Lecture 4: Scientific QA

25 Spring CSCE-689 SPTP: NLP4Science

Presented by Yichen Tao
01/30/2025

Refresh from previous lecture...

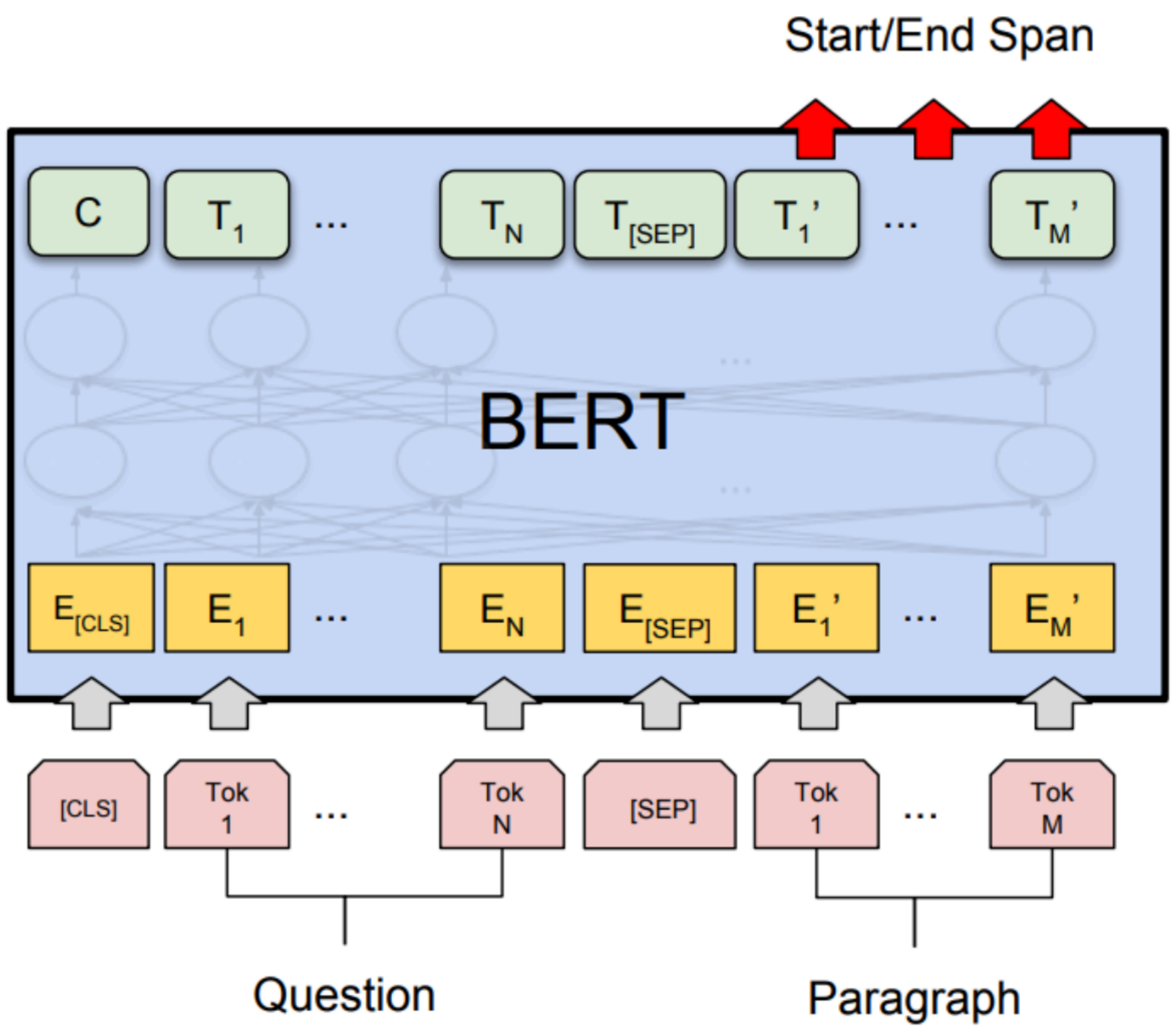
- Encoder-only Transformers: BERT
- (Extractive) question answering -> ternary classification

Context: Corynebacterium minutissimum is the bacteria that leads to cutaneous eruptions of erythrasma and is the most common cause of interdigital foot infections. It is found mostly in occluded intertriginous areas such as the axillae, …

Question: Which bacteria causes erythrasma?

⇩

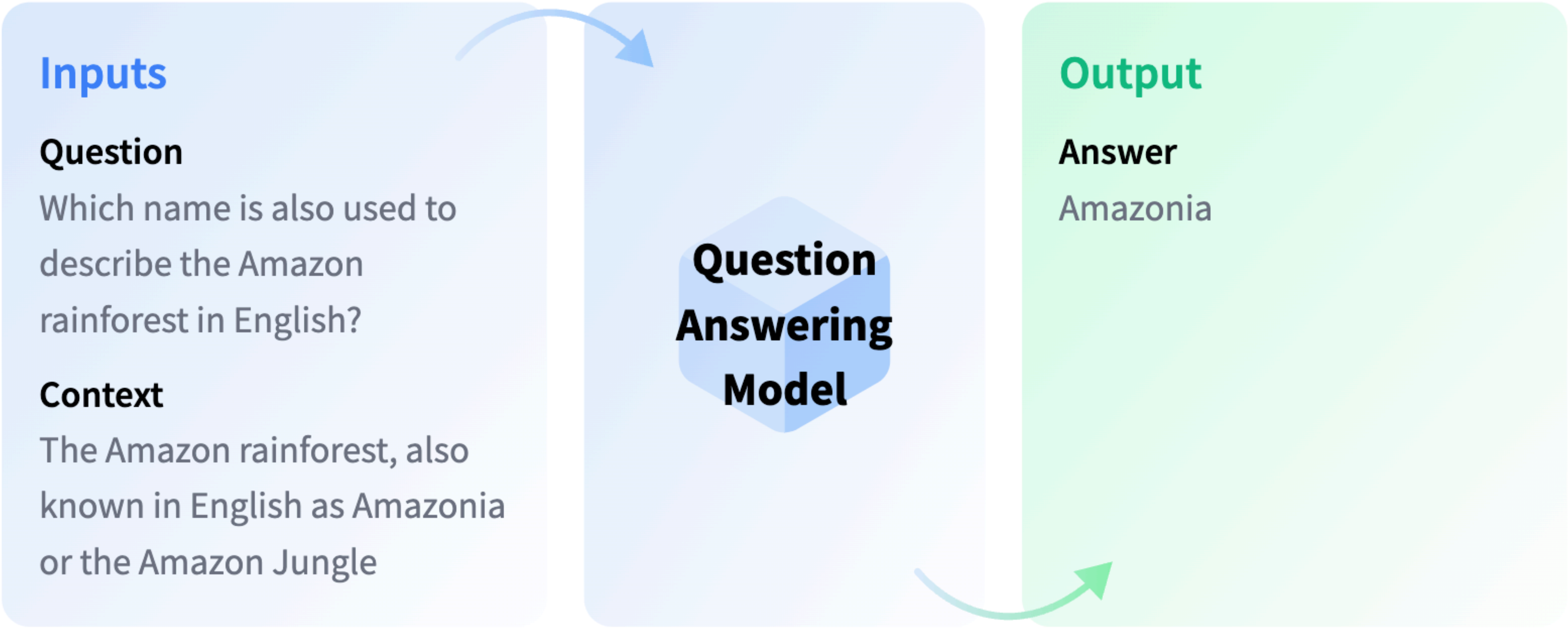
Answer: <u>Corynebacterium minutissimum</u> is the bacteria that leads to cutaneous …



BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
https://yuzhang-teaching.github.io/CSCE689-S25/L2.pdf

Question Answering Task Overview

- **Extractive QA:** The model extracts the answer from a context.
- **Generative QA:** The model generates the answer, usually with decoder-only architectures.
  - Open Generative QA: with context example
  - Close Generative QA: without context



https://huggingface.co/tasks/question-answering

Agenda

- **Pseudo QA (classification):** PubMedQA
- **Close generative question answering:** MetaMath
- **Open generative question answering:** Better to Ask in English

Agenda

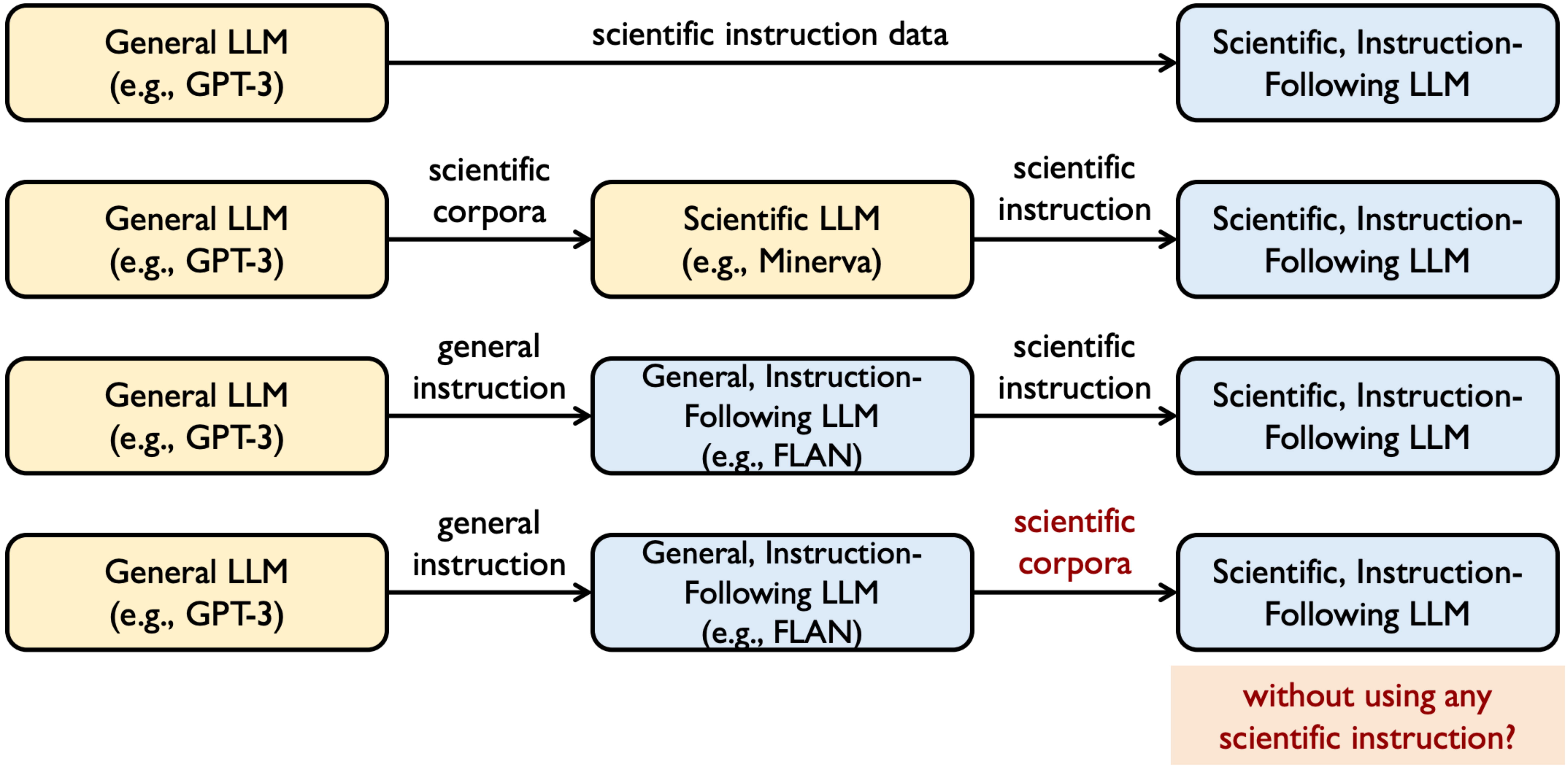
- **Pseudo QA (classification):** PubMedQA
- **Close generative question answering:** MetaMath
- **Open generative question answering:** Better to Ask in English

Why we need QA?

- **OpenAI’s 5 steps to AGI**
  - L1: conversational AI
  - L2: reasoner: solving complex problems requiring advanced reasoning
- **Math == Reasoning?**

Domain specific LLMs

- Need domain specific dataset for SFT



https://yuzhang-teaching.github.io/CSCE689-S25/L3.pdf

How to build a QA dataset?

- **Common solution 1:**
  - Use experienced annotator to manually create Q/A pair
  - *Very expensive!*
- **Common solution 2:**
  - Use existing models to perform extractive QA task on corpora
  - *Can a new model trained on this dataset outperform the model creating it?*
  - *The extracted QA pair does not require reasoning to answer*
- **Therefore… no large dataset require reasoning available!**

How to build a QA dataset?

- **Hybrid Solution (1+2)?**
  - Use PubMed dataset as corpora
  - **Step 1:** manually label small portion of data
  - **Step 2:** automatically collect unlabelled data, and use agreement rate with the manual portion to validate collection
  - **Step 3:** automatically convert titles/abstracts to question

Statistic	PQA-L	PQA-U	PQA-A
Number of QA pairs	1.0k	61.2k	211.3k
Prop. of yes (%)	55.2	–	92.8
Prop. of no (%)	33.8	–	7.2
Prop. of maybe (%)	11.0	–	0.0
Avg. question length	14.4	15.0	16.3
Avg. context length	238.9	237.3	238.0
Avg. long answer length	43.2	45.9	41.0

Table 1: PubMedQA dataset statistics.

PubMedQA: A Dataset for Biomedical Research Question Answering, EMNLP '19

How to build a QA dataset?

- **3:** automatically convert titles to question
  - Using fixed rule-based conversion based on [1]
  - Why not LLM?

Original Statement Title	Converted Question	Label	%
Spontaneous electrocardiogram alterations ***predict*** ventricular fibrillation in Brugada syndrome.	***Do*** spontaneous electrocardiogram alterations ***pre-dict*** ventricular fibrillation in Brugada syndrome?	***yes***	92.8
Liver grafts from selected older donors ***do not have*** significantly more ischaemia reperfusion injury.	***Do*** liver grafts from selected older donors ***have*** sig-nificantly more ischaemia reperfusion injury?	***no***	7.2

Table 2: Examples of automatically generated instances for PQA-A. Original statement titles are converted to questions and answers are automatically generated according to the negation status.

[1] Stanford CoreNLP parser (Manning et al., 2014)

How to show the dataset is useful?

- Fine-tune a few models using PubMedQA dataset
- Show the model is stronger compared to baseline

{BioBERT, ESIM, BiLSTM}

PQA-A

PQA-L

PQA-A w/ Long answer

PQA-L w/ Long answer

Inference on PQA-U

Filter PQA-U where any logit > threshold

Definition:
- **PQA-L (1k):** Human annotator labeled
- **PQA-U (61k):** Unlabeled
- **PQA-A (211k):** Generated question using rule-based label

BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics ‘20

Experimental results of SFT with PubMedQA

Model	Final Phase Only		Single-phase		Phase I + Final		Phase II + Final		Multi-phase	
	Acc	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**
Majority	55.20	23.71	–	–	–	–	–	–	–	–
Human (single)	78.00	72.19	–	–	–	–	–	–	–	–
w/o A.S.										
Shallow Features	53.88	36.12	<u>57.58</u>	31.47	57.48	37.24	56.28	<u>40.88</u>	53.50	39.33
BiLSTM	55.16	23.97	55.46	39.70	58.44	40.67	52.98	33.84	59.82	<u>41.86</u>
ESIM w/ BioELMo	53.90	32.40	61.28	42.99	61.96	43.32	60.34	44.38	62.08	45.75
BioBERT	56.98	28.50	66.44	47.25	66.90	46.16	66.08	50.84	67.66	52.41
w/ A.S.										
Shallow Features	53.60	35.92	57.30	30.45	55.82	35.09	56.46†	40.76	55.06†	40.67†
BiLSTM	55.22†	23.86	55.96†	40.26†	<u>61.06</u>†	41.18†	54.12†	34.11†	58.86	41.06
ESIM w/ BioELMo	53.96†	31.07	62.68†	43.59†	<u>63.72</u>†	47.04†	60.16	45.81†	<u>63.72</u>†	<u>47.90</u>†
BioBERT	57.28†	28.70†	66.66†	46.70†	67.24†	46.21†	66.44†	51.41†	<u>**68.08**</u>†	<u>**52.72**</u>†

Table 5: Main results on PQA-L test set under reasoning-required setting. A.S.: additional supervision. †with A.S. is better than without A.S. Underlined numbers are model-wise best performance, and bolded numbers are global best performance. All numbers are percentages.

Agenda

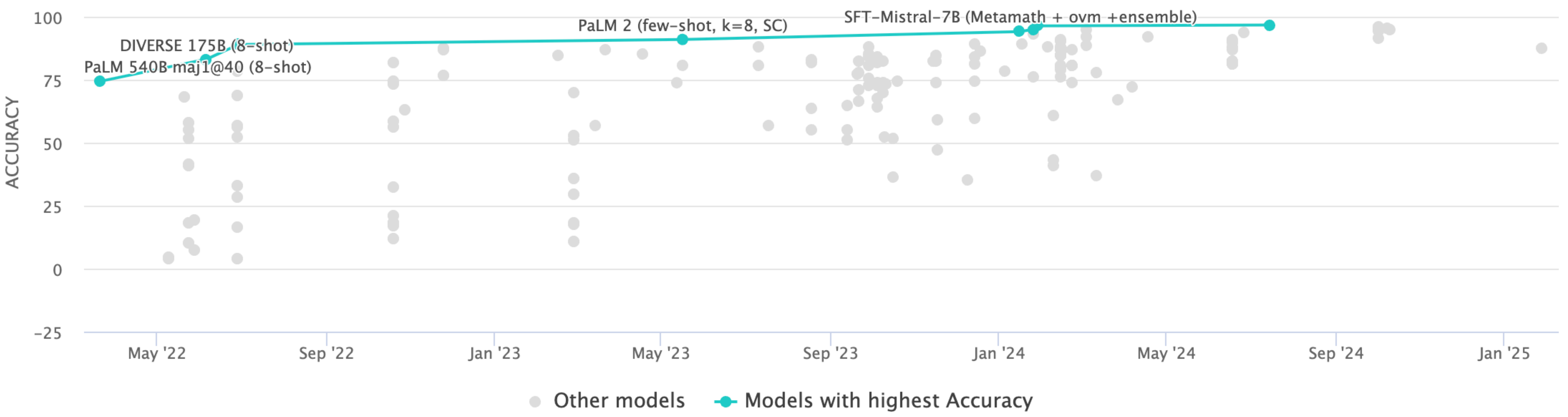
- **Pseudo QA (classification):** PubMedQA
- **Close generative question answering:** MetaMath
- **Open generative question answering:** Better to Ask in English

Myth of math problem solving

- **OpenAI's 5 steps to AGI**
  - L1: conversational AI
  - L2: reasoner: solving complex problems requiring advanced reasoning
- **Math == Reasoning?**

Myth of math problem solving

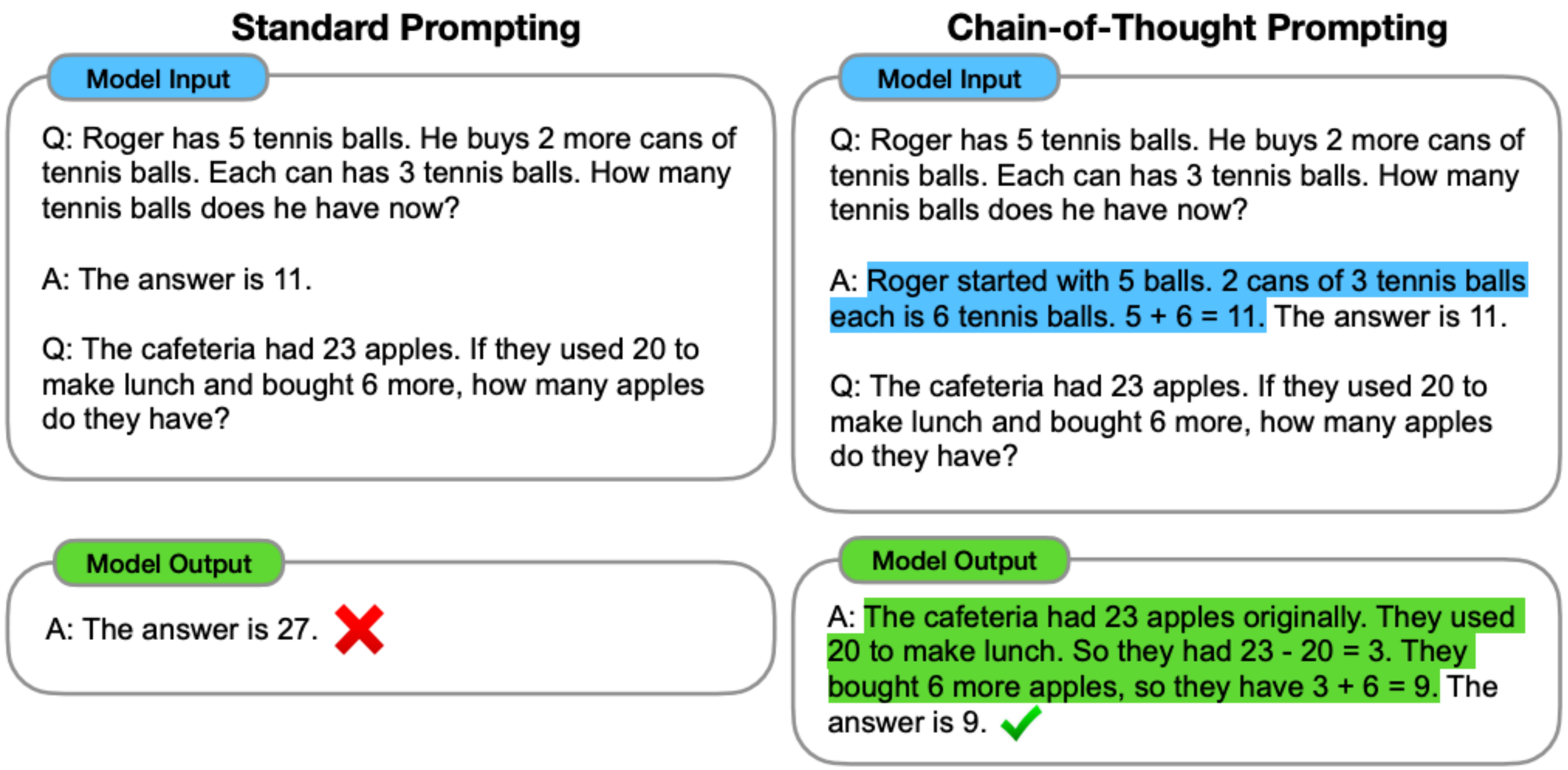
- GSM8K: High school math problems, taking 2-8 steps to solve.

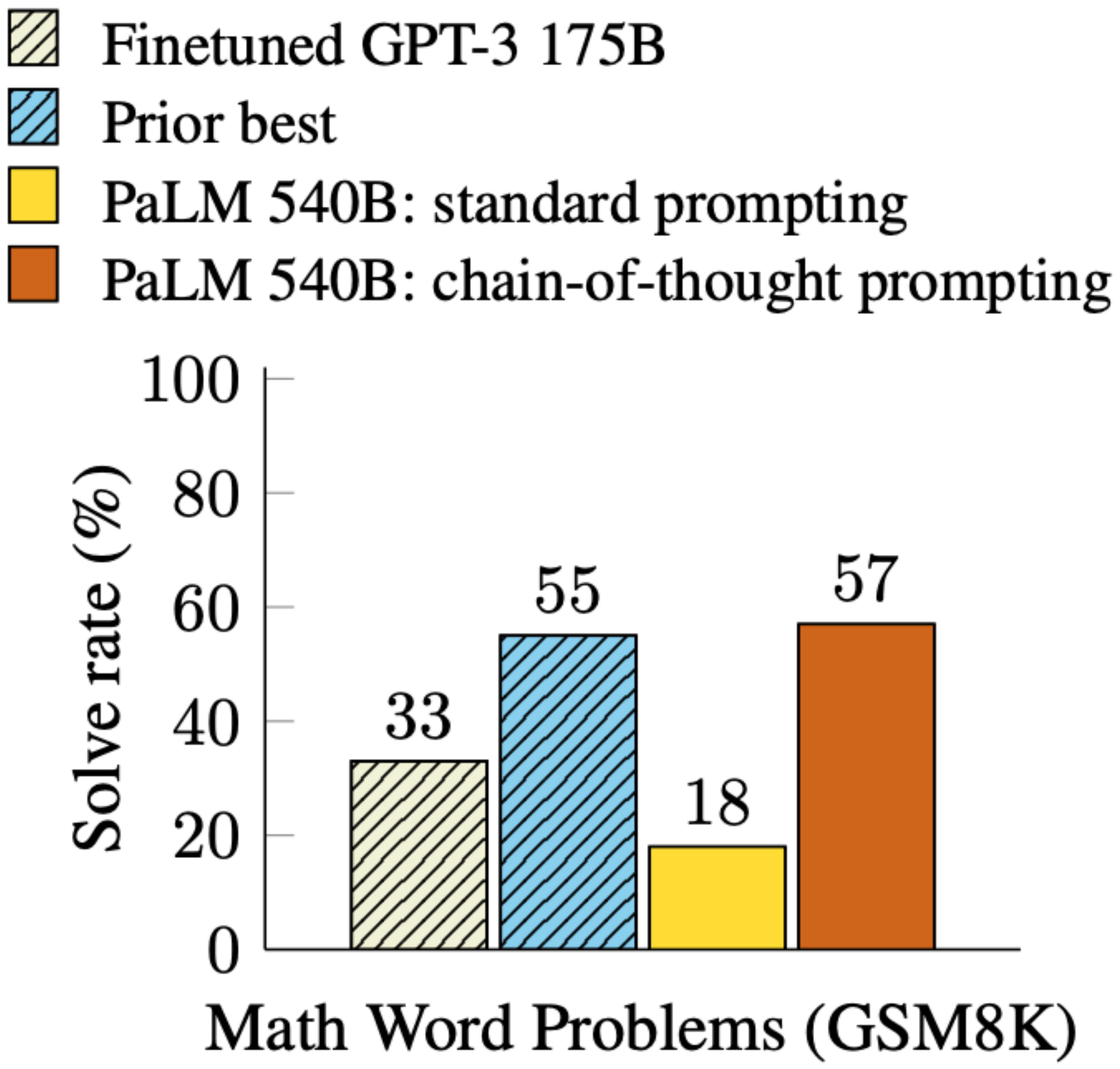


https://paperswithcode.com/sota/arithmetic-reasoning-on-gsm8k

Chain-of-thought prompting

- Add intermediate steps in ICL example
- Or even easier, add “show your steps” in prompt

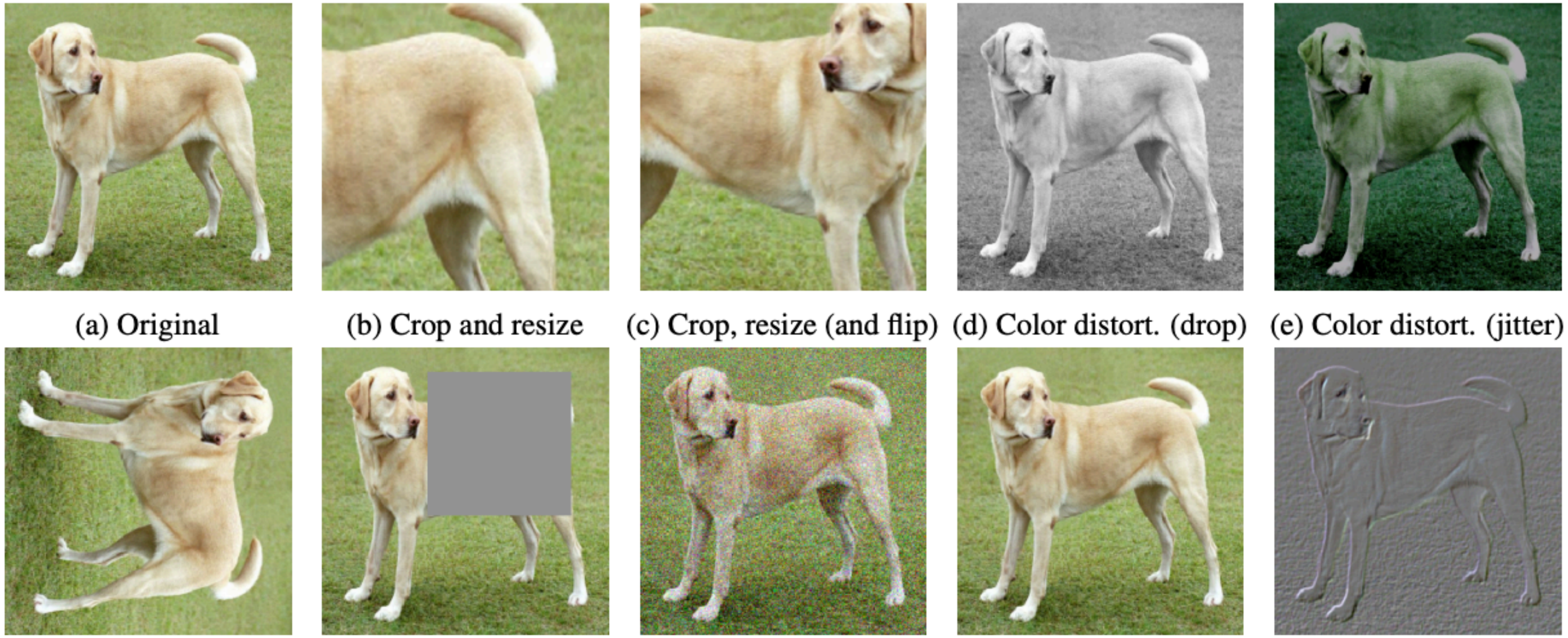




Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, NeurIPS ‘22

A lesson from CV: data augmentation

- Augment one datapoint with different ‘view’ of it



(a) Original (b) Crop and resize (c) Crop, resize (and flip) (d) Color distort. (drop) (e) Color distort. (jitter)

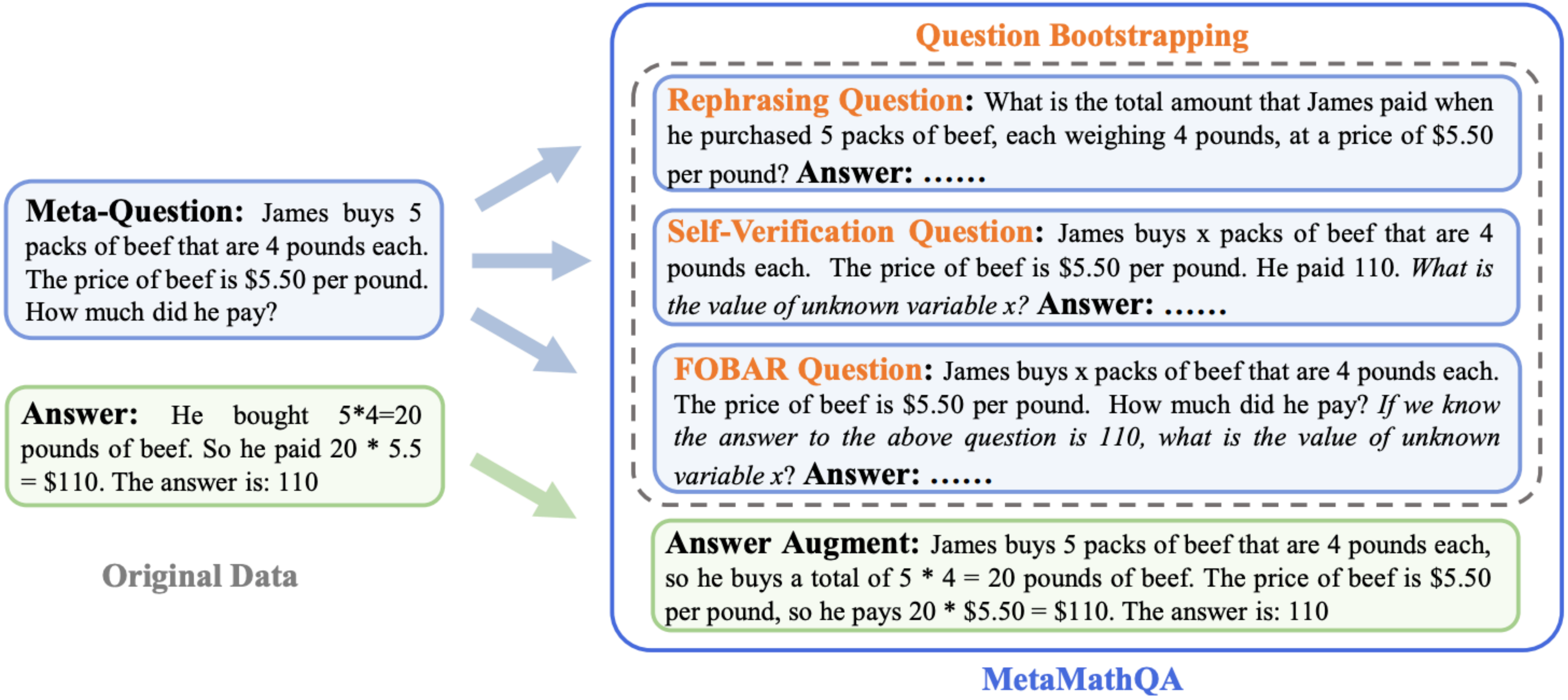
(f) Rotate $\{90°, 180°, 270°\}$ (g) Cutout (h) Gaussian noise (i) Gaussian blur (j) Sobel filtering

Figure 4. Illustrations of the studied data augmentation operators. Each augmentation can transform data stochastically with some internal parameters (e.g. rotation degree, noise level). Note that we *only* test these operators in ablation, the *augmentation policy used to train our models* only includes *random crop (with flip and resize)*, *color distortion*, and *Gaussian blur*. (Original image cc-by: Von.grzanka)

A Simple Framework for Contrastive Learning of Visual Representations, ICML ‘20

MetaMath: augmenting math QAs

- Augment one datapoint with different 'view' of it



metamath: bootstrap your own mathematical questions for large language models, ICLR '24

How?

- Question bootstrpping with LLM

Example A.1: Prompt for Rephrasing GSM8K Questions

You are an AI assistant to help me rephrase questions. Follow the given examples.

Question: Olivia has $23. She bought five bagels for $3 each. How much money does she have left?
Rephrase the above question: What is the amount of money that Olivia has left after purchasing five bagels for $3 each, if she initially had $23?

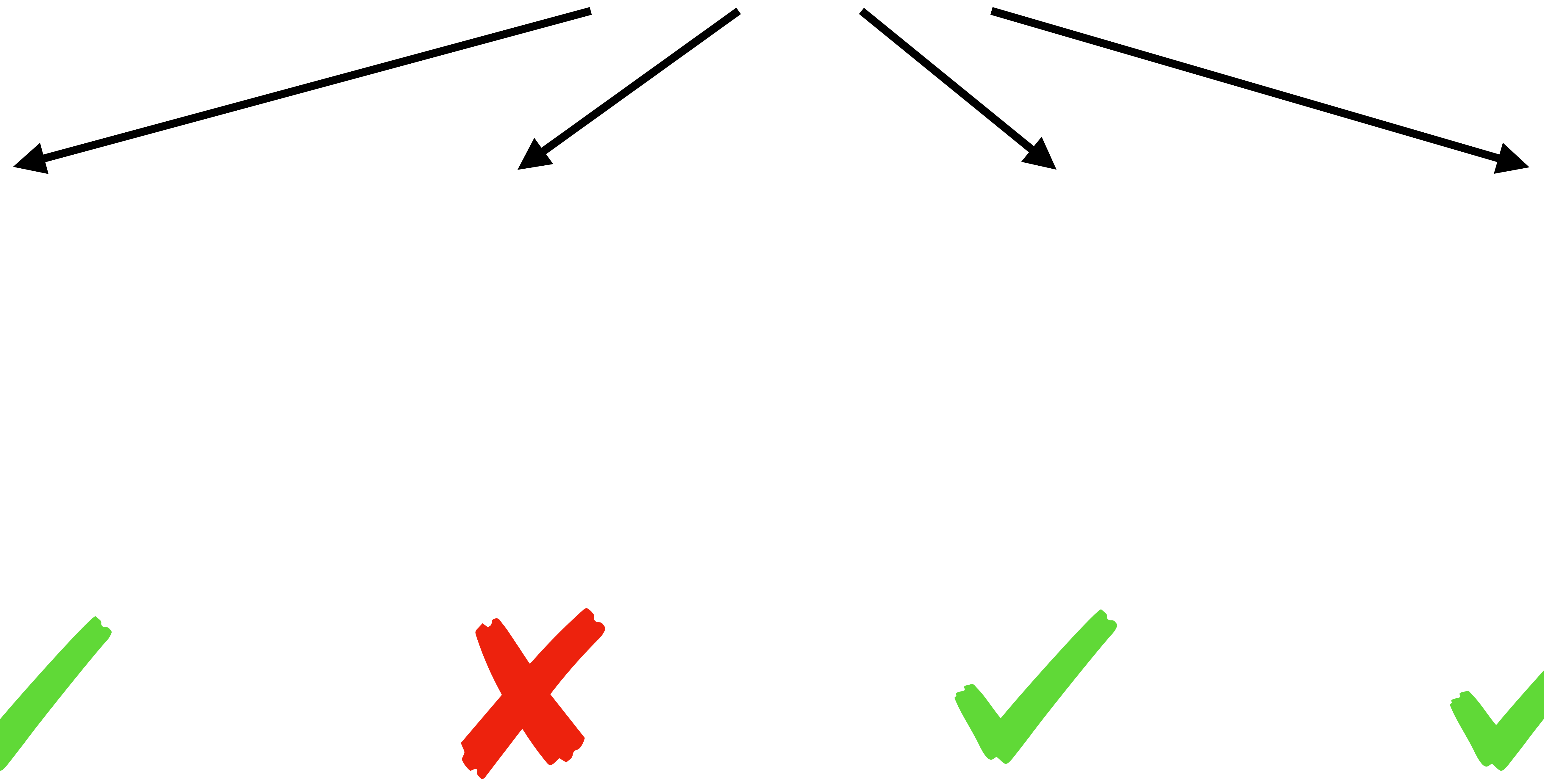
Question: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of wednesday?
Rephrase the above question: After losing 23 golf balls on Tuesday and an additional 2 on Wednesday, how many golf balls does Michael have left if he initially had 58 golf balls?

metamath: bootstrap your own mathematical questions for large language models, ICLR '24

How?

- Answer bootstrapping with LLM

Answer this question? x4



metamath: bootstrap your own mathematical questions for large language models, ICLR '24

Experiment results of MetaMath

- One augmentation bring 18% accuracy increase on GSM8K

Method	GSM8K						MATH					
	AnsAug	Rep.	SV	FOBAR	GSM8K	MATH	AnsAug	Rep.	SV	FOBAR	GSM8K	MATH
SFT [70]	✗	✗	✗	✗	41.6	3.0	✗	✗	✗	✗	13.8	4.7
MetaMath	✓	✗	✗	✗	59.6	4.4	✓	✗	✗	✗	28.4	12.9
	✗	✓	✗	✗	59.7	4.4	✗	✓	✗	✗	30.4	12.4
	✓	✓	✗	✗	60.6	4.4	✓	✓	✗	✗	29.1	15.3
	✓	✓	✓	✓	**64.4**	**5.7**	✓	✓	✓	✓	**34.6**	**17.7**

Table 3: Effect of different question augmentation with LLaMA-2-7B finetuned on GSM8K or MATH.

metamath: bootstrap your own mathematical questions for large language models, ICLR '24

Beyond data augmentation

- Reinforcement learning technique
  - Use LLM to generate reward for its own response [1]

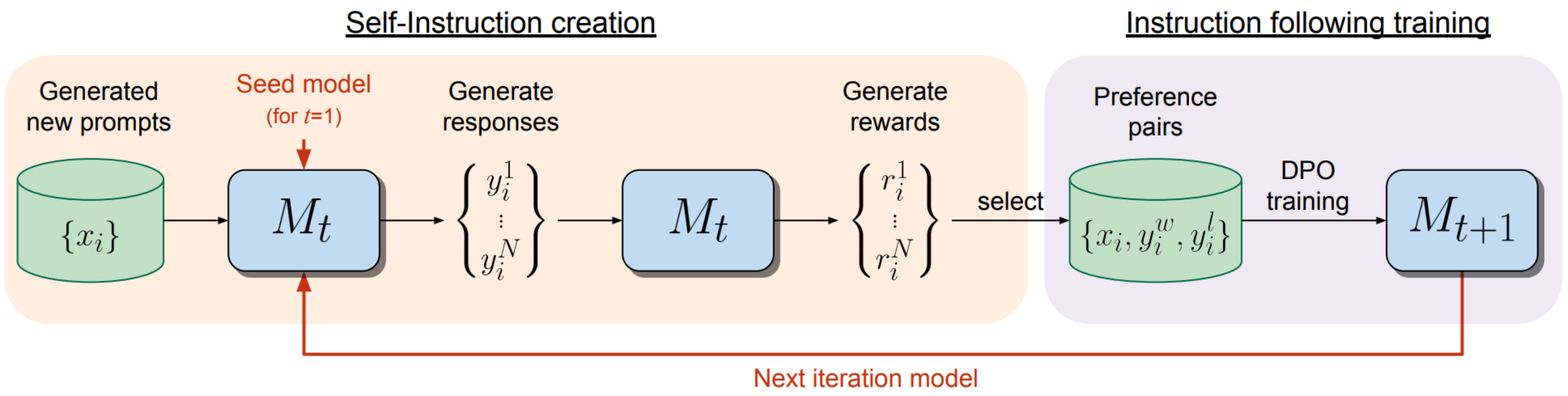


Figure 1: **Self-Rewarding Language Models.** Our self-alignment method consists of two steps: (i) *Self-Instruction creation*: newly created prompts are used to generate candidate responses from model M_t, which also predicts its own rewards via LLM-as-a-Judge prompting. (ii) *Instruction following training*: preference pairs are selected from the generated data, which are used for training via DPO, resulting in model M_{t+1}. This whole procedure can then be iterated resulting in both improved instruction following and reward modeling ability.

[1] Self-rewarding language models, ICML '24

Beyond data augmentation

- Reinforcement learning technique
  - Use LLM to generate rational for Q/A pair, and reward correct rationale [2]

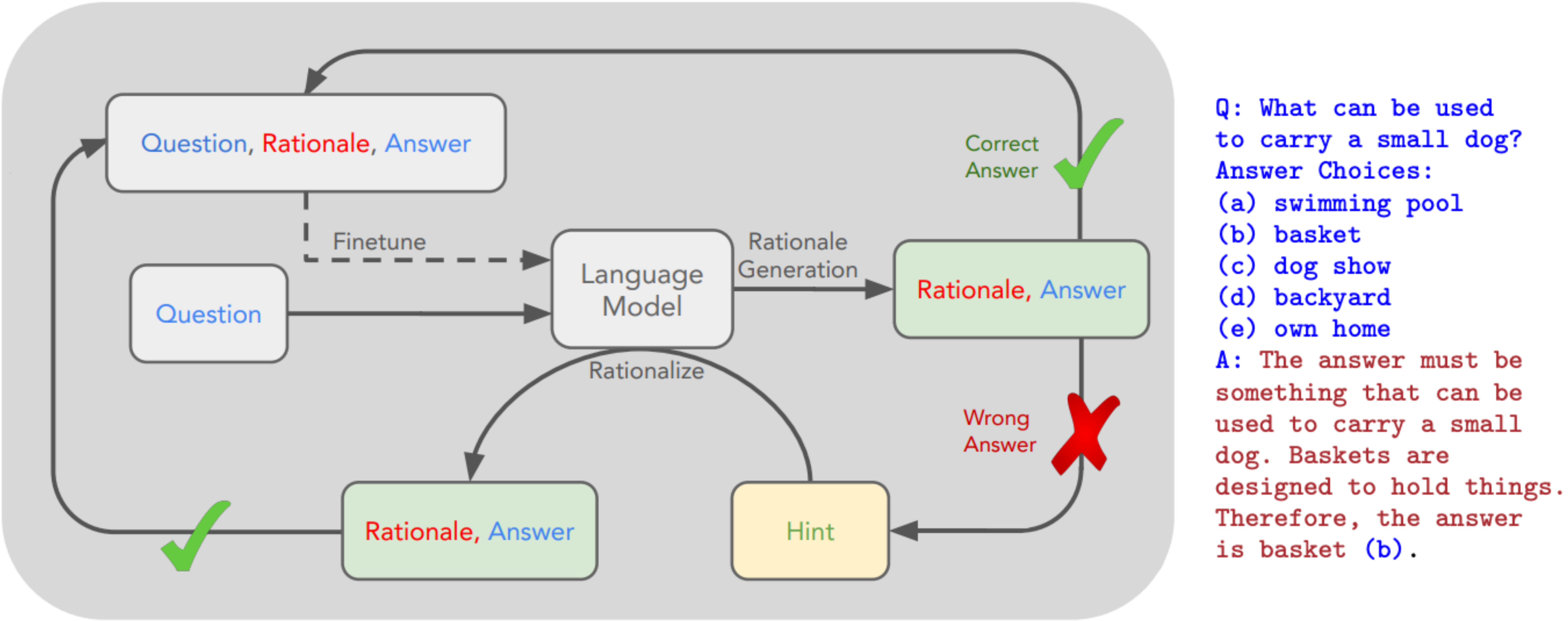
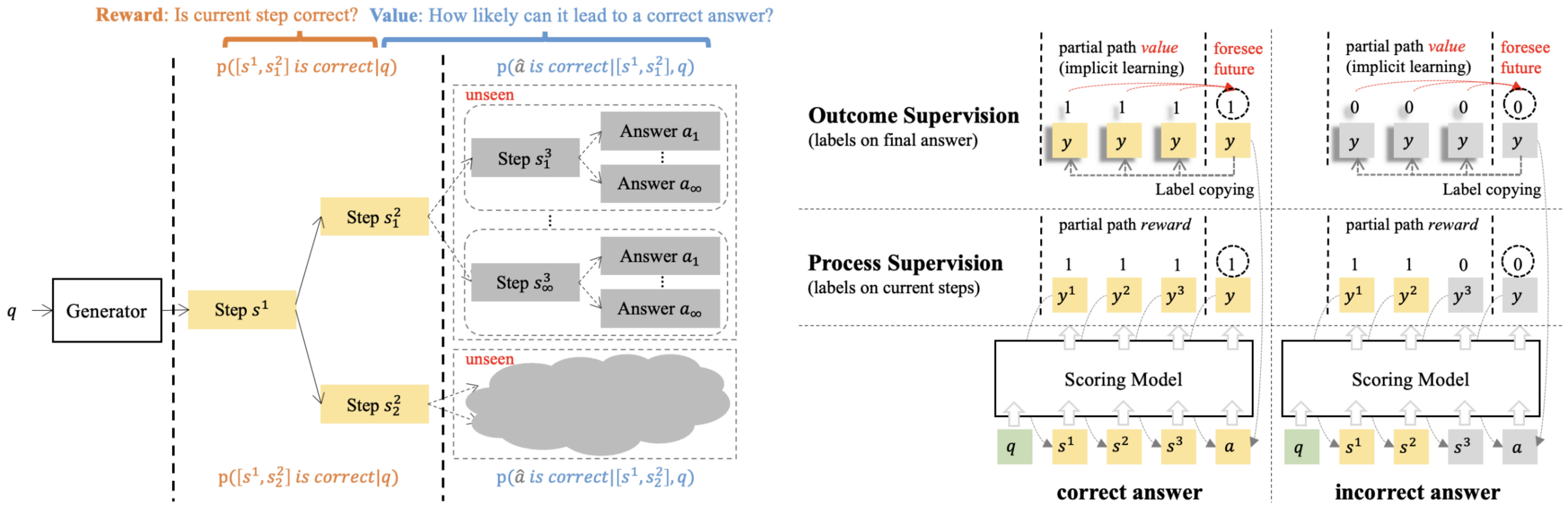


Figure 1: An overview of STaR and a STaR-generated rationale on CommonsenseQA. We indicate the fine-tuning outer loop with a dashed line. The questions and ground truth answers are expected to be present in the dataset, while the rationales are generated using STaR.

[2] STaR: Self-Taught Reasoner Bootstrapping Reasoning With Reasoning, NeruIPS '22

Beyond data augmentation

- Value model to detect early stage error in chain-of-thought [3]



(a) Reward and value

(b) Outcome supervision and process supervision on training models to evaluate *complete paths*

Figure 1: (a): When evaluating partial paths (here for the first two steps), reward focuses on the current states, while value focuses on the unseen future outcomes. (b): Given a question q and a solution path $[s^1, \cdots, s^m, a]$, models are trained to predict path correctness (circled output scalar on the last token). Outcome supervision replicates the final answer's correctness label across all steps (indicated by shaded labels), causing the model to implicitly learn to foresee the future, predicting values for partial paths. By contrast, process supervision details per-step correctness labels, causing the model to learn to predict step-level correctness, i.e. reward. Correct steps and answers are colored in yellow and incorrect ones in grey.

[3] OVM, Outcome-supervised Value Models for Planning in Mathematical Reasoning, NAACL '24

Beyond data augmentation

- LLM with RL
  - GPT-o1 achieves 83.3% accuracy on AIME@24
  - DeepSeek-R1 achieves 79.8% accuracy on AIME@24
  - GPT-4o achieves 13.4% on AIME@24

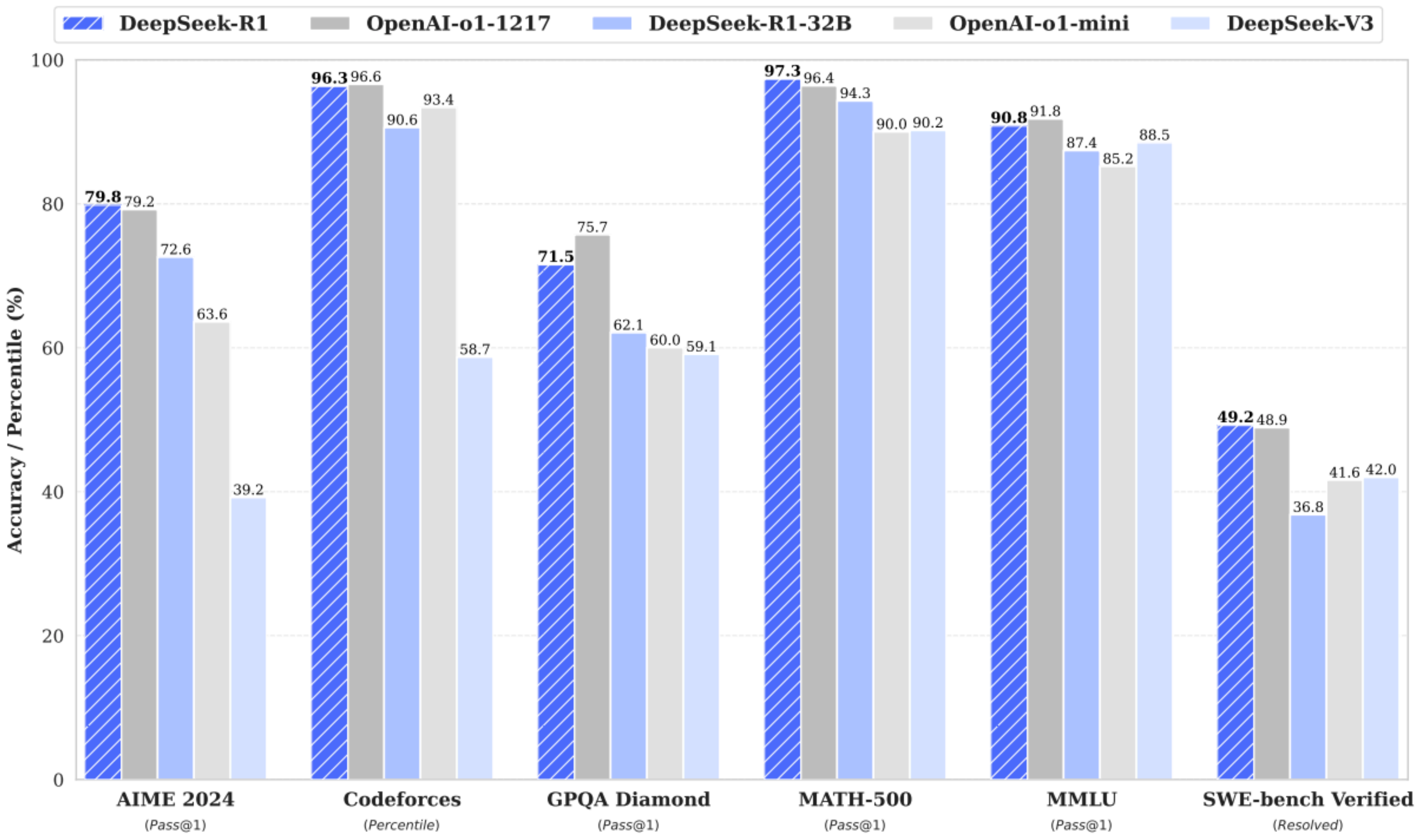


Figure 1 | Benchmark performance of DeepSeek-R1.

DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

Takeaways

- Reasoning capability is important milestone to AGI
- Both data augmentation and inference time techniques (CoT, ensemble, etc.) widely used to improve reasoning capability for LLMs
- RL based methods further improve LLM's reasoning capability

Agenda

- **Pseudo QA (classification):** PubMedQA
- **Close generative question answering:** MetaMath
- **Open generative question answering:** Better to Ask in English

Important question for generative QA

- **How to evaluate a generated sentence?**
- With ground truth label (discussed last lecture)
  - Traditional NLP metrics
    - ROUGE, BLEU
  - LLM metrics
    - BERTScore: use BERT's [CLS] token embedding to calculate cosine similarity
    - GPT: prompt to get a score, more of a rating

Important question for generative QA

- **How to evaluate a generated sentence?**
- Without ground truth label (i.e. consistency, human preference, pick correct answer, etc.)
  - LLM-as-a-judge
  - Agreement rate on par with human-human agreement rate
  - Also works for BERTScore / GPTScore usecase

Question: If the FED buys bonds in the secondary market
(A) the money supply will increase. (B) the money supply will decrease. (C) the money supply will not be affected. (D) the discount rate would be affected. Output the answer only.

Assistant A:

The Federal Reserve buys bonds in the secondary market to increase the money supply.

Assistant B:

(A) the money supply will increase.

Follow-up Question: How does it affect my daily life? Give 3 examples.

Assistant A:

The Federal Reserve buys bonds in the secondary market to increase the money supply. This affects your daily life by:
1. Increasing the money supply.
2. Increasing the money supply and lowering the discount rate.
3. Increasing the money supply and lowering the discount rate and increasing the money supply.

Assistant B:

When the Federal Reserve buys bonds in the secondary market, it can have several effects on the economy that can impact your daily life. Here are three examples:

1. Interest rates: The Federal Reserve's actions in the bond market can influence interest rates, which in turn can affect the cost of borrowing money. For example, if the Federal Reserve buys bonds and drives up demand for them, interest rates on things like mortgages, car loans, and credit cards may go up, making it more expensive to borrow money.
2. Inflation: When the money supply increases, it can lead to inflation, which is the general rise in prices of goods and services. This can impact your daily life by making it more expensive to buy things like groceries, gas, and other necessities.
3. Employment: The Federal Reserve's actions in the bond market can also have an impact on employment. For example, if the Federal Reserve's actions lead to economic growth and increased demand for goods and services, it can create more job opportunities and boost employment.

GPT-4 Judgment:

Assistant A provided an incorrect response to the user's question about how the Federal Reserve buying bonds in the secondary market affects daily life. The answer given is repetitive and lacks clear examples of how the action impacts daily life.
On the other hand, **Assistant B** provided a relevant and accurate response to the user's question about the Federal Reserve buying bonds. The answer includes three clear examples of how the action impacts daily life, such as interest rates, inflation, and employment.

Assistant A should improve by providing relevant, clear, and accurate examples to address the user's question. **Assistant B**'s response is more helpful, relevant, and detailed.

Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, NeruIPS '23

Better to Ask in English

- **Framework for identifying Cross-lingual LLM performance degradation**

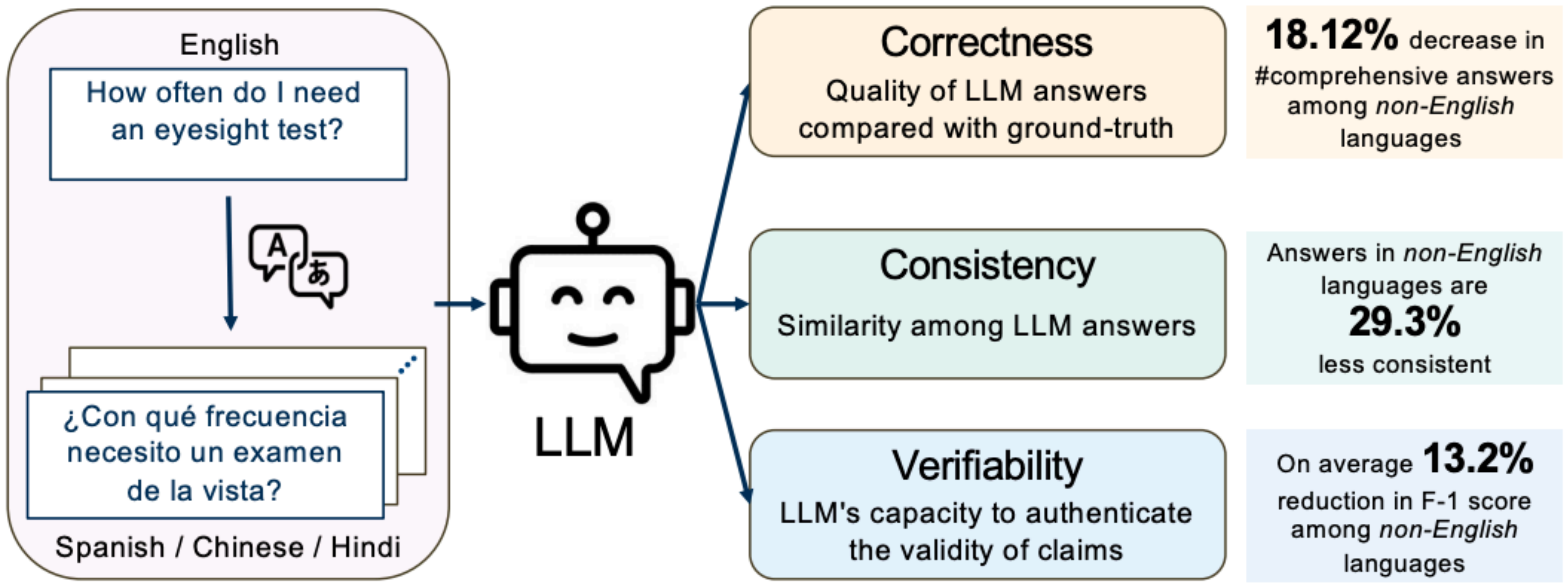


Figure 1: We present XLINGEVAL, a comprehensive framework for assessing cross-lingual behaviors of LLMs for high risk domains such as healthcare. We present XLINGHEALTH, a cross-lingual benchmark for healthcare queries.

Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, NeruIPS '23

Better to Ask in English

- **Correctness:** LLM evaluating Q/GT pair with generated answer in different language

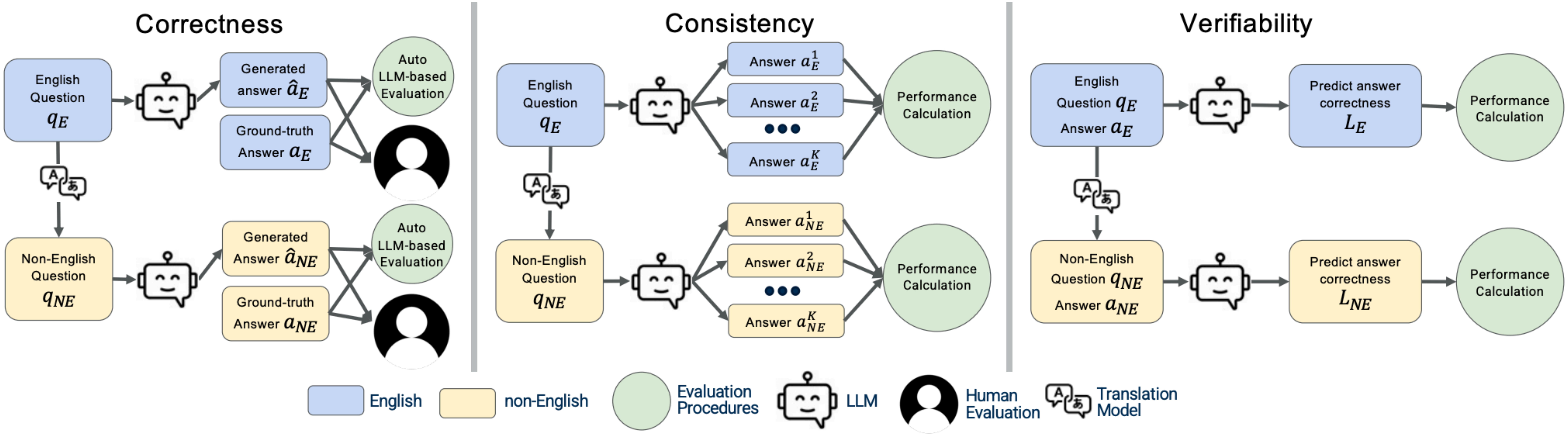


Figure 2: Evaluation pipelines for correctness, consistency, and verifiability criteria in the XLingEval framework.

Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, NeruIPS '23

Better to Ask in English

- **Consistency:** LLM evaluating if list of answers are consistent

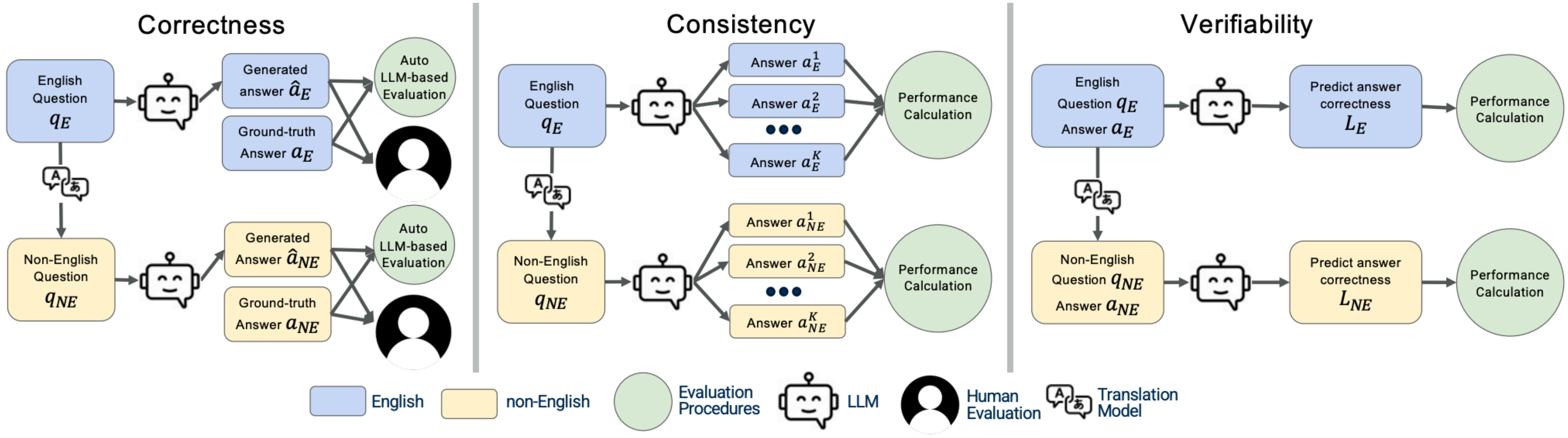


Figure 2: Evaluation pipelines for correctness, consistency, and verifiability criteria in the XLingEval framework.

Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, NeruIPS '23

Better to Ask in English

- **Verifiability:** LLM decide if the answer is relevant to question

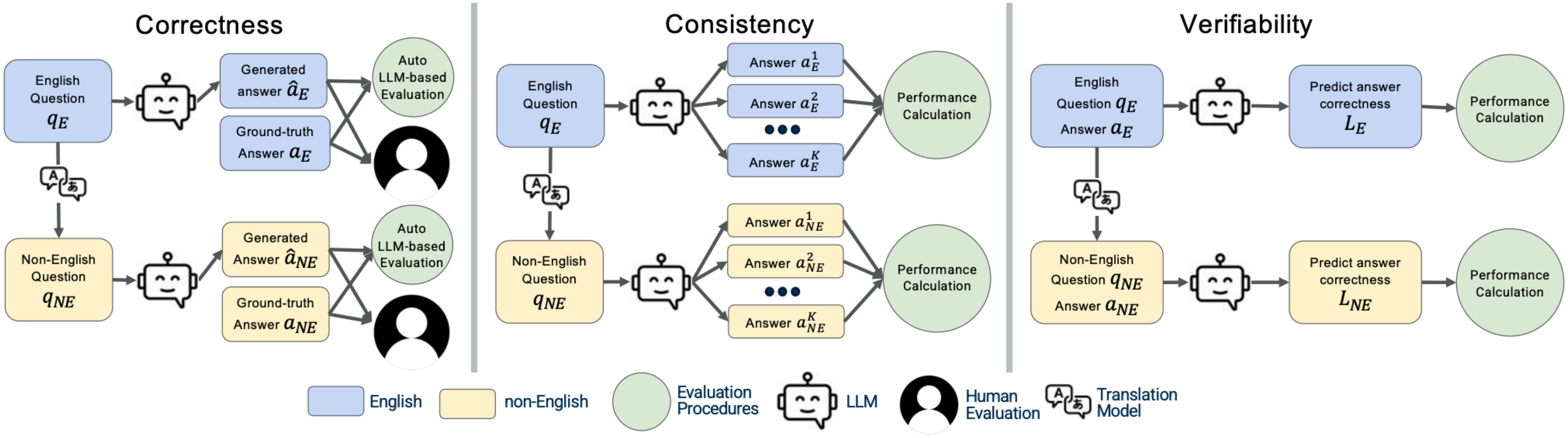


Figure 2: Evaluation pipelines for correctness, consistency, and verifiability criteria in the XLINGEVAL framework.

Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, NeruIPS '23

Experiment results: Better to Ask in English

- **Correctness:** human verification is introduced for a subset of full test set to ensure the quality of automated evaluation

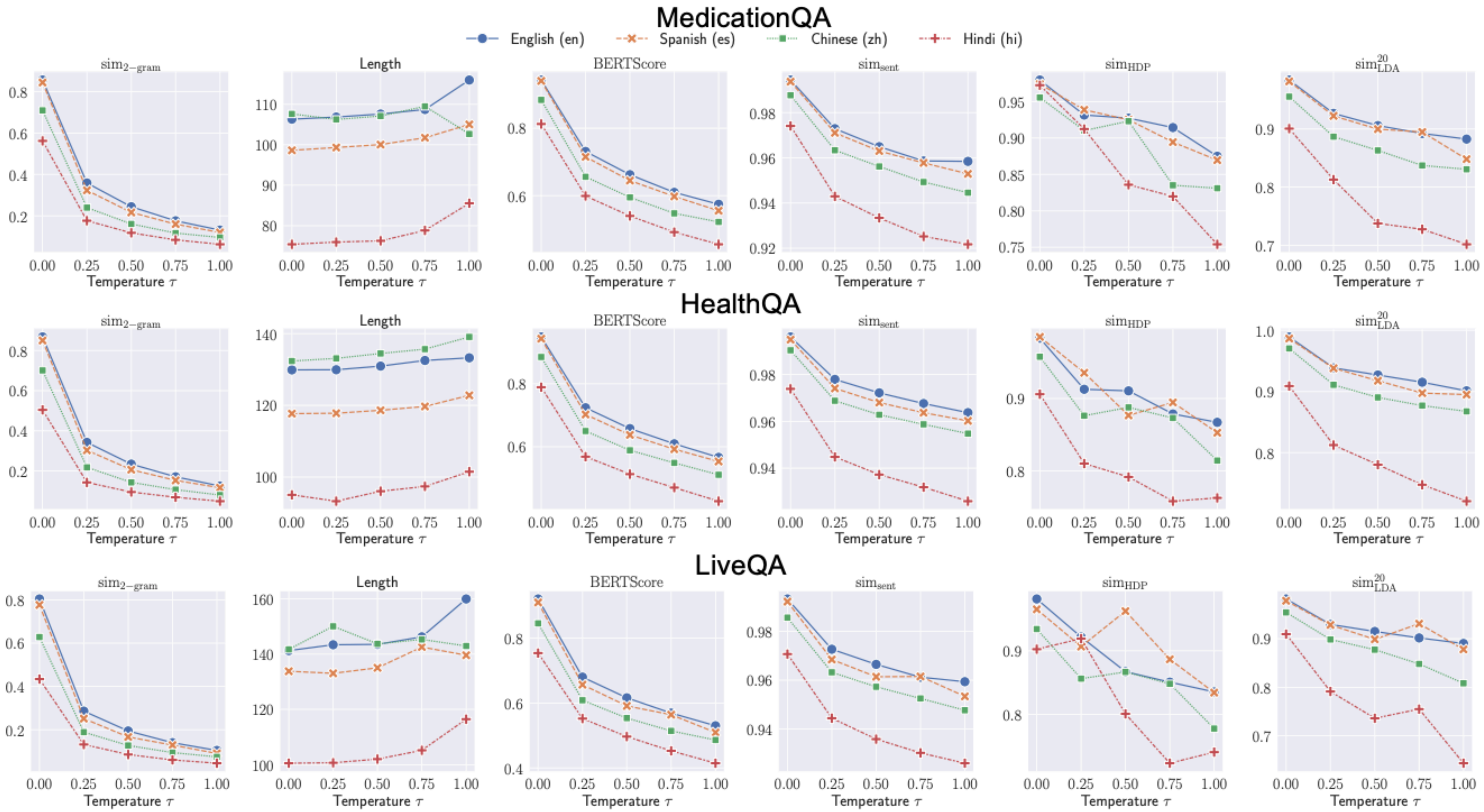
Table 1: Automated correctness evaluation in four languages: English (en), Spanish (es), Chinese (zh), and Hindi (hi) for GPT-3.5. Each number represents the number of answers assigned to the respective label in the dataset.

Information Comparison (LLM Answer vs ground-truth Answer)	HealthQA				LiveQA				MedicationQA			
	en	es	zh	hi	en	es	zh	hi	en	es	zh	hi
More comprehensive and appropriate	1013	891	878	575	226	213	212	142	618	547	509	407
Less comprehensive and appropriate	98	175	185	402	3	12	16	59	18	50	41	125
Neither contradictory nor similar	20	63	57	110	14	20	14	32	49	70	92	107
Contradictory	3	5	14	47	3	1	4	13	5	23	48	51

Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, NeruIPS '23

Experiment results: Better to Ask in English

- **Consistency:** degradation is common for all languages except English



Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, NeruIPS '23

Experiment results: Better to Ask in English

Figure 4: Results of LiveQA on metrics of the verifiability experiment, including macro precision, macro recall, macro F1-score, accuracy, and area under the curve (AUC). Each column represents a distinct metric. The x- and y-axis of each heatmap represents varying languages and temperatures τ, respectively. The results for the other datasets are in the Appendix (Figure A3)

Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, NeruIPS '23

Discussion: Better to Ask in English

- **Circulating argument?** Is it a fair comparison to use a `performance degraded model` to evaluate result, which is later used to show there will be a performance degradation across different languages?
- **Choice of metric?** Contextualized metric like BERTScore is designed for English. Length metric might not make sense.
  - BERT is pretrained on BookCorpus (EN only) and EN Wikipedia.
  - Think of the example of rewriting math questions. Rewriting can have very different length!
- **Would hallucination change depending on the language?**
  - Hallucination: LLM generates seemingly plausible yet factually unsupported content

A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, Information Systems Vol. 1

Questions? Comments?