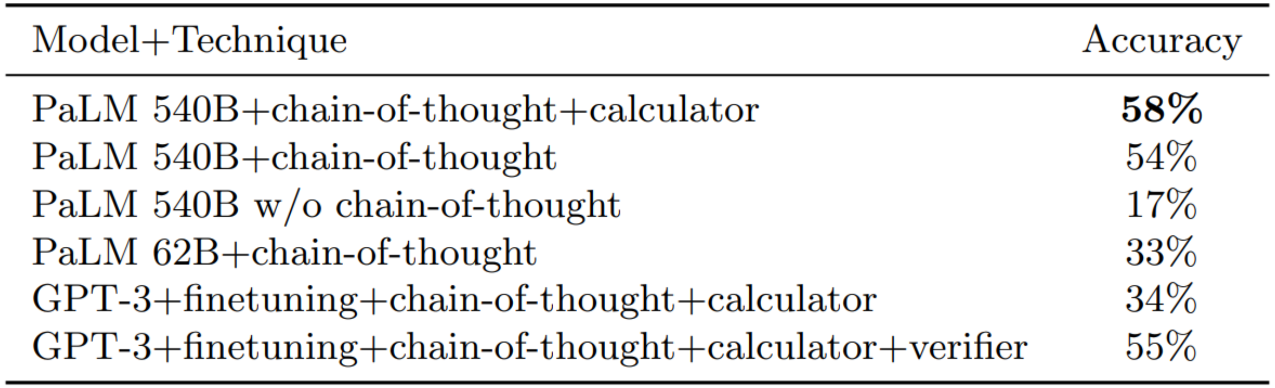
Jason Wei et al. Chain of thought prompting elicits reasoning in large language models. 2022.
Takeshi Kojima et al. Large language models are zero-shot reasoners. NeurIPS. 2022.
Karthik Valmeekam et al. Large language models still can't plan (a benchmark for llms on planning and reasoning about change). NeurIPS Workshop FMDM. 2022.

Part 6.

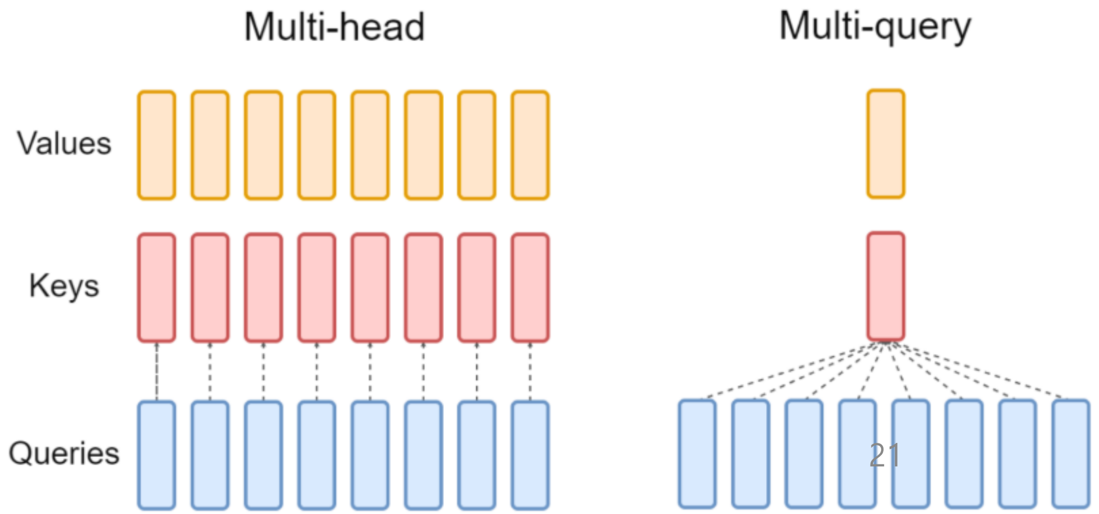
Appendix

- **PaLM (Chowdhery et al., 2022)**
  - SwiGLU Activation
    - Significantly increase quality compared to standard ReLU, GeLU, or Swish activations
  - Parallel Layers
    - Parallel formulation $$y = x + \text{MLP}(\text{LayerNorm}(x)) + \text{Attention}(\text{LayerNorm}(x))$$
    - Results in roughly 15% faster training speed at large scales
    - 8B scale but no quality degradation at 62B scale
  - RoPE embeddings (Su et al., 2021)
  - No biases were used in any of the dense kernels or layer norms

Dataset: GSM8K



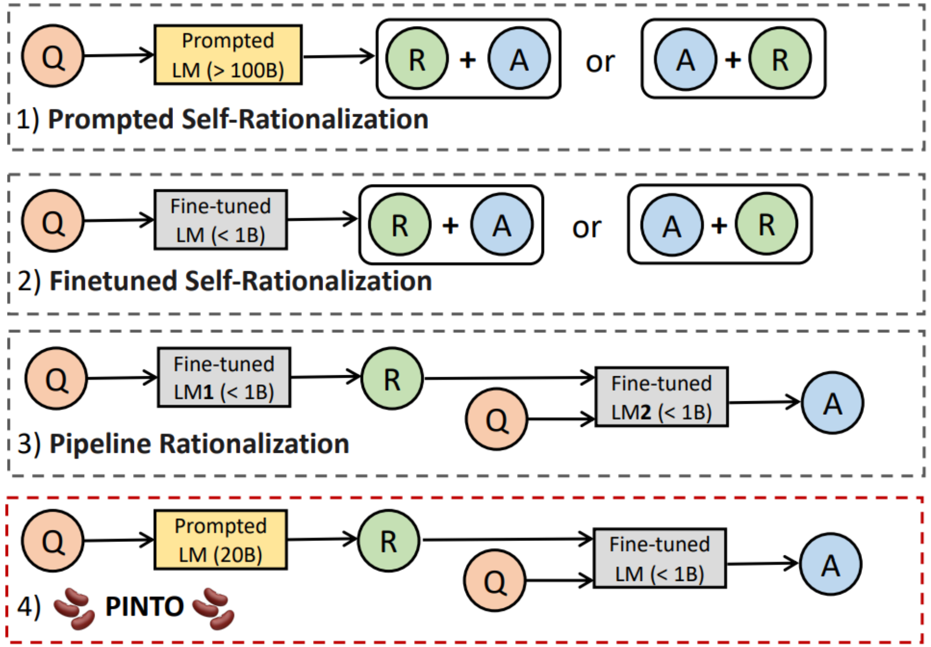
Model+Technique	Accuracy
PaLM 540B+chain-of-thought+calculator	**58%**
PaLM 540B+chain-of-thought	54%
PaLM 540B w/o chain-of-thought	17%
PaLM 62B+chain-of-thought	33%
GPT-3+finetuning+chain-of-thought+calculator	34%
GPT-3+finetuning+chain-of-thought+calculator+verifier	55%

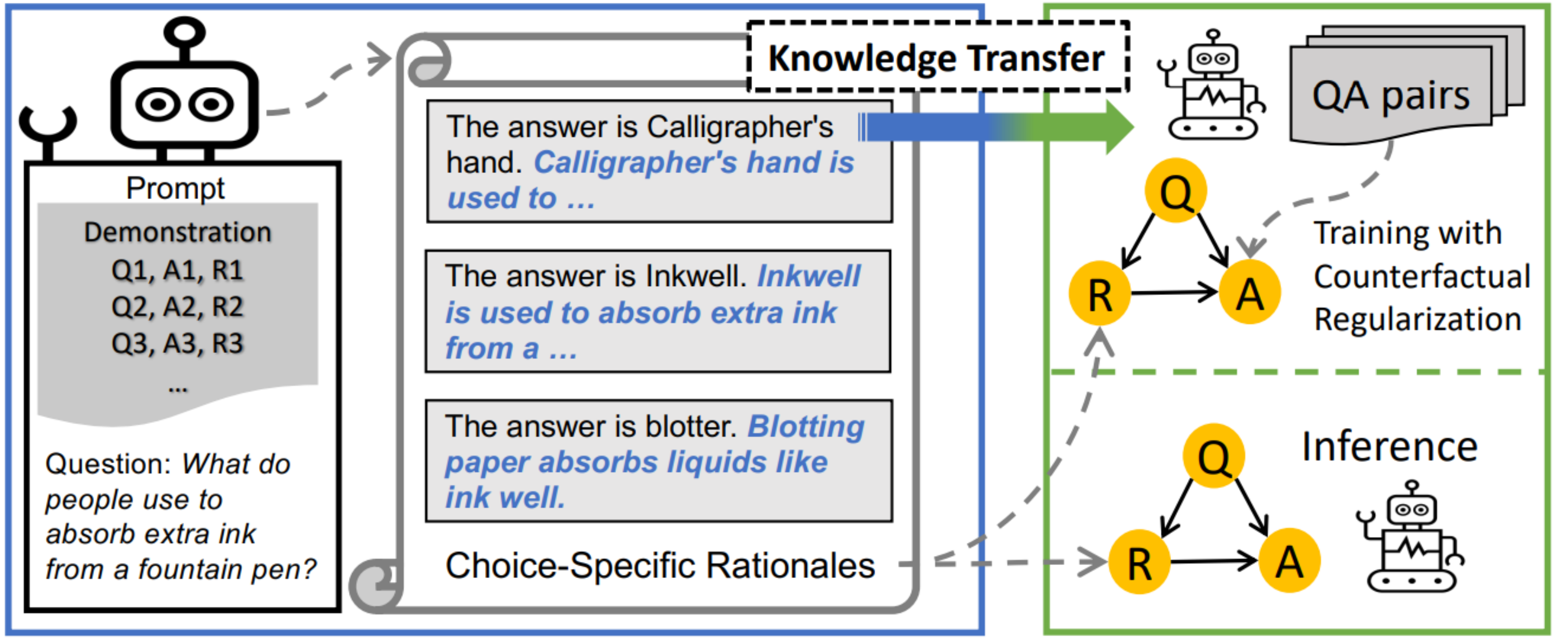


Aakanksha Chowdhery et al. Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311. 2022.

Appendix

- **Pinto TUNING (Wang et al., 2022a)**
  - Prompted RatIonalizing with CouNTerfactual ReasOning (PINTO)
    - Rationalizing module is frozen during fine-tuning, which drastically reduces training costs
    - Knowledge of reasoning module (smaller LM) is transferred from the rationalizing module




Prompt
Demonstration
Q1, A1, R1
Q2, A2, R2
Q3, A3, R3
...
Question: What do people use to absorb extra ink from a fountain pen?
Knowledge Transfer
The answer is Calligrapher's hand. Calligrapher's hand is used to ...
The answer is Inkwell. Inkwell is used to absorb extra ink from a ...
The answer is blotter. Blotting paper absorbs liquids like ink well.
Choice-Specific Rationales
QA pairs
Q
R
A
Training with Counterfactual Regularization
Inference
Q
R
A

a) **Rationalizing Module**:
frozen, prompted *medium-scale* LM

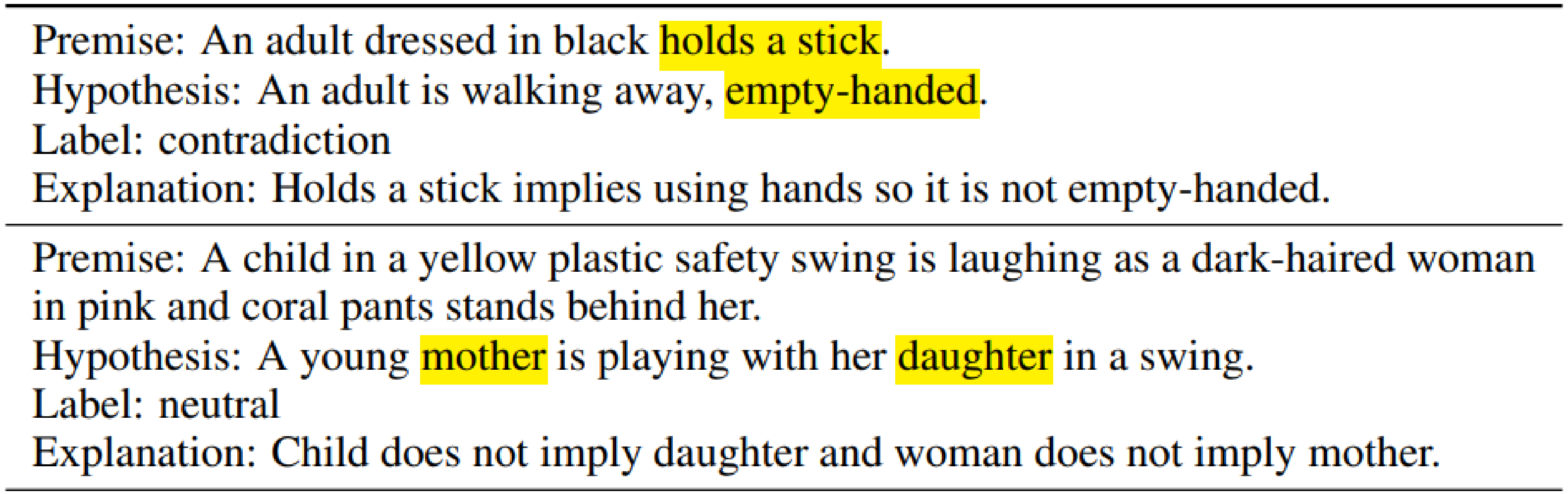
b) **Reasoning Module**:
fine-tuned *small-scale* LM

Peifeng Wang et al. Pinto: Faithful language reasoning using prompt-generated rationales. ICLR. 2023.

Part 6.

Appendix

- **e-SNLI (Camburu et al., 2018)**
  - Exploit and generate explanations for the task of recognizing textual entailment
  - Free-form natural language explanations, as opposed to formal language, for a series of reasons
    - Natural language is readily comprehensible to an end-user who needs to assert a model's reliability
    - Eliminate the additional effort of learning to produce formal language, making it simpler to collect such datasets
    - Natural language justifications might eventually be mined from existing large-scale free-form text



Premise: An adult dressed in black holds a stick.
Hypothesis: An adult is walking away, empty-handed.
Label: contradiction
Explanation: Holds a stick implies using hands so it is not empty-handed.

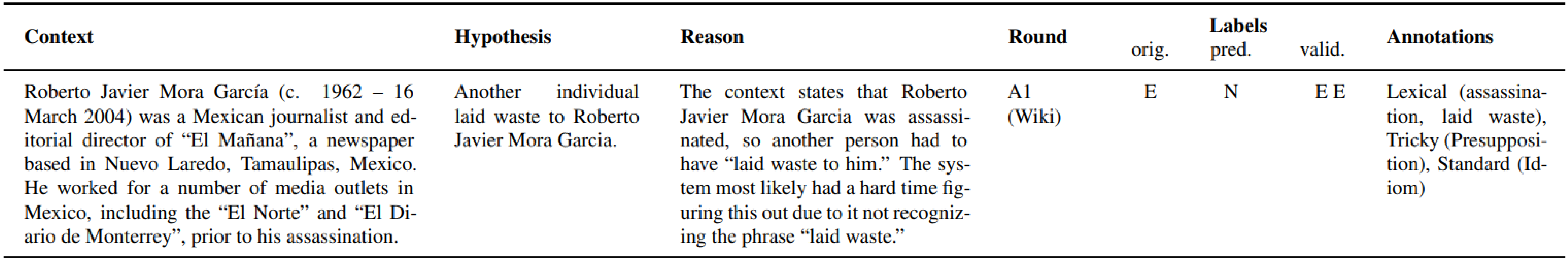
Premise: A child in a yellow plastic safety swing is laughing as a dark-haired woman in pink and coral pants stands behind her.
Hypothesis: A young mother is playing with her daughter in a swing.
Label: neutral
Explanation: Child does not imply daughter and woman does not imply mother.

Oana-Maria Camburu et al. e-snli: Natural language inference with natural language explanations. NeurIPS. 2018.

Part 6.

Appendix

- **ANLI (Nie et al., 2020)**
  - Adversarial NLI
    - Both benchmark longevity and robustness issues
  - Adversarial human-andmodel-in-the-loop solution for NLU dataset collection
    - In the first stage, our current best models cannot determine the correct label for
    - After each round, we train a new model and set aside a new test set



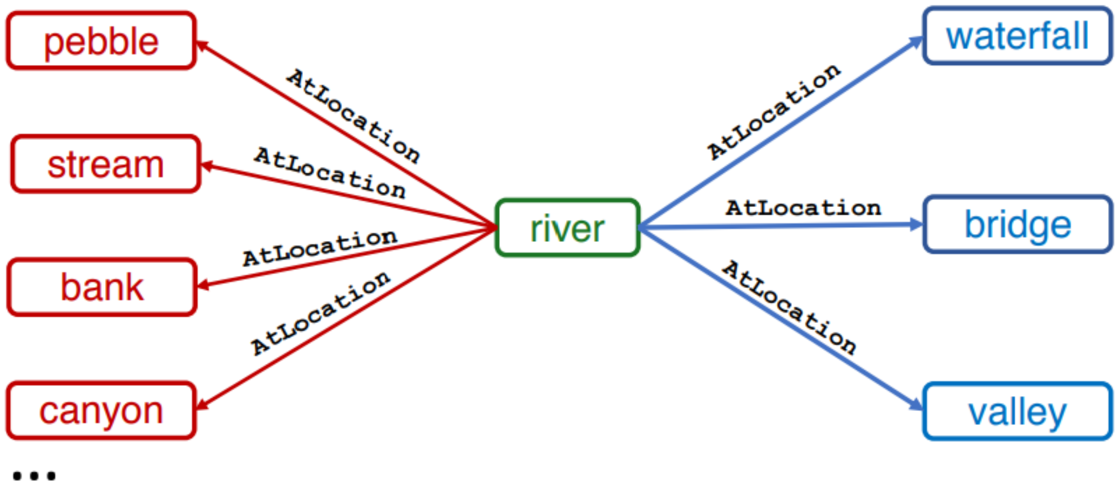
Context	Hypothesis	Reason	Round	Labels			Annotations
				orig.	pred.	valid.	
Roberto Javier Mora García (c. 1962 – 16 March 2004) was a Mexican journalist and editorial director of “El Mañana”, a newspaper based in Nuevo Laredo, Tamaulipas, Mexico. He worked for a number of media outlets in Mexico, including the “El Norte” and “El Diario de Monterrey”, prior to his assassination.	Another individual laid waste to Roberto Javier Mora Garcia.	The context states that Roberto Javier Mora Garcia was assassinated, so another person had to have “laid waste to him.” The system most likely had a hard time figuring this out due to it not recognizing the phrase “laid waste.”	A1 (Wiki)	E	N	E E	Lexical (assassination, laid waste), Tricky (Presupposition), Standard (Idiom)

Yixin Nie et al. Adversarial NLI: A new benchmark for natural language understanding. ACL. 2020.

Appendix

- **CSQA (Talmor et al., 2018)**
  - Five-choice QA dataset that tests general commonsense about the daily concepts
  - Generate commonsense questions at scale by asking crowd workers to author questions that describe the relation between concepts from CONCEPTNET
  - Only that particular target concept is the answer, while the other two distractor concepts are not

a) Sample ConceptNet for specific subgraphs



b) Crowd source corresponding natural language questions and two additional distractors

*Where on a **river** can you hold a cup upright to catch water on a sunny day?*

✓ **waterfall**, ✗ bridge, ✗ valley, ✗ pebble, ✗ mountain

*Where can I stand on a **river** to see water falling without getting wet?*

✗ waterfall, ✓ **bridge**, ✗ valley, ✗ stream, ✗ bottom

*I'm crossing the **river**, my feet are wet but my body is dry, where am I?*

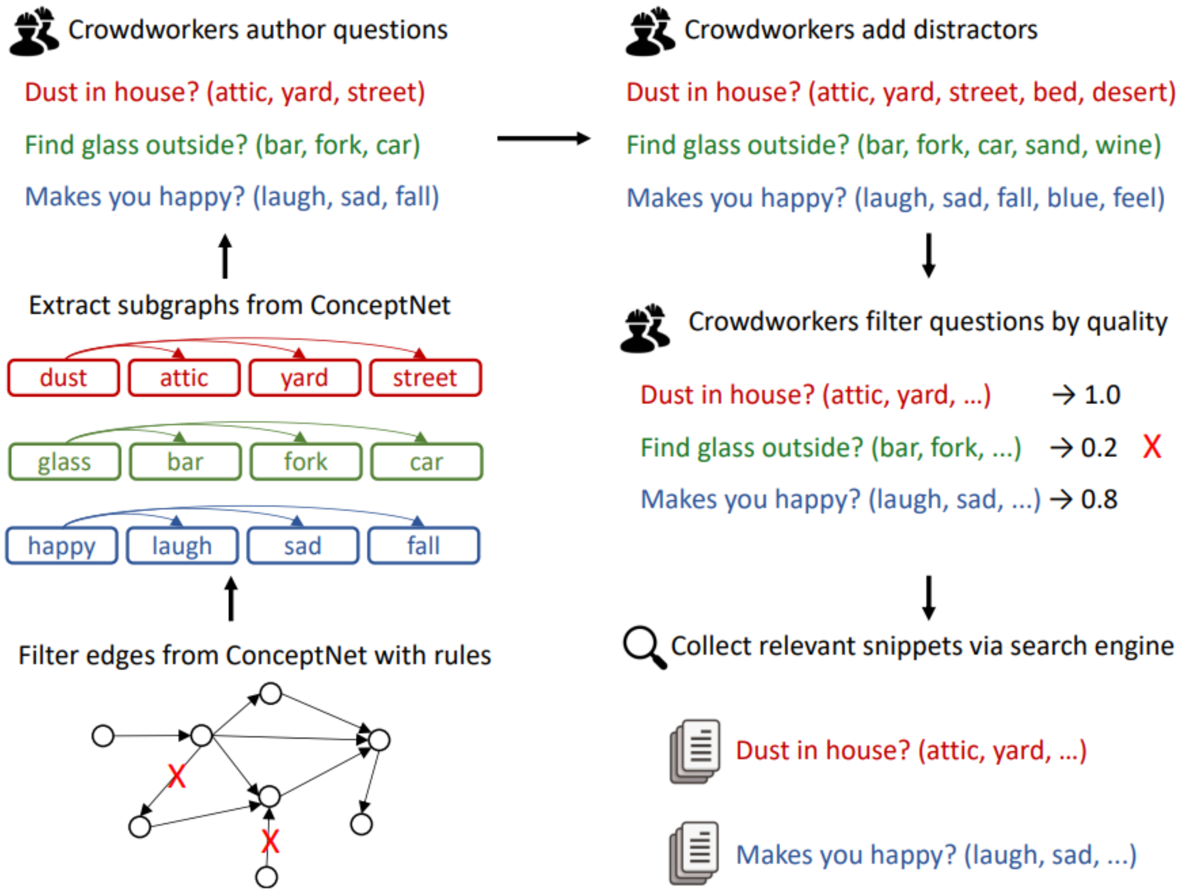
✗ waterfall, ✗ bridge, ✓ **valley**, ✗ bank, ✗ island

Alon Talmor et al. Commonsenseqa: A question answering challenge targeting commonsense knowledge. 2018.

Part 6.

Appendix

- **CSQA (Talmor et al., 2018)**
  - Dataset Generation



Crowdworkers author questions

Dust in house? (attic, yard, street)

Find glass outside? (bar, fork, car)

Makes you happy? (laugh, sad, fall)

Extract subgraphs from ConceptNet

Filter edges from ConceptNet with rules

Crowdworkers add distractors

Dust in house? (attic, yard, street, bed, desert)

Find glass outside? (bar, fork, car, sand, wine)

Makes you happy? (laugh, sad, fall, blue, feel)

Crowdworkers filter questions by quality

Dust in house? (attic, yard, ...) → 1.0

Find glass outside? (bar, fork, ...) → 0.2 X

Makes you happy? (laugh, sad, ...) → 0.8

Collect relevant snippets via search engine

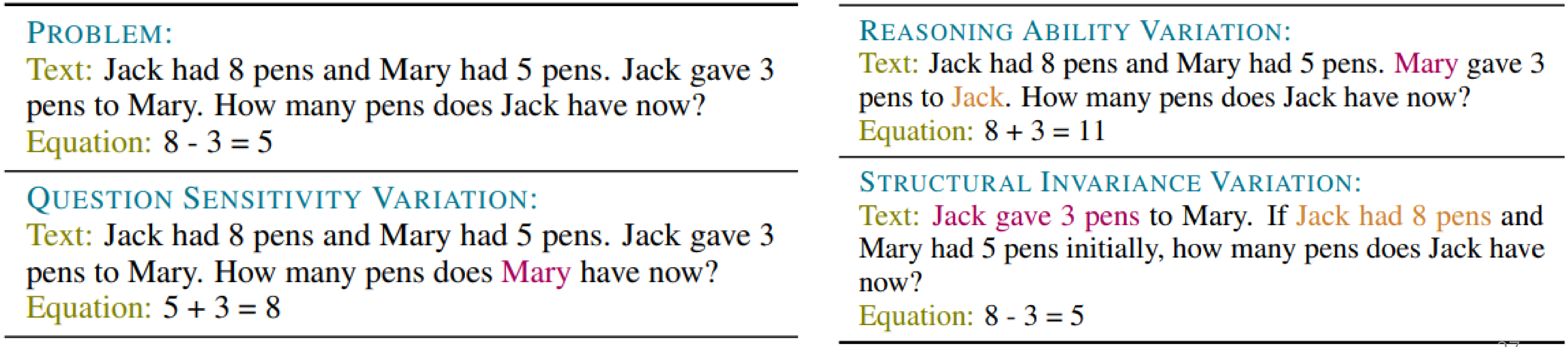
Dust in house? (attic, yard, ...)

Makes you happy? (laugh, sad, ...)

Alon Talmor et al. Commonsenseqa: A question answering challenge targeting commonsense knowledge. 2018.

Appendix

- **SVAMP (Patel et al., 2021)**
  - Existing methods use shallow heuristics to solve a majority of word problems
    - Robustly solve even the simplest of MWPs suggesting that the above belief is not well-founded
  - Provide concrete evidence
    - Existing methods use shallow heuristics to solve a majority of word problems
  - Question text has been removed leaving only the narrative
    - Rely on superficial patterns present in the narrative of the MWP
    - Achieve high accuracy without even looking at the question



PROBLEM:
Text: Jack had 8 pens and Mary had 5 pens. Jack gave 3 pens to Mary. How many pens does Jack have now?
Equation: 8 - 3 = 5

QUESTION SENSITIVITY VARIATION:
Text: Jack had 8 pens and Mary had 5 pens. Jack gave 3 pens to Mary. How many pens does Mary have now?
Equation: 5 + 3 = 8

REASONING ABILITY VARIATION:
Text: Jack had 8 pens and Mary had 5 pens. Mary gave 3 pens to Jack. How many pens does Jack have now?
Equation: 8 + 3 = 11

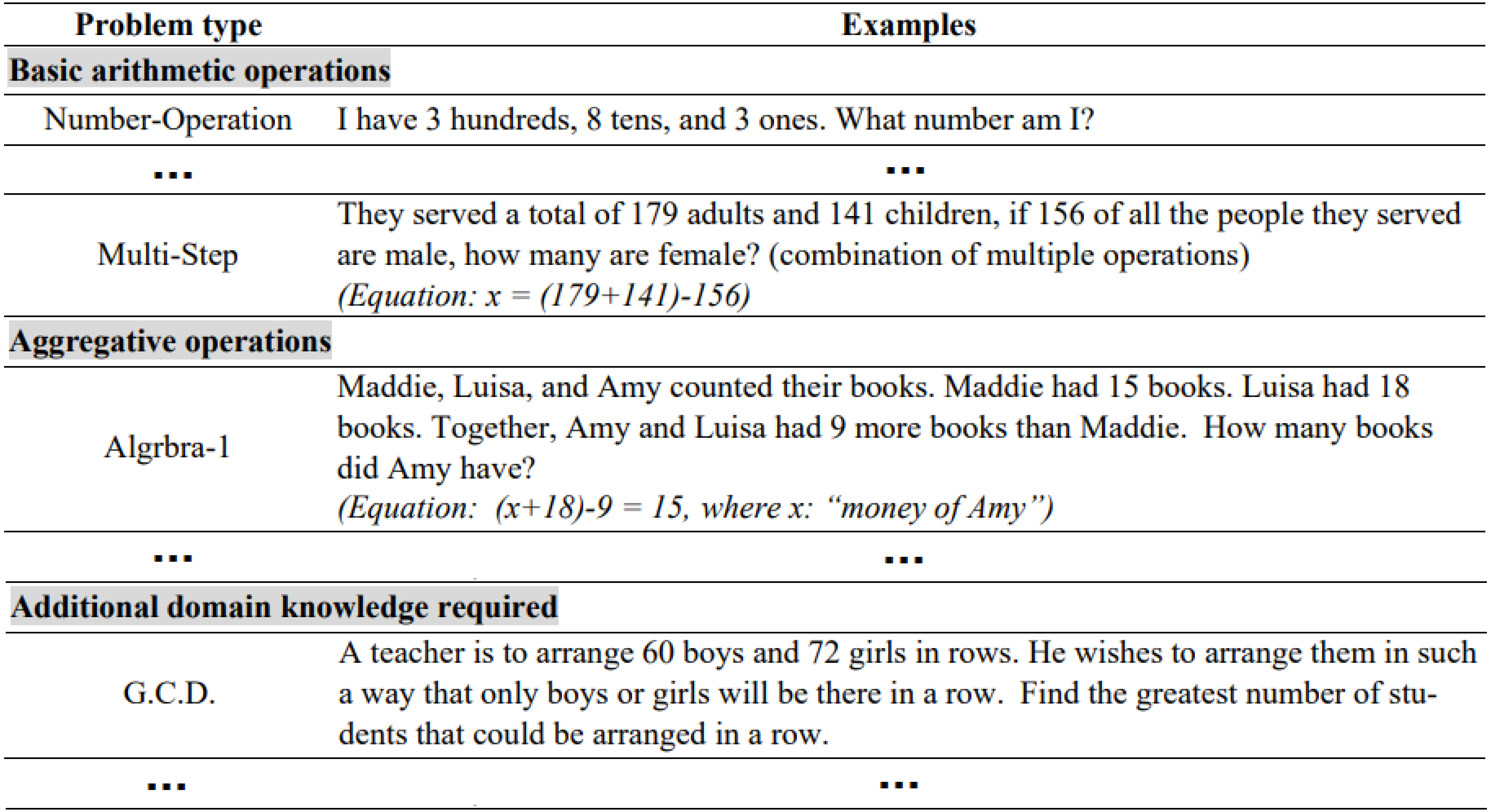
STRUCTURAL INVARIANCE VARIATION:
Text: Jack gave 3 pens to Mary. If Jack had 8 pens and Mary had 5 pens initially, how many pens does Jack have now?
Equation: 8 - 3 = 5

Arkil Patel et al. Are NLP models really able to solve simple math word problems? NAACL. 2021.

Part 6.

Appendix

- **ASDiv (Academia Sinica Diverse MWP Dataset) (miao et al.,**



Problem type	Examples
Basic arithmetic operations	
Number-Operation	I have 3 hundreds, 8 tens, and 3 ones. What number am I?
...	...
Multi-Step	They served a total of 179 adults and 141 children, if 156 of all the people they served are male, how many are female? (combination of multiple operations) *(Equation: x = (179+141)-156)*
Aggregative operations	
Algrbra-1	Maddie, Luisa, and Amy counted their books. Maddie had 15 books. Luisa had 18 books. Together, Amy and Luisa had 9 more books than Maddie. How many books did Amy have? *(Equation: (x+18)-9 = 15, where x: "money of Amy")*
...	...
Additional domain knowledge required	
G.C.D.	A teacher is to arrange 60 boys and 72 girls in rows. He wishes to arrange them in such a way that only boys or girls will be there in a row. Find the greatest number of students that could be arranged in a row.
...	...

Shen-yun Miao et al. A diverse corpus for evaluating and developing english math word problem solvers. ACL. 2020.