Teaching AI to Answer Questions with Reasoning that Actually Helps You

Nishant Balepur

Jordan Boyd-Graber (Chair)
Rachel Rudinger (Co-chair)
Fumeng Yang (Department Representative)
David Weintrop (Dean's Representative)
Shi Feng (External Member, George Washington University)

UNIVERSITY OF MARYLAND

Why do we ask questions?

Goal: Learn Something New

What does “LLM” mean?

Goal: Solve a Multi-Step Problem

How can I get my refund?

Goal: Receive Tailored Advice

How do I get 5 professors in the same room / time?

Goal: Recall Forgotten Information

Who gave that proposal talk with too many emojis?

Why do we ask questions?

Goal: Learn Something New

What does “LLM” mean?

An LLM is trained on...

Goal: Solve a Multi-Step Problem

How can I get my refund?

Please hold for 3 hours

Goal: Receive Tailored Advice

How do I get 5 professors in the same room / time?

It’s impossible

Goal: Recall Forgotten Information

Who gave that proposal talk with too many emojis?

You’re listening to it now

The Central Idea of Question Answering Research:

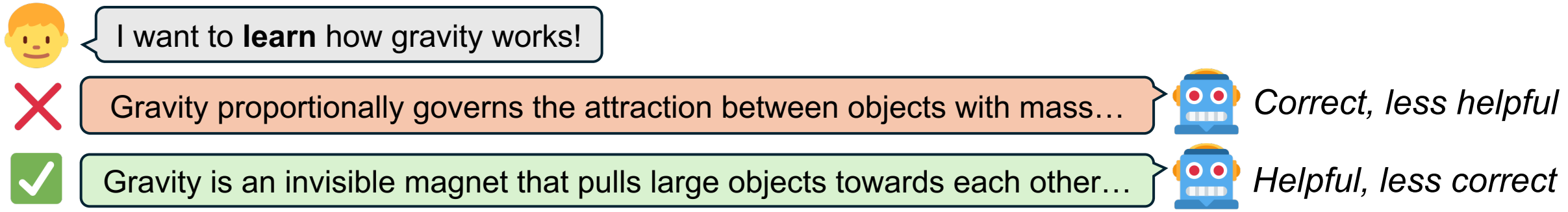
Building systems that answer questions and are **helpful** for these goals

What do we mean by helpfulness?

When helpfulness is discussed in NLP, it's ambiguous:

> *Our goal is not to define or prescribe what 'helpful' and 'harmless' mean …, so for the most part we simply let our crowdworkers interpret these concepts* [1]

Correctness ⇏ Helpfulness



Our definition:

> *A question answering system is **helpful** if it provides responses that enable users to maximally achieve their goals*

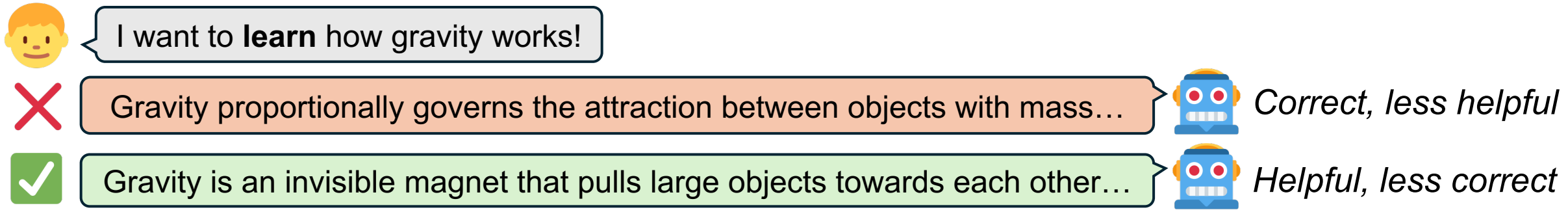
[1] Bai, et al. (2022). *Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback*

What do we mean by helpfulness?

When helpfulness is discussed in NLP, it's ambiguous:

> *Our goal is not to define or prescribe what 'helpful' and 'harmless' mean …, so for the most part we simply let our crowdworkers interpret these concepts* [1]

Correctness ⇏ Helpfulness



Our definition:

> *A question answering system is **helpful** if it provides responses that enable users to maximally achieve their goals, as intended by the system designer* ←

[1] Bai, et al. (2022). *Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback*

What goals did prior QA systems help users achieve?

1) Domain-Specific QA

User Goal: Help experts query DBs

What are the plag analyses for breccias?

LUNAR

Development: Expert approval

What are the plag analyses for breccias?

SELECT plagioclase FROM breccias

LGTM!

User Goal

Development

2) Open-Domain QA

User Goal: Help users find info online

Who wrote the novel Don Quixote?

Web search

Development: Correctness

Who wrote the novel Don Quixote?

Document Retrieval

Span Selection

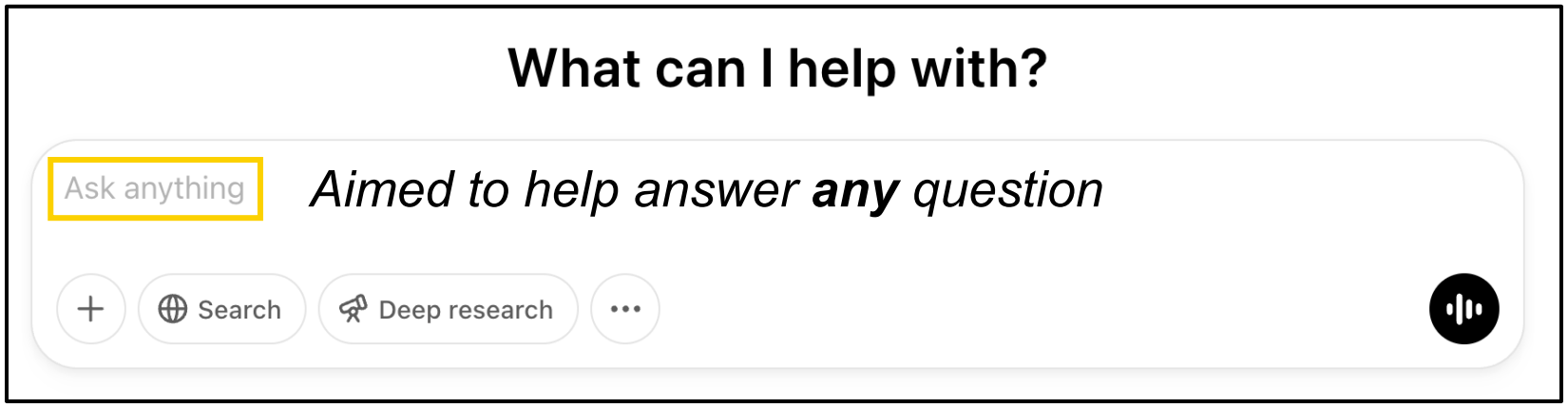
Corpus

Ground Truth Ans.

[1] Progress in natural language understanding

[2] The trec-8 question answering track report

But now, we have Large Language Models



Do ***user goals*** with LLMs match LLM ***development***?

User goals with Large Language Models

Users can now ask LLMs questions to support goals beyond correctness

Goal: Learning

What's a memorable way I can remember the capital of France?

A fun and memorable way to remember that **Paris** is the capital of France is:

🔷 **"A Pair (Paris) of Fancy Pants (France)"**

Goal: Factual Verifiability

What's the capital of France? I need to verify this with official sources

The capital of France is **Paris.**

To verify this with an official source, you can refer to:

- The French government's official website: www.gouvernement.fr
- The CIA World Factbook: France Country Profile
- United Nations data: data.un.org

Users don't just expect correct answers, but reasoning chains* to help them

*Disclaimer: "reasoning" does not imply true reasoning/faithfulness, but the utility of these generations

Development with Large Language Models

LLMs are developed to be strong text generation systems

Pre-Training

Learn to mimic human-written texts (next-token prediction)

Post-Training

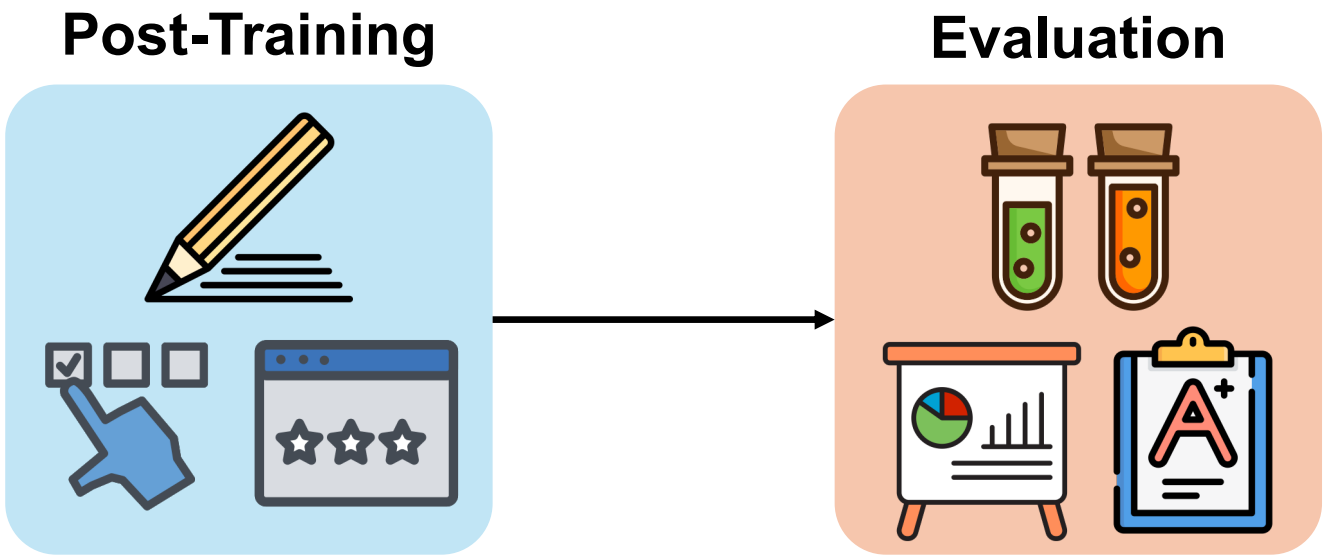
Fine-tuned to be safe, follow instructions, use a certain style, …

Evaluation

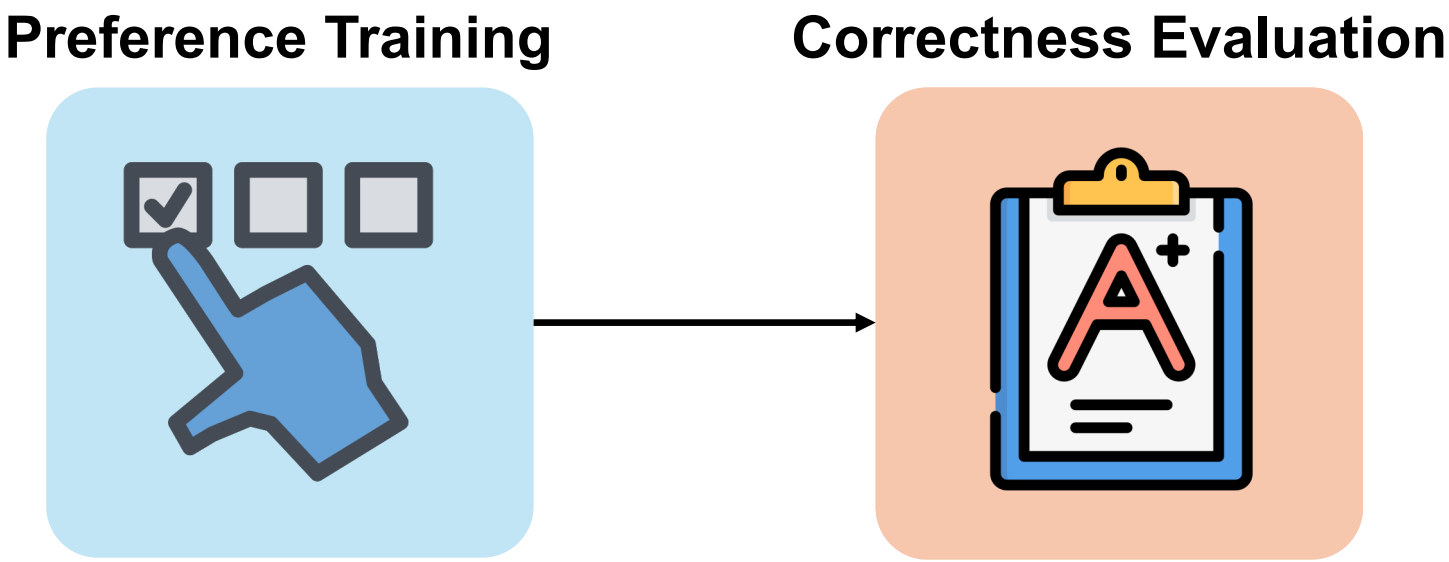
Check the model's knowledge, ability to help users, …

Is LLM development aligned with helping users?

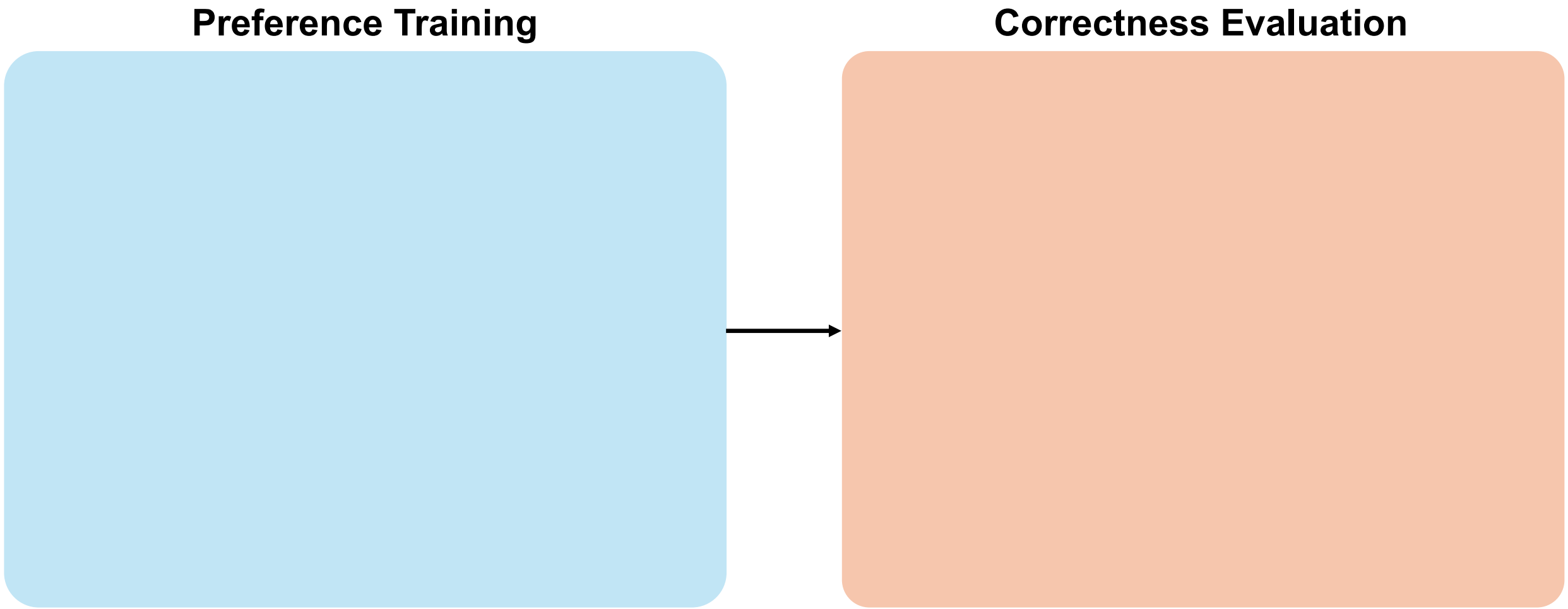
Is LLM development aligned with helping users?



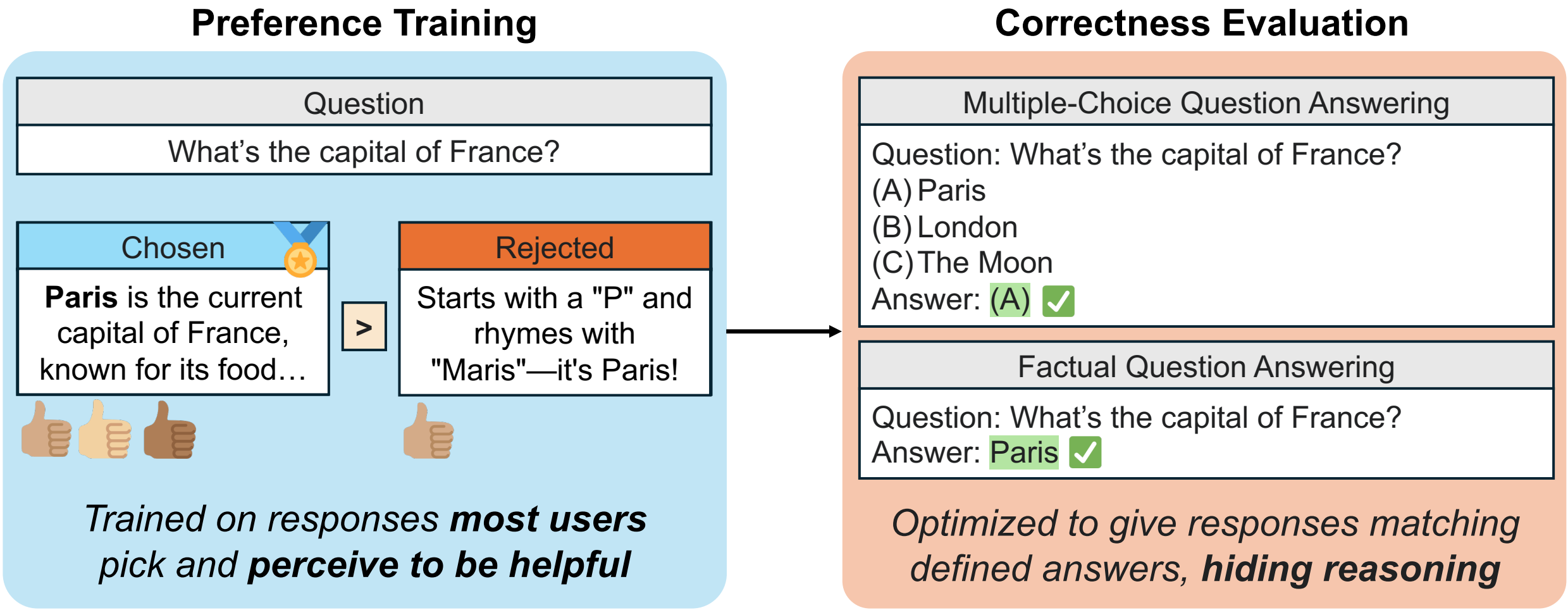
Is LLM development aligned with helping users?



Is LLM development aligned with helping users?



Is LLM development aligned with helping users?



Is LLM development aligned with helping users? **No**

Preference Training

Question
What's the capital of France?

Chosen	>	Rejected
Paris is the current capital of France, known for its food…		Starts with a "P" and rhymes with "Maris"—it's Paris!

*Trained on responses **most users** pick and **perceive to be helpful***

Correctness Evaluation

Multiple-Choice Question Answering

Question: What's the capital of France?
(A) Paris
(B) London
(C) The Moon
Answer: (A) ✅

Factual Question Answering

Question: What's the capital of France?
Answer: Paris ✅

*Optimized to give responses matching defined answers, **hiding reasoning***

⚠ Poor proxies for true helpfulness ⚠

Is LLM development aligned with helping users? **No**

Part I: Correctness Evaluation

Multiple-Choice Question Answering

Question: What's the capital of France?
(A) Paris
(B) London
(C) The Moon
Answer: (A) ✅

Factual Question Answering

Question: What's the capital of France?
Answer: Paris ✅

*Optimized to give responses matching defined answers, **hiding reasoning***

Part II: Preference Training

Question

What's the capital of France?

Chosen

Paris is the current capital of France, known for its food…

>

Rejected

Starts with a "P" and rhymes with "Maris"—it's Paris!

*Trained on responses **most users** pick and **perceive to be helpful***

QA weaknesses in evaluations (**Part I**) can inform preference training solutions (**Part II**)!

Is LLM development aligned with helping users? **No**

Part I: Correctness Evaluation

Multiple-Choice Question Answering

Question: What's the capital of France?
(A) Paris
(B) London
(C) The Moon
Answer: (A) ✅

Factual Question Answering

Question: What's the capital of France?
Answer: Paris ✅

Optimized to give responses matching defined answers, ***hiding reasoning***

Part II: Preference Training

Question

What's the capital of France?

Chosen

Paris is the current capital of France, known for its food…

>

Rejected

Starts with a "P" and rhymes with "Maris"—it's Paris!

Trained on responses ***most users*** *pick and* ***perceive to be helpful***

[1] Process of Elimination (ACL 2024, Findings)
[2] Reverse Question Answering (NAACL 2025)

[3] Mnemonic Generation (EMNLP 2024)
[4] Plan Helpfulness for Multi-Step QA (Proposed)
[5] Personalized Preferences in QA (Proposed)

Is LLM development aligned with helping users? **No**

Part I: Correctness Evaluation

Multiple-Choice Question Answering

Question: What's the capital of France?
(A) Paris
(B) London
(C) The Moon
Answer: (A) ✅

Factual Question Answering

Question: What's the capital of France?
Answer: Paris ✅

*Optimized to give responses matching defined answers, **hiding reasoning***

Part II: Preference Training

Question

What's the capital of France?

Chosen

Paris is the current capital of France, known for its food…

>

Rejected

Starts with a "P" and rhymes with "Maris"—it's Paris!

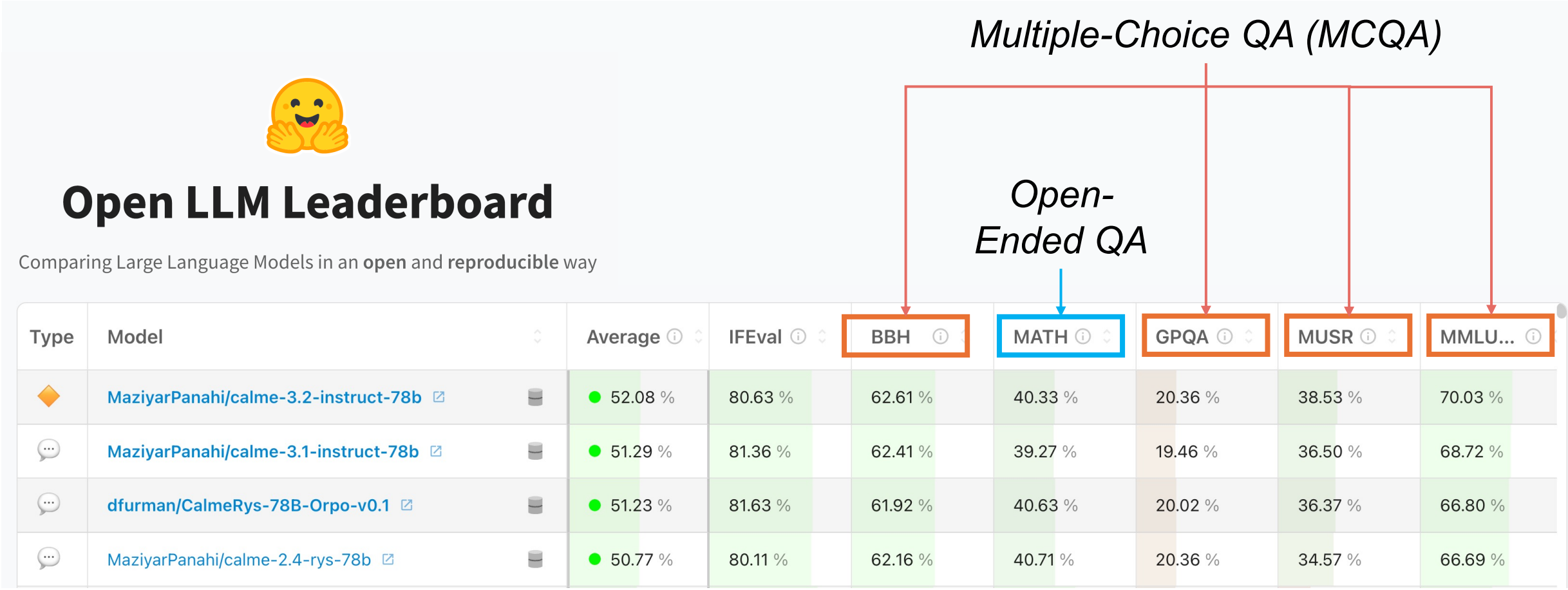
*Trained on responses **most users** pick and **perceive to be helpful***

[1] Process of Elimination (ACL 2024, Findings)
[2] Reverse Question Answering (NAACL 2025)

[3] Mnemonic Generation (EMNLP 2024)
[4] Plan Helpfulness for Multi-Step QA (Proposed)
[5] Personalized Preferences in QA (Proposed)

Correctness-based QA is standard for LLM evaluations

Simple, mirrors human testing, claimed to measure knowledge + reasoning



Type	Model	Average	IFEval	BBH	MATH	GPQA	MUSR	MMLU...
🔶	MaziyarPanahi/calme-3.2-instruct-78b	52.08 %	80.63 %	62.61 %	40.33 %	20.36 %	38.53 %	70.03 %
💬	MaziyarPanahi/calme-3.1-instruct-78b	51.29 %	81.36 %	62.41 %	39.27 %	19.46 %	36.50 %	68.72 %
💬	dfurman/CalmeRys-78B-Orpo-v0.1	51.23 %	81.63 %	61.92 %	40.63 %	20.02 %	36.37 %	66.80 %
💬	MaziyarPanahi/calme-2.4-rys-78b	50.77 %	80.11 %	62.16 %	40.71 %	20.36 %	34.57 %	66.69 %

MCQA Task Format

Given a question and set of choices, LLMs generate the letter of the correct answer

Direct Answer
Question: What's the capital of France? (A) London (B) Paris Answer: (B)

MCQA Task Format

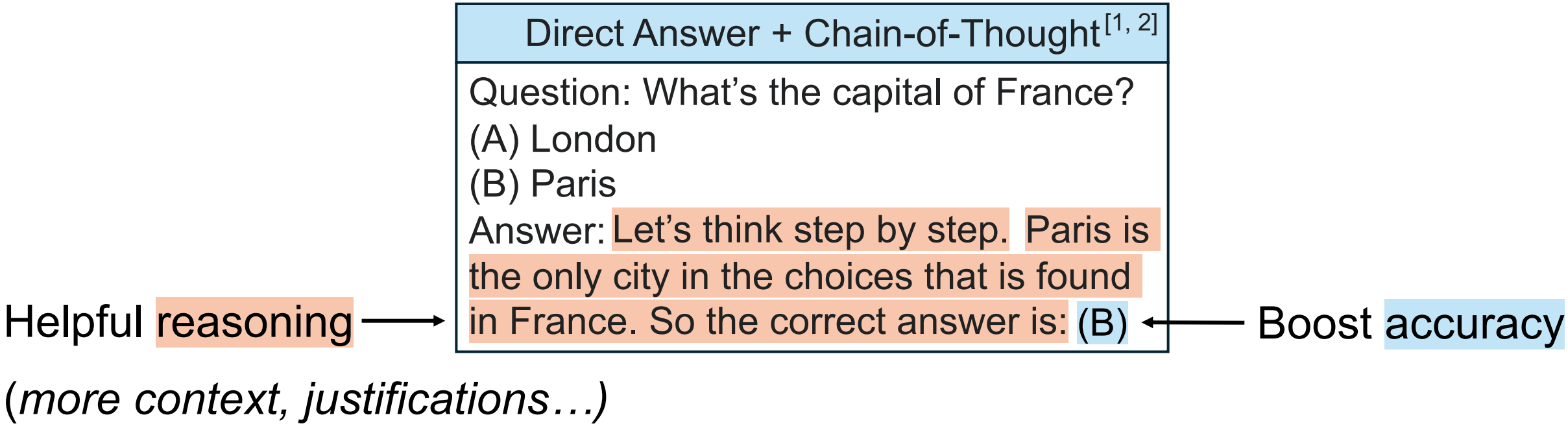
Given a question and set of choices, LLMs generate the letter of the correct answer

Direct Answer + Chain-of-Thought[1, 2]
Question: What's the capital of France? (A) London (B) Paris Answer: Let's think step by step. (B)

[1] Wei et. al., Chain-of-Thought Prompting Elicits Reasoning in Large Language Model (2022)
[2] Kojima et. al., Large Language Models are Zero-Shot Reasoners (2022)

MCQA Task Format

Given a question and set of choices, LLMs generate the letter of the correct answer



[1] Wei et. al., Chain-of-Thought Prompting Elicits Reasoning in Large Language Model (2022)

[2] Kojima et. al., Large Language Models are Zero-Shot Reasoners (2022)

A New MCQA Task Format

Given a question and set of choices, LLMs generate the letter of the ~~correct~~ incorrect answer

Direct Answer + Chain-of-Thought
Question: What's the capital of France? (A) London (B) Paris Answer: Let's think step by step. Paris is the only city in the choices that is found in France. So the correct answer is: (B)

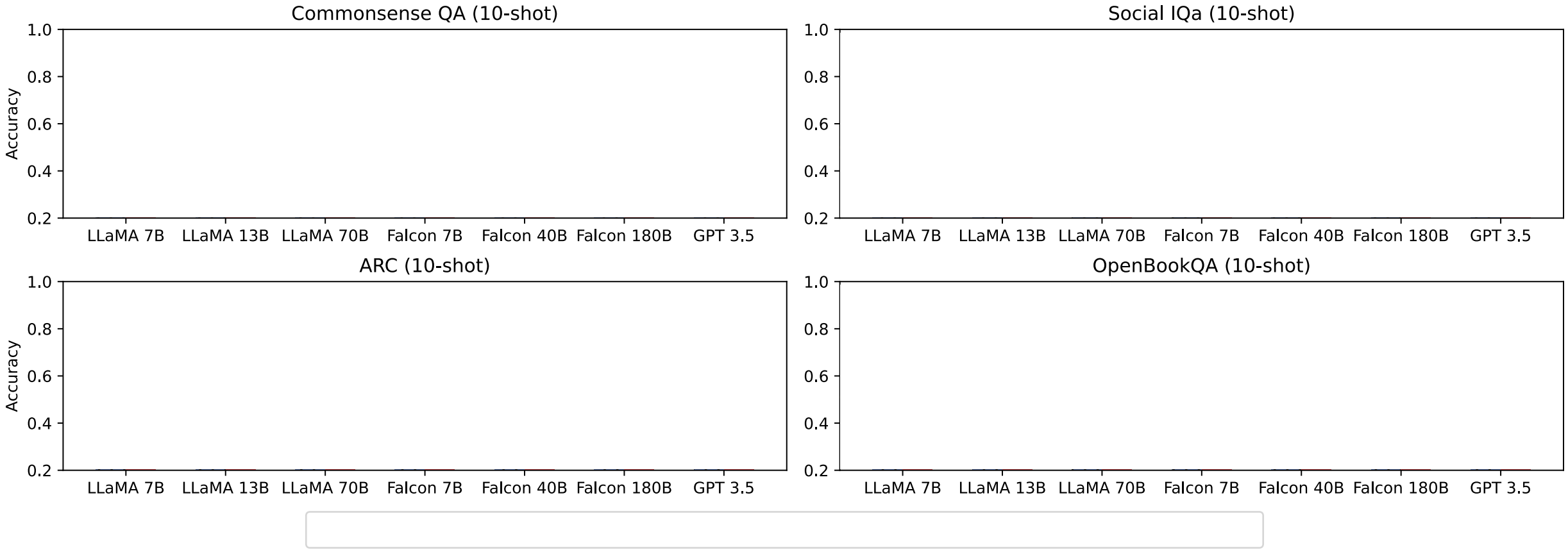
Process of Elimination + Chain-of-Thought
Question: What's the capital of France? (A) London (B) Paris Answer: Let's think step by step.

Goal: Can LLMs adapt their reasoning to explain why choices are wrong?

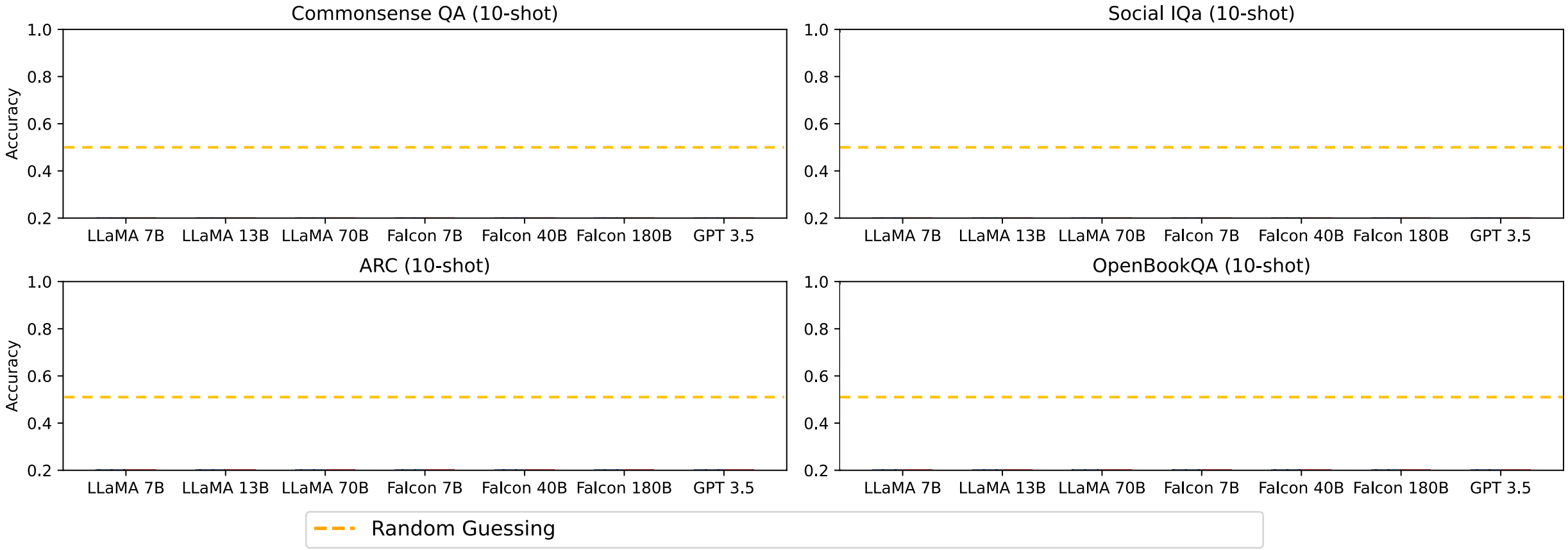
- Signal for personalization, clarify student errors, …

Implementation: LLM prompts have instructions for each task + 10 demonstrations

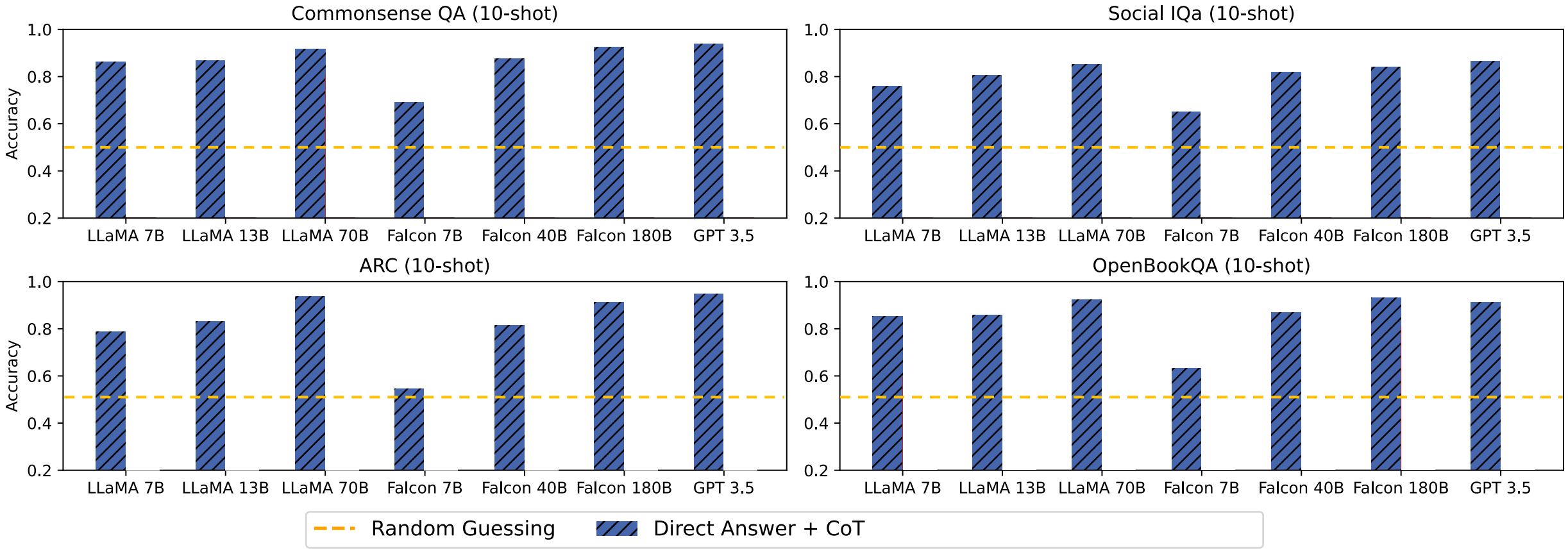
LLM Accuracy with Direct Answer vs PoE



LLM Accuracy with Direct Answer vs PoE on **two-choice MCQs**

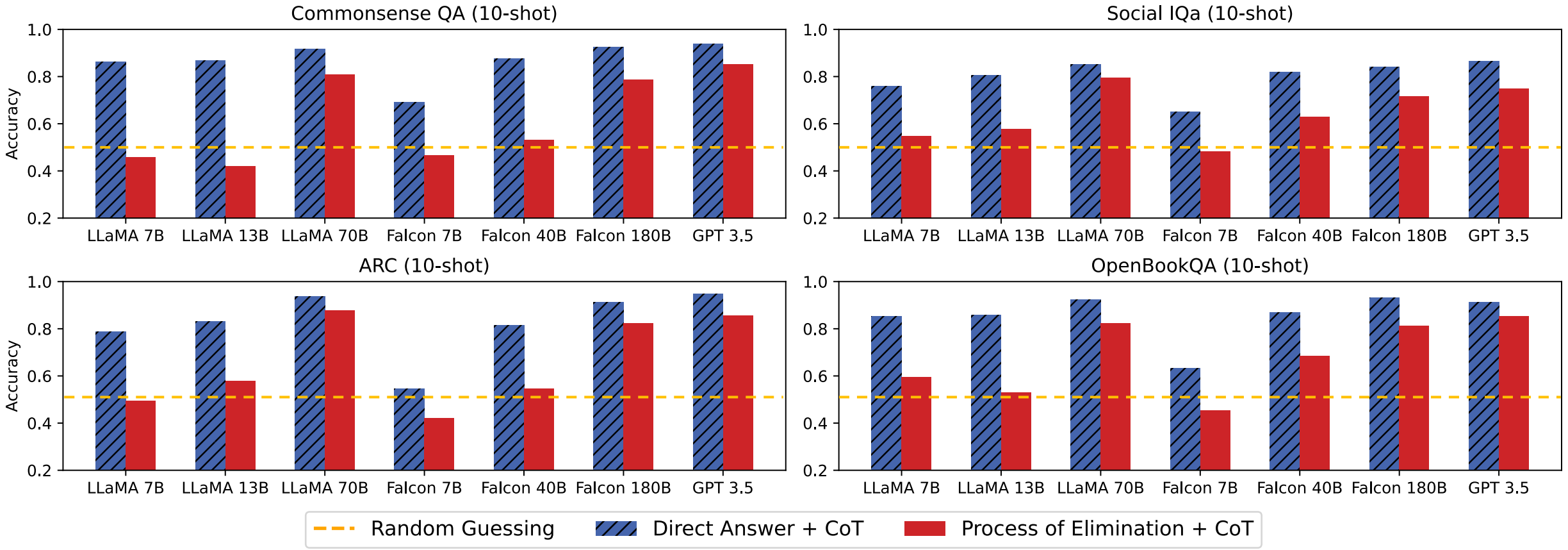


LLM Accuracy with Direct Answer vs PoE on **two-choice MCQs**



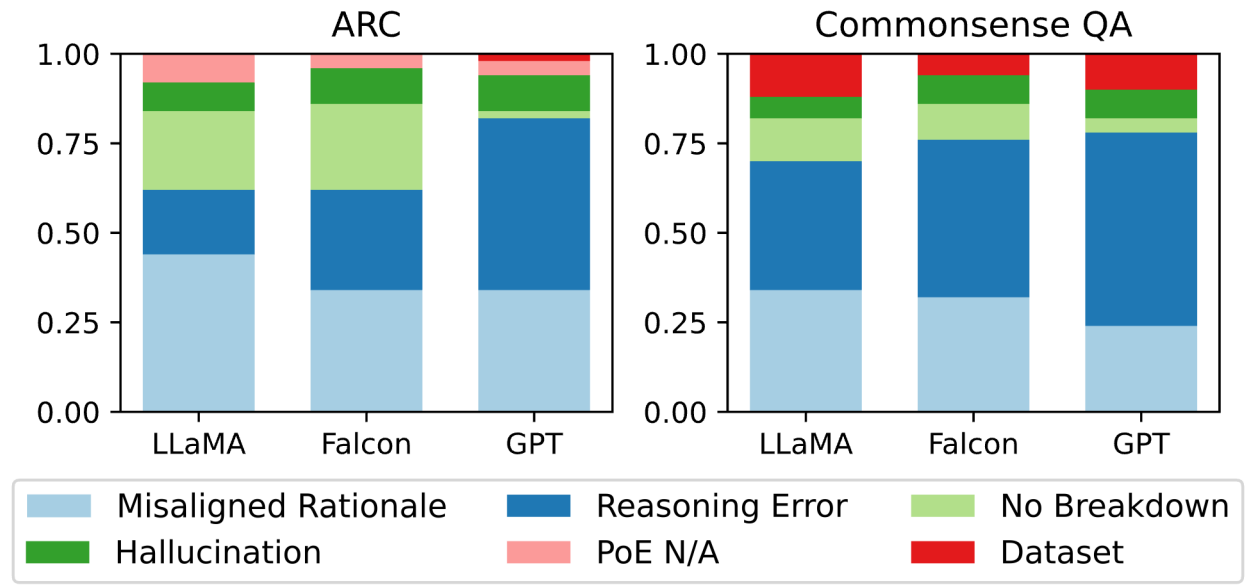
- LLMs can reason why answers are right…

LLM Accuracy with Direct Answer vs PoE on **two-choice MCQs**

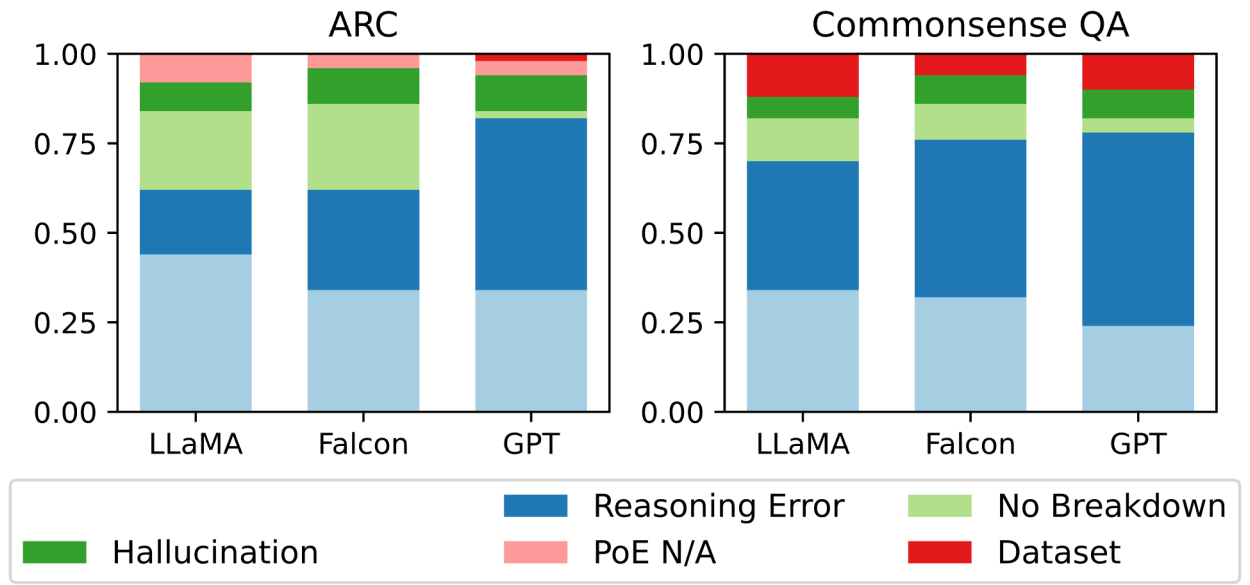


➢ LLMs can reason why answers are right… but can't reason why alternatives are wrong! 😬

Why does Process of Elimination Fail?



Why does Process of Elimination Fail?

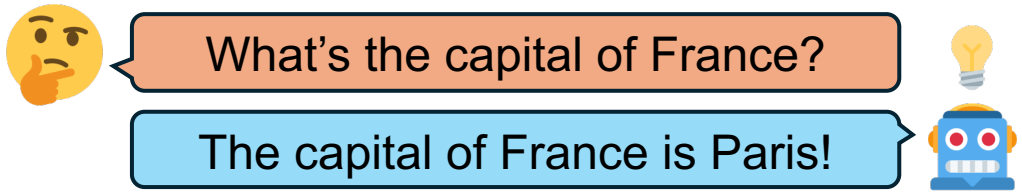


LLMs overfit to one reasoning type without adaptability…

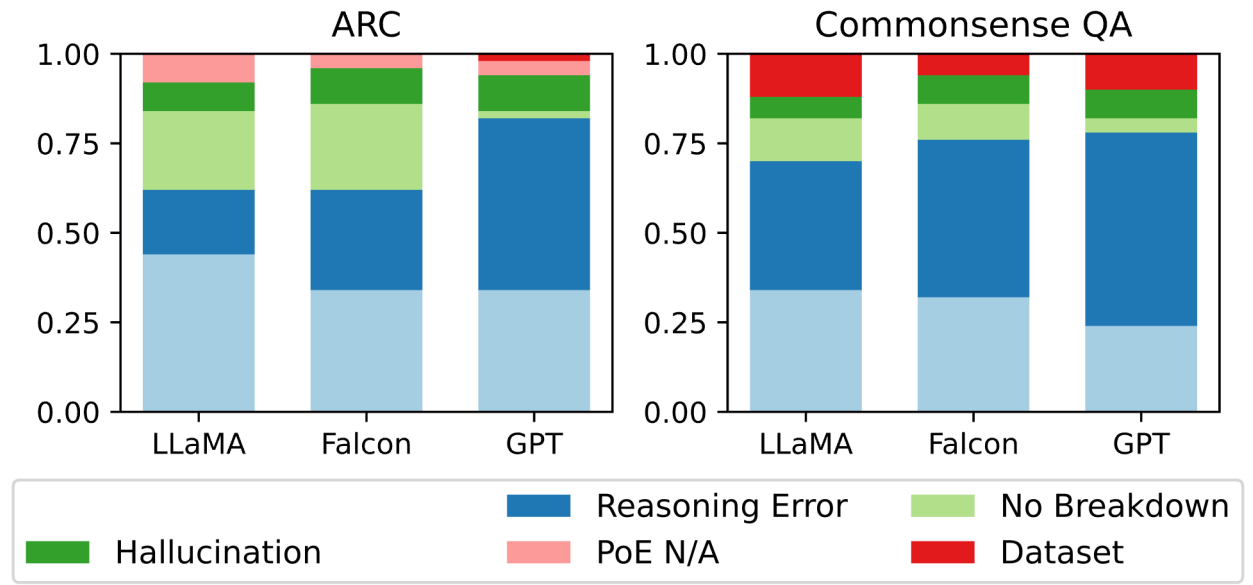
Misaligned Rationale

Which molecule does not have a carbon-nitrogen bond?
(A) nucleic acid (B) carbohydrate

...so can they personalize to user needs?



Why does Process of Elimination Fail?

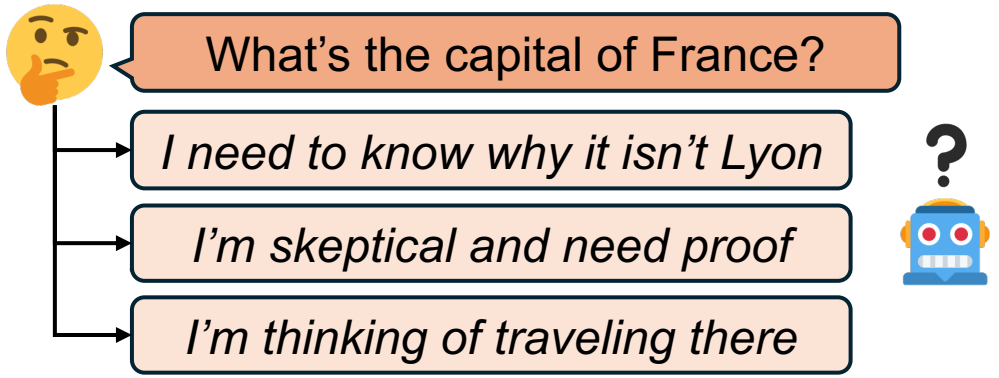


LLMs overfit to one reasoning type without adaptability…

Misaligned Rationale

Which molecule does not have a carbon-nitrogen bond?
(A) nucleic acid (B) carbohydrate
… There is no nitrogen in carbohydrates. So the incorrect answer is “carbohydrate” which is choice (B)

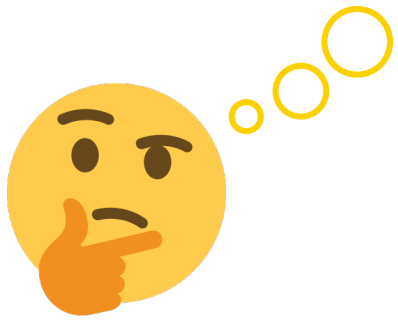
...so can they personalize to user needs?



What about reasoning in open-ended QA (no choices)?

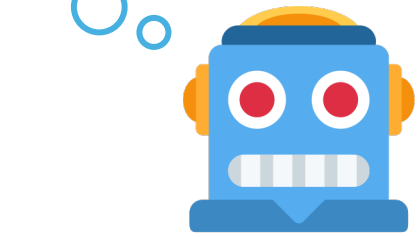
What about reasoning in open-ended QA (no choices)?

Question Answering
Question: How many years apart were the parts of *Don Quixote* published? Answer: 10 years



This is sometimes framed as **deductive reasoning:**

Reaching ***the*** output conclusion (answer)
based on input premises (question)



But what about other reasoning types?

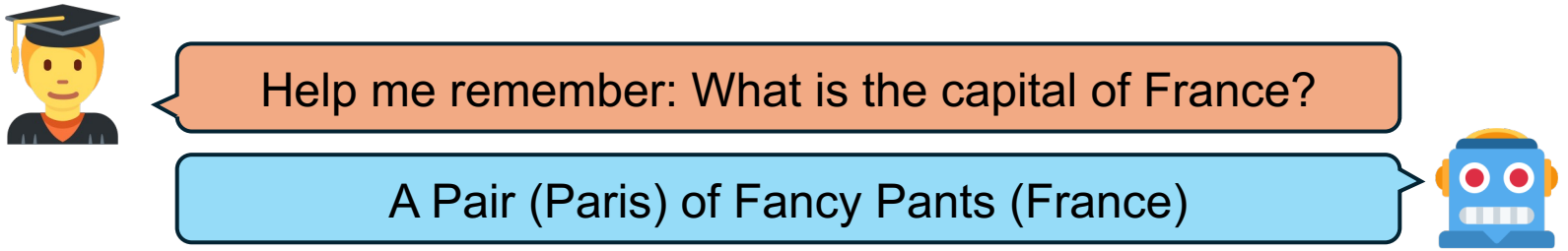
Deductive	Inductive	Abductive
Deriving conclusions based on premises	Generalizing from previous observations	Giving the best explanation for a given outcome
How many years apart were the parts of Don Quixote published?	*Does Don Quixote think all large structures are giants?*	*Why would Sancho ever be friends with Don Quixote?*

But what about other reasoning types?

Abductive

Giving the best explanation for a given outcome *by reasoning over **many** possible explanations*

← Untested in QA w/ only one right answer, but important!



Giving the most helpful response for a query…

But what about other reasoning types?

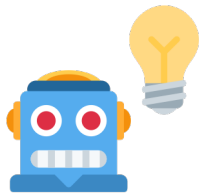
Abductive

Giving the best explanation for a given outcome
*by reasoning over **many** possible explanations*

Untested in QA w/ only one right answer, but important!

Help me remember: What is the capital of France?

A Pair (Paris) of Fancy Pants (France)



Imagine a Parrot (Paris) flying over the Eiffel Tower

How don't you know? We've used this example 5 times

Giving the most helpful response for a query…
*by reasoning over **many** possible responses*

How can we test abduction in question answering?

Question Answering
Question: How many years apart were the two parts of *Don Quixote* published? Answer: 10 years

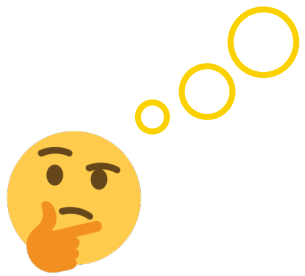
How can we test abduction in ***reverse*** question answering?

Question Answering
Question: How many years apart were the two parts of *Don Quixote* published? Answer: 10 years

*Accurate if you **deduce** <u>the</u> correct answer*

Reverse Question Answering
Task: Give me a question with the answer "10 years" Question: How many years did the Trojan War last?

*Accurate if you **adduce** <u>any</u> valid question*
(via LLM verifier)



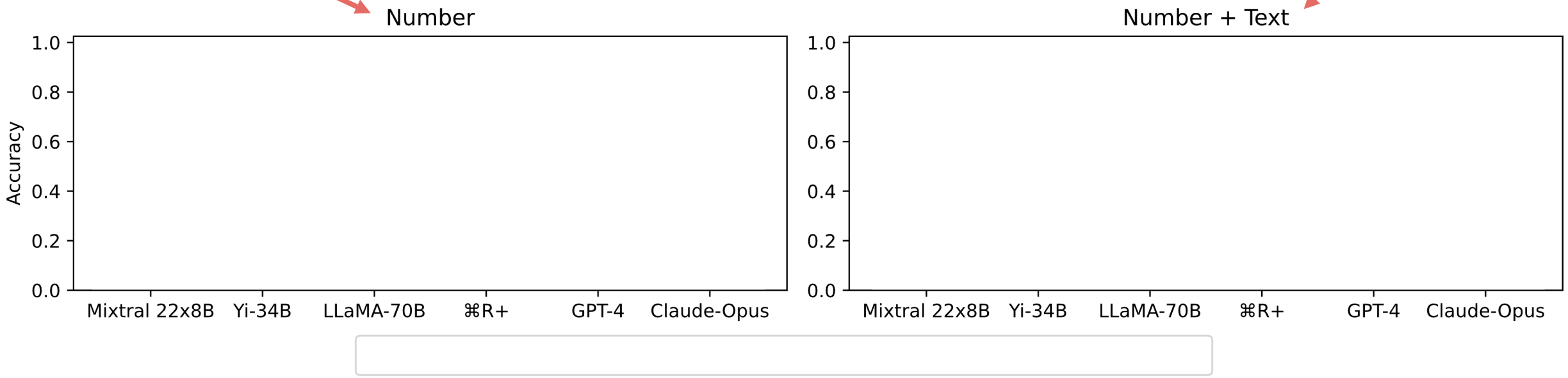
Our Goal: Assess LLM abduction via RQA

(LLM prompts are 0-shot, with formatting constraints)

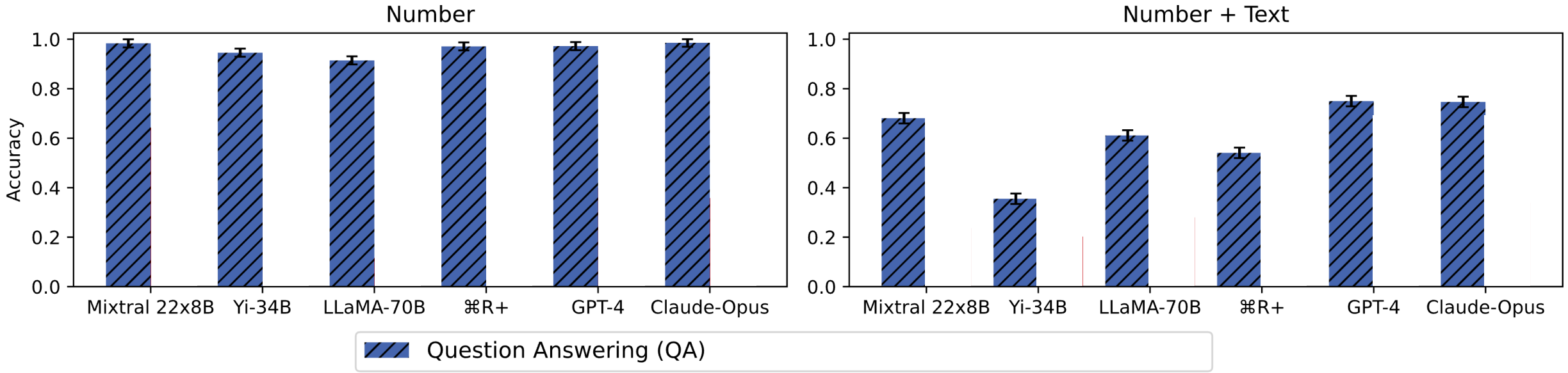
Are LLMs accurate question generators?

(e.g. What is 12 times 12? 144)

(e.g. How long is a century? 100 years)

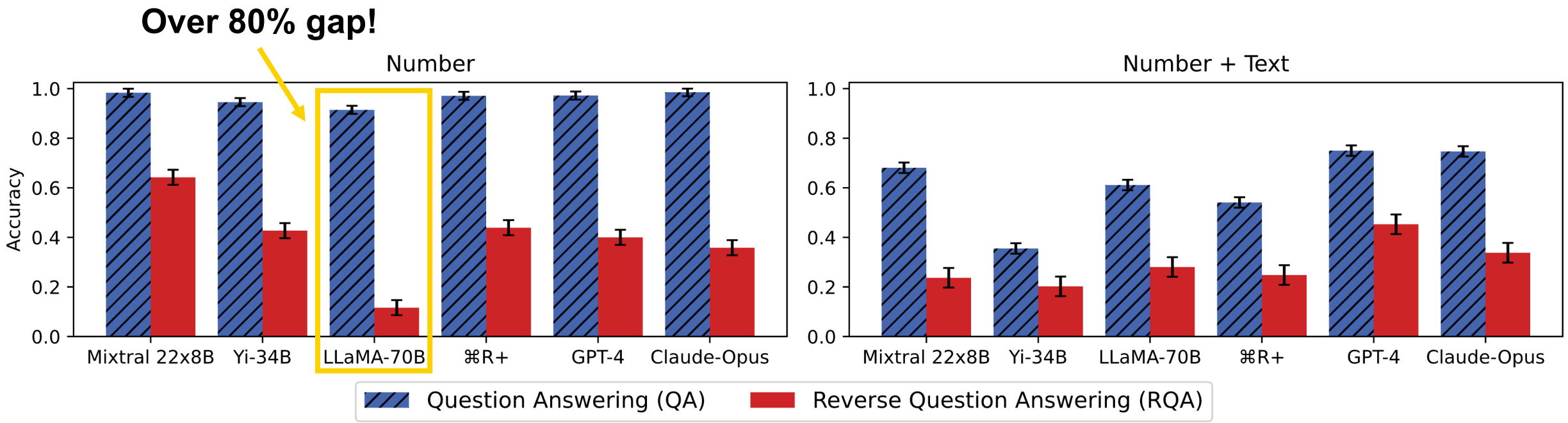


Are LLMs accurate question generators?



- LLMs appear fairly accurate in QA/deduction

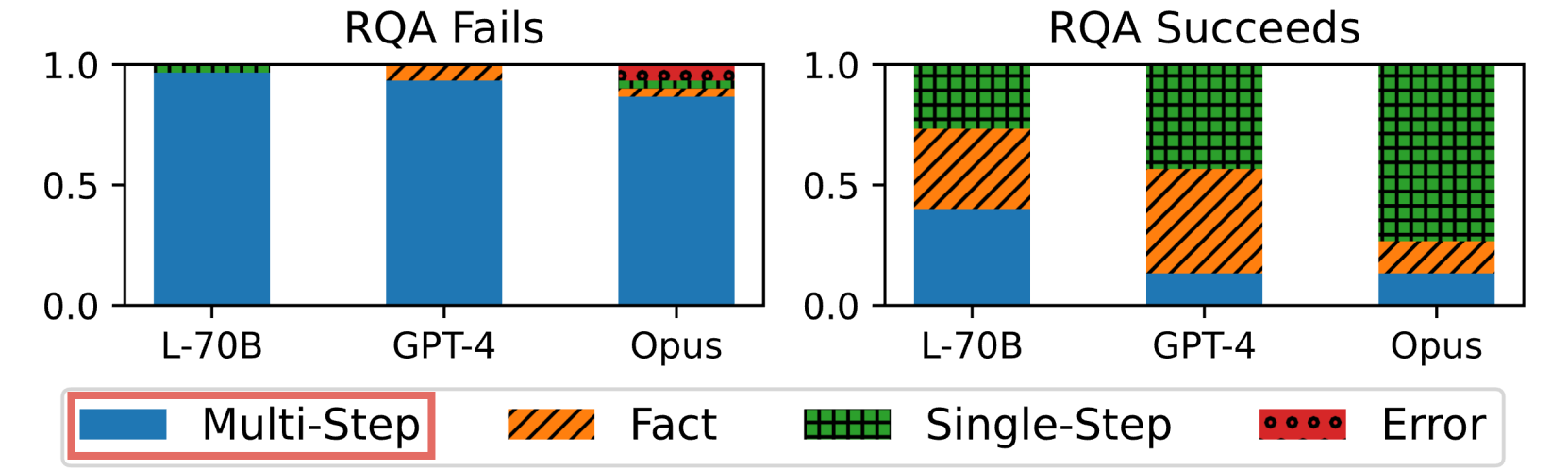
Are LLMs accurate question generators?



- LLMs appear fairly accurate in QA/deduction
- But are significantly weaker at numerical RQA/abduction!

When might RQA fail?

We analyze questions for numbers when RQA fails and categorize them:



Generate a question for “437”
Question: What is the sum of the numbers of legs of a group of 23 cats, 12 humans, and 1 spider?

(it’s actually 124)

Generate a question for “756”
Question: What is the sum of the numbers from 1 to 27, inclusive?

(it’s actually 378)

Look complex, but are in fact unhelpful

When might RQA fail?

We speculate: could this be due to preference training?

Looks helpful, but isn't (complexity bias) [1,2]

Prompt
Generate a question for "756"

Response 1
Question: What is the sum of the numbers from 1 to 27, inclusive?

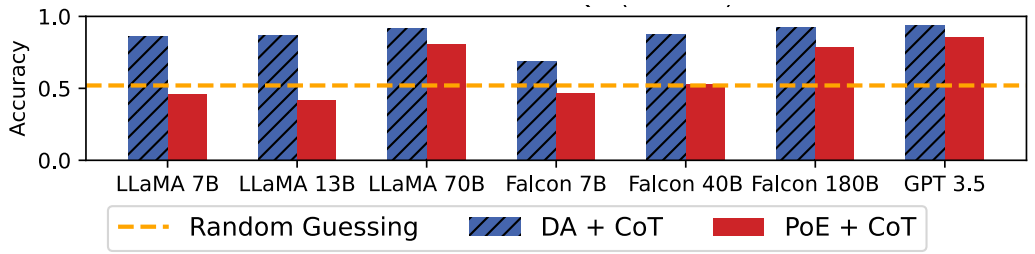
Response 2
Question: What is 755 + 1?

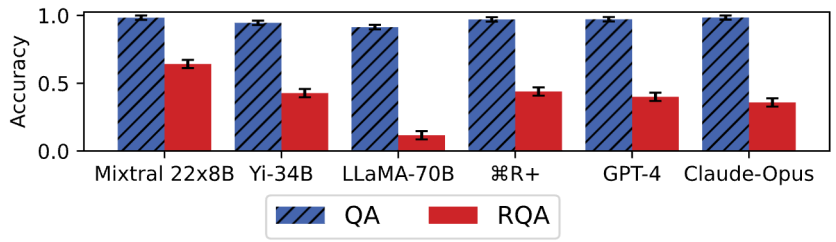
Judgment Biases:

[1] Toward a positive theory of consumer choice
[2] Miswanting: Some problems in the forecasting of future affective states

Conclusion: Answering Questions w/ **Reasoning** that Truly Helps

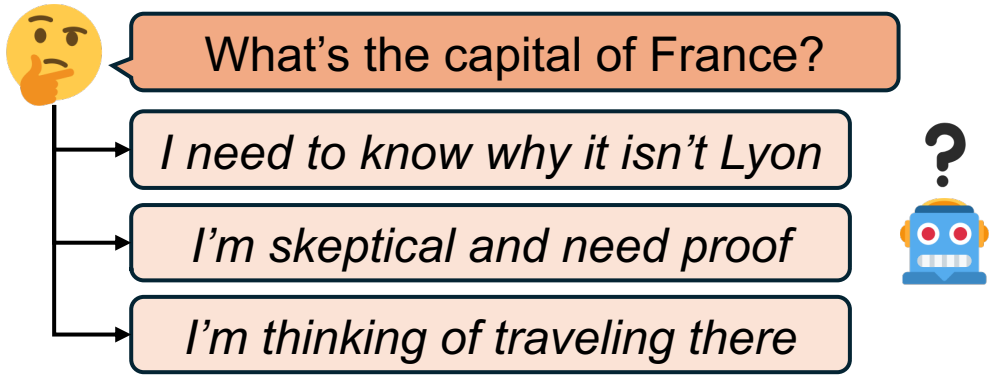
QA correctness does not fully assess the helpfulness of LLM reasoning





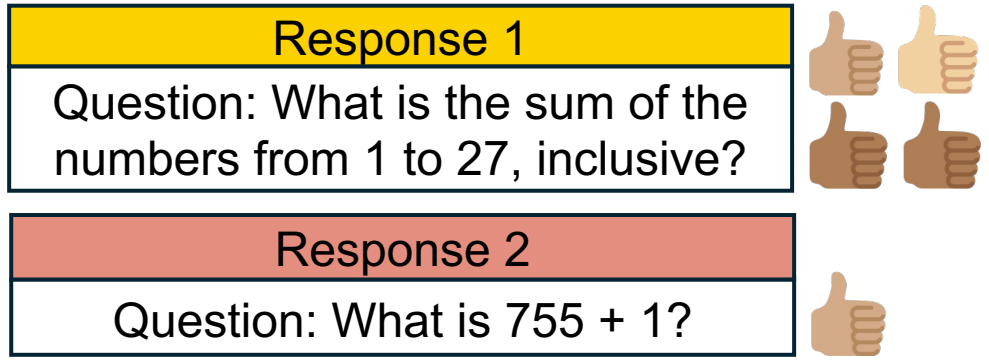
Two insights for developing models:

PoE: LLMs fail to adapt reasoning



*Can LLMs **personalize** to user needs?*

RQA: LLM questions just ***look*** helpful



*Can users predict what is **truly helpful**?*

Is LLM development aligned with helping users? **No**

Part I: Correctness Evaluation

Multiple-Choice Question Answering

Question: What's the capital of France?
(A) Paris
(B) London
(C) The Moon
Answer: (A) ✅

Factual Question Answering

Question: What's the capital of France?
Answer: Paris ✅

*Optimized to give responses matching defined answers, **hiding reasoning***

Part II: Preference Training

Question

What's the capital of France?

Chosen

Paris is the current capital of France, known for its food…

>

Rejected

Starts with a "P" and rhymes with "Maris"—it's Paris!

*Trained on responses **most users** pick and **perceive to be helpful***

[1] Process of Elimination (ACL 2024, Findings)
[2] Reverse Question Answering (NAACL 2025)

[3] Mnemonic Generation (EMNLP 2024)
[4] Plan Helpfulness for Multi-Step QA (Proposed)
[5] Personalized Preferences in QA (Proposed)

Is LLM development aligned with helping users? **No**

Part I: Correctness Evaluation

Multiple-Choice Question Answering

Question: What's the capital of France?
(A) Paris
(B) London
(C) The Moon
Answer: (A) ✓

Factual Question Answering

Question: What's the capital of France?
Answer: Paris ✓

*Optimized to give responses matching defined answers, **hiding reasoning***

Part II: Preference Training

Question

What's the capital of France?

Chosen

Paris is the current capital of France, known for its food…

>

Rejected

Starts with a "P" and rhymes with "Maris"—it's Paris!

*Trained on responses most **users** pick and **perceive to be helpful***

[1] Process of Elimination (ACL 2024, Findings)
[2] Reverse Question Answering (NAACL 2025)

[3] Mnemonic Generation (EMNLP 2024)
[4] Plan Helpfulness for Multi-Step QA (Proposed)
[5] Personalized Preferences in QA (Proposed)

Recall: How can QA systems help users achieve their goals?

Goal: Learn Something New

What does “LLM” mean?

Goal: Solve a Multi-Step Problem

How can I get my refund?

Goal: Receive Tailored Advice

How do I get 5 professors in the same room / time?

Goal: Recall Forgotten Information

Who gave that proposal talk with too many emojis?

Recall: How can QA systems help users achieve their goals?

Vocabulary Learning →

Goal: Learn Something New

What does “LLM” mean?

Students are expected to learn 1000+ vocab terms for the GRE 😴

johnyboyblablablublu • 6y ago •

I would say electronic apps suffice for decent GRE score. I used three apps- magoosh flash cards, magoosh vocabulary, and galvanize (I would highly recommend this, it will also help understand the meaning of words).

Generally, aim for around new 1000-1200 words in your vocabulary. More than that just felt like dumping a lot of words in a lot less time.

5 Reply ···

You may need to know **1,000+** GRE vocab words to be ready for whatever could come your way on test day. Even if you already have a broad vocabulary, you probably will still need to learn at least a few hundred vocab words over the course of your GRE test prep. Dec 29, 2023

TTP GRE Blog
https://gre.blog.targettestprep.com › how-to-learn-voca...

How to Learn Vocabulary for GRE Verbal - TTP GRE Blog

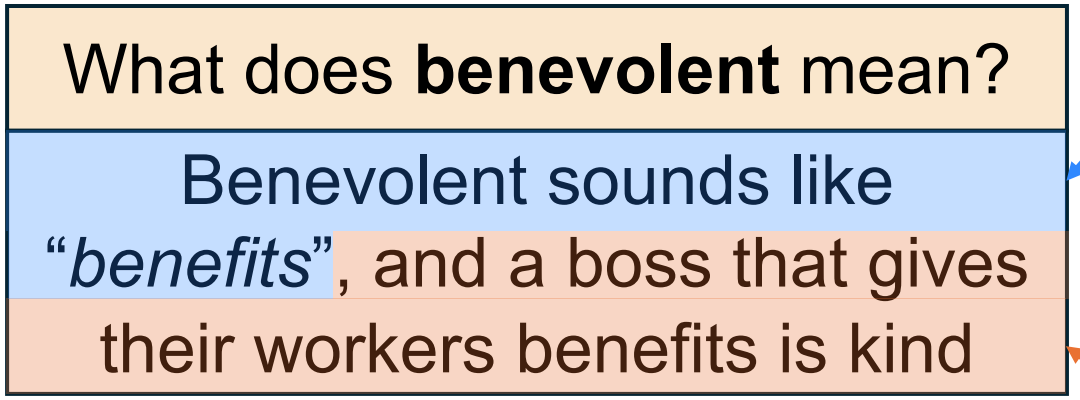
ToomuchLes
Posted August 29, 2013

Hey all!
So.. I've been studying GRE words for nearly 2 months now, and as of right now, I've memorized Kaplan GRE Vocab Flashcards (500), Manhattan Prep GRE 500 Essential Words, and Barrows 2nd Edition GRE Words (500). In total, with synonyms and additional words on the comment section, I probably memorized more than 2,500 GRE Words (minimum), and when I mean 'memorized' I mean I can recite definition, synonyms and everything without a pause. As you can see, Im very proud of such an achievement lol.

How can systems make studying vocab more effective?

*Mnemonics must have **correct** answers (definitions)*

*and **helpful** reasoning (keywords/explanations)*



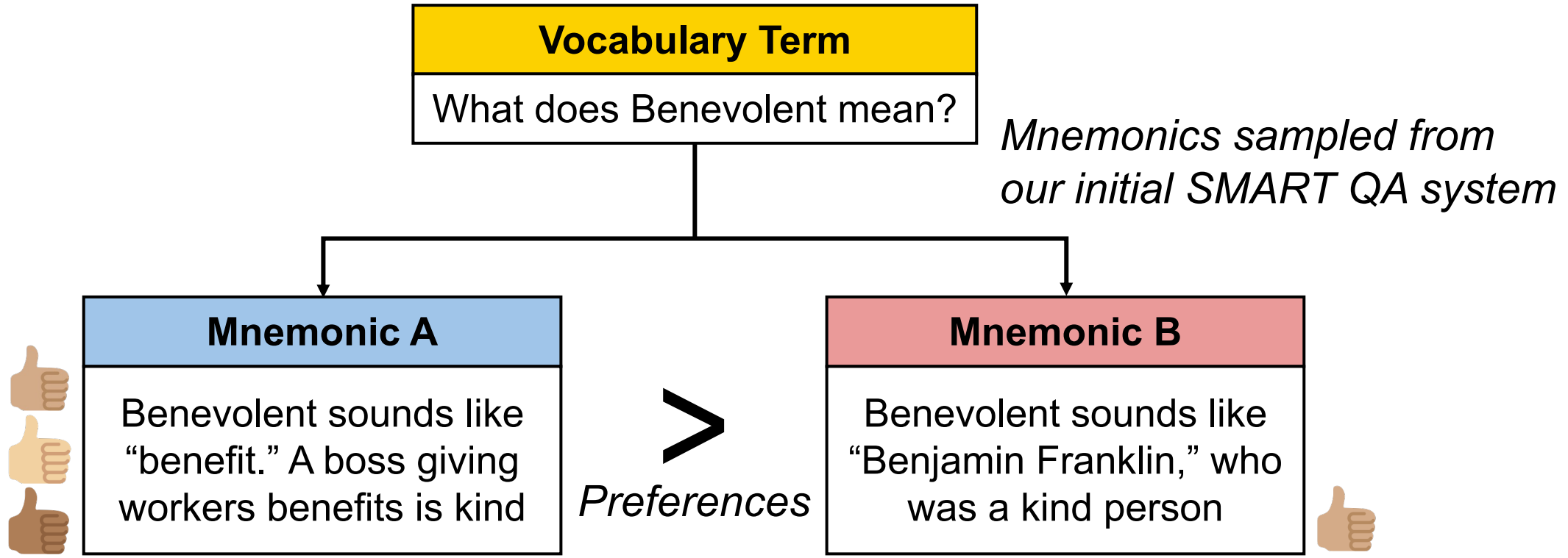
Let's use mnemonic devices!

1) Link term to a **simpler** + **similarly sounding** keyword

2) **Explain the link** between the keyword + term

Goal: Build a mnemonic generator (SMART) to help students learn vocab

Teach SMART which mnemonics are **helpful** via preference training



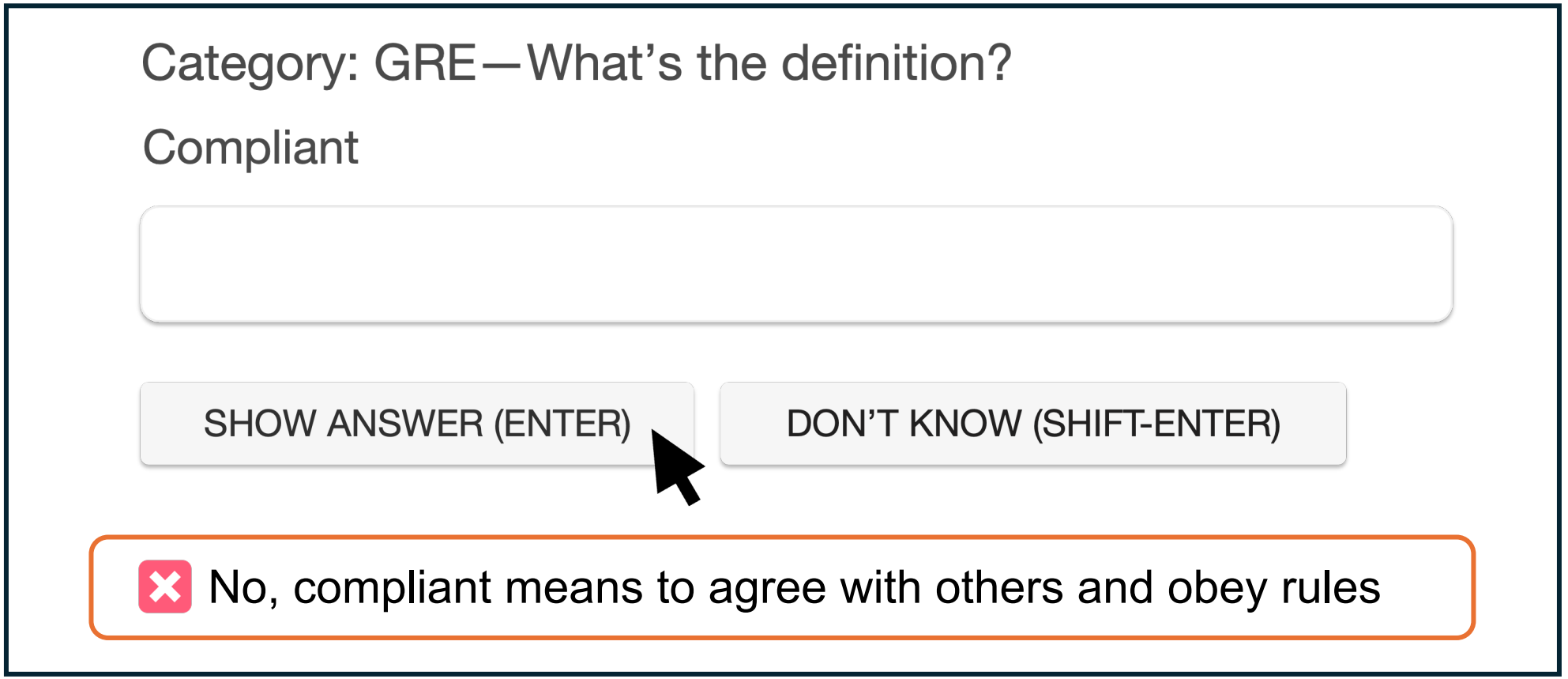
Preference training assumes users can predict what is helpful…
…but RQA leads us to speculate this may not always be the case!

Testing this assumption: collecting feedback on mnemonics

47 learners study with mnemonics from the initial SMART model in a flashcard app:

Testing this assumption: collecting feedback on mnemonics

47 learners study with mnemonics from the initial SMART model in a flashcard app: [1]

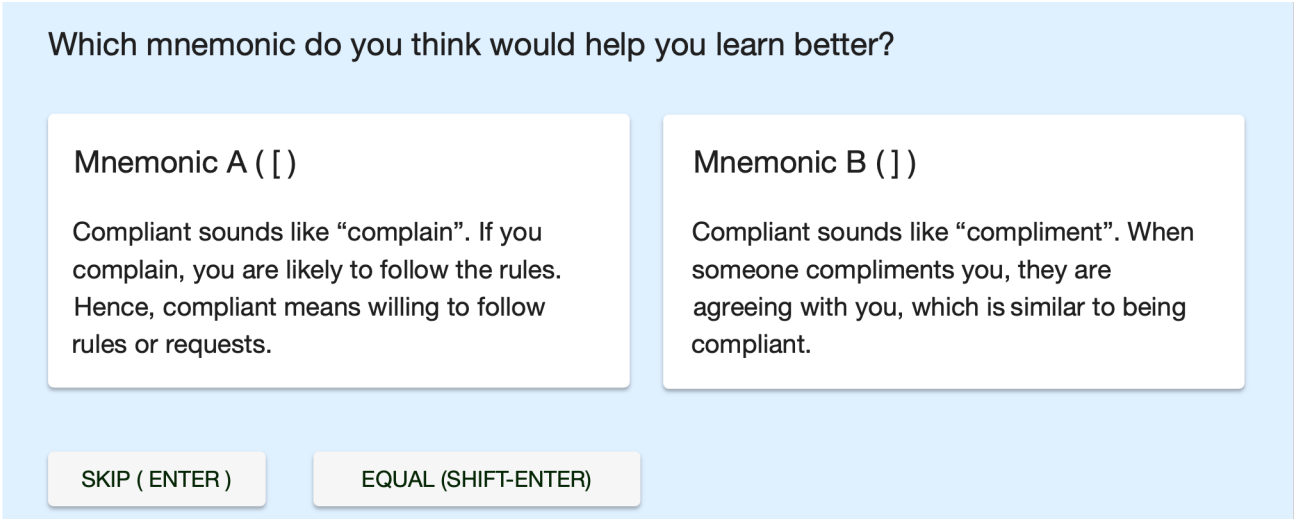


[1] KARL: Knowledge-Aware Retrieval and Representations aid Retention and Learning in Students

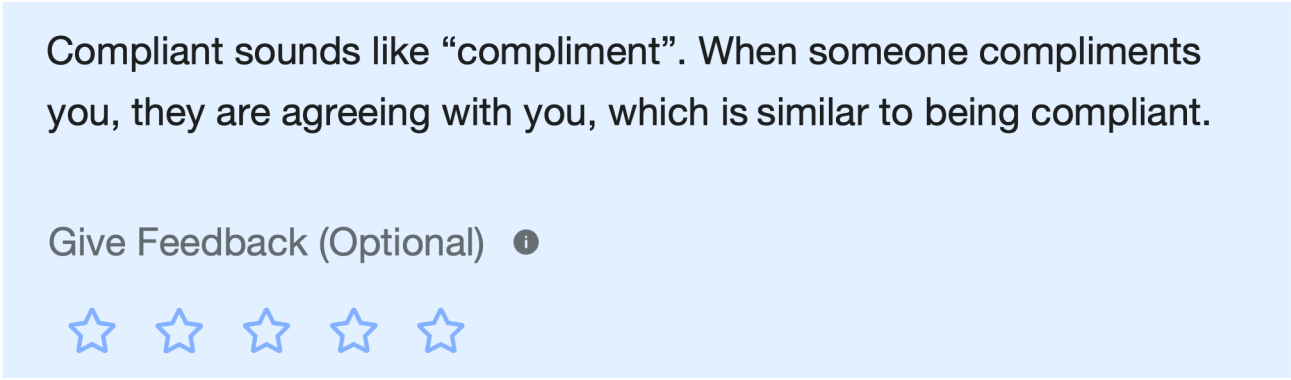
1) **Expressed Preferences**: What users ***think*** will help them

Standard way of collecting preference data

Pairwise Comparisons



Likert Ratings



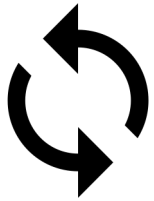
2) **Observed Preferences**: What ***truly*** helps user goals (learning)

Benevolent

Mnemonic

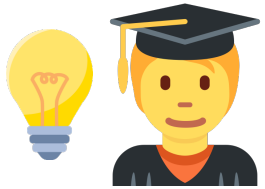
Benevolent sounds like “benefit.” A boss giving workers benefits is kind

Idea: **More helpful** mnemonics need **less iterations** to study



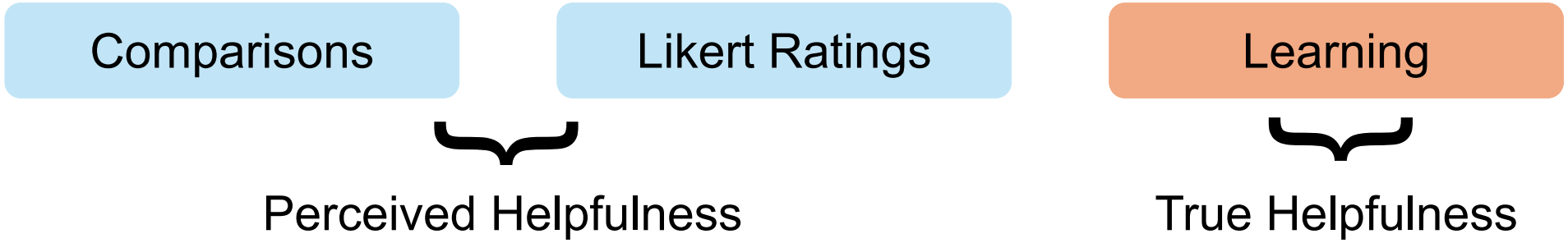
Keep Studying ***N*** *times*

Short-Term Learning *(downstream goal)*



Benevolent means “well meaning and kind”

Our preference types

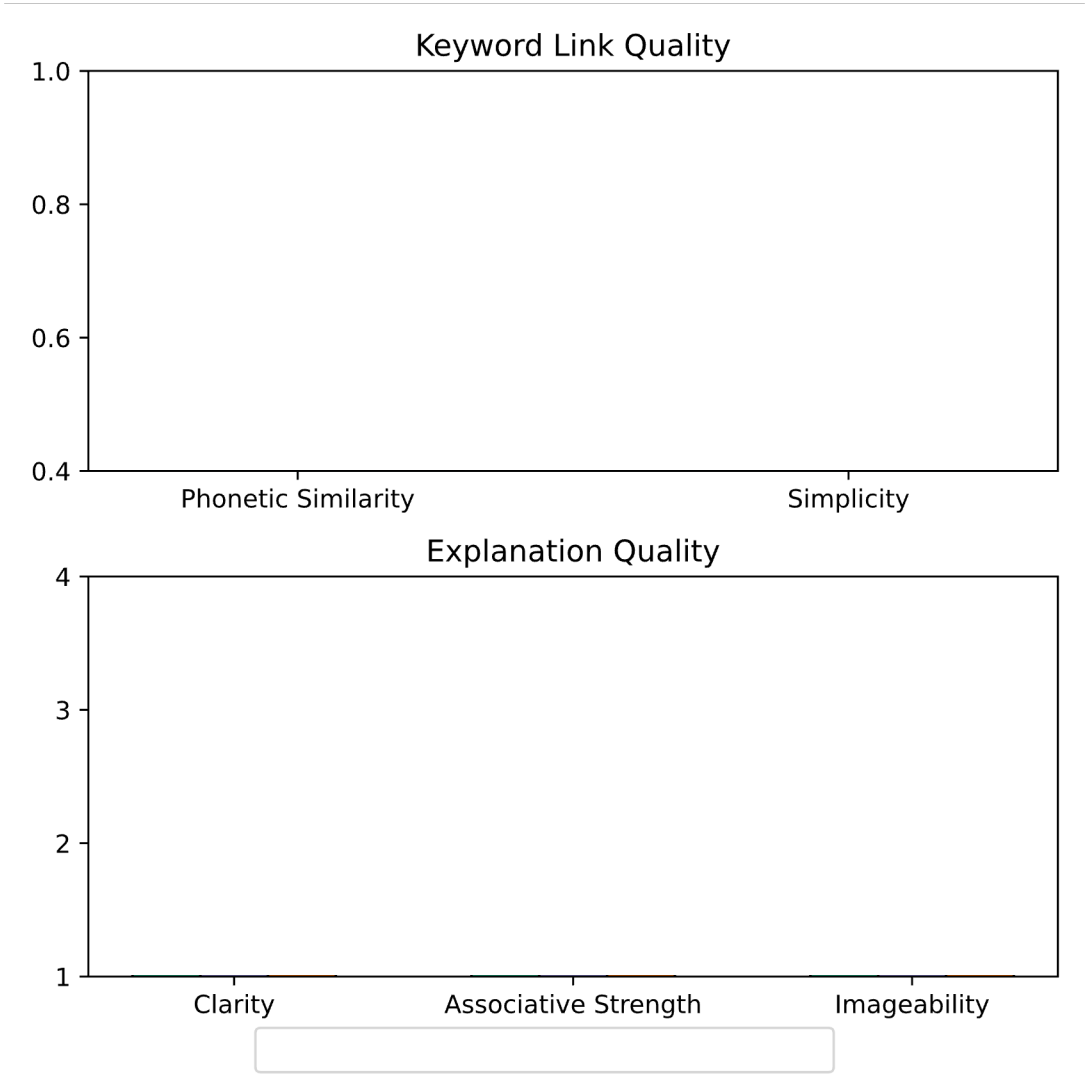


Our preference types have large disagreements!

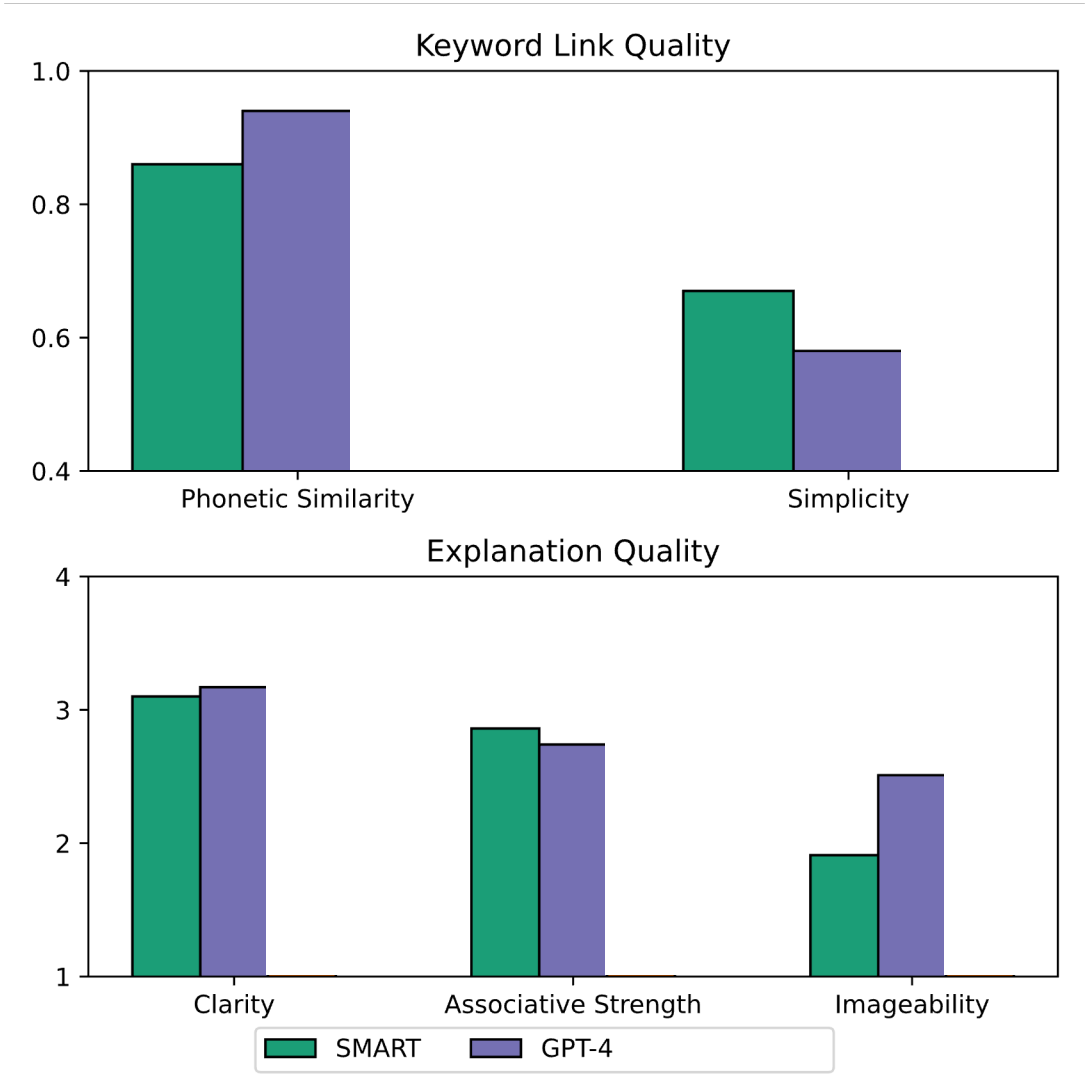
	Comparisons	Likert Ratings	Learning
Comparisons	–	–	–
Likert Ratings	67.5%	–	–
Learning	40.7%	50.7%	–

- Students struggle to predict what actually helps them!

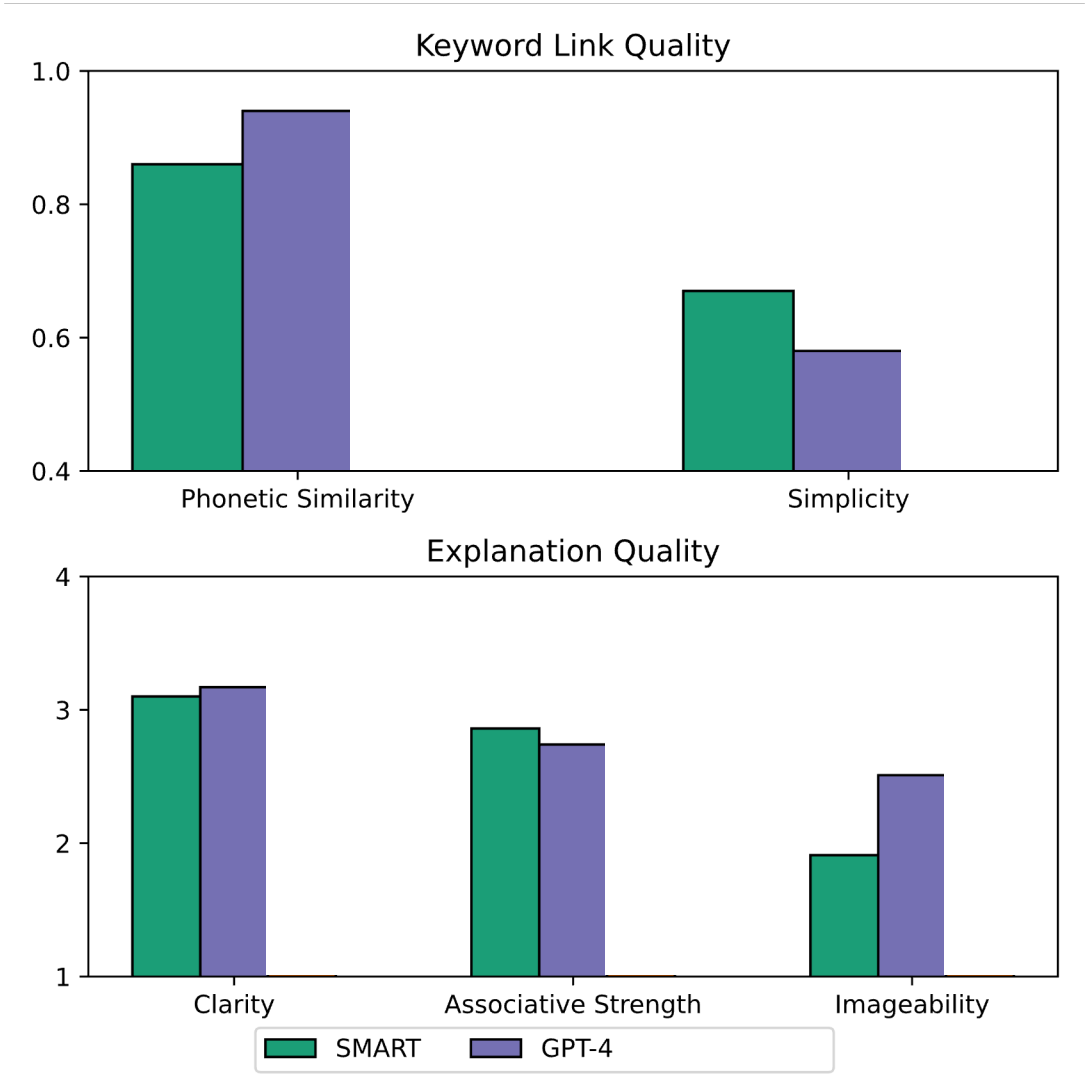
What do two experts (one post-doc, one professor) think?



What do two experts (one post-doc, one professor) think?



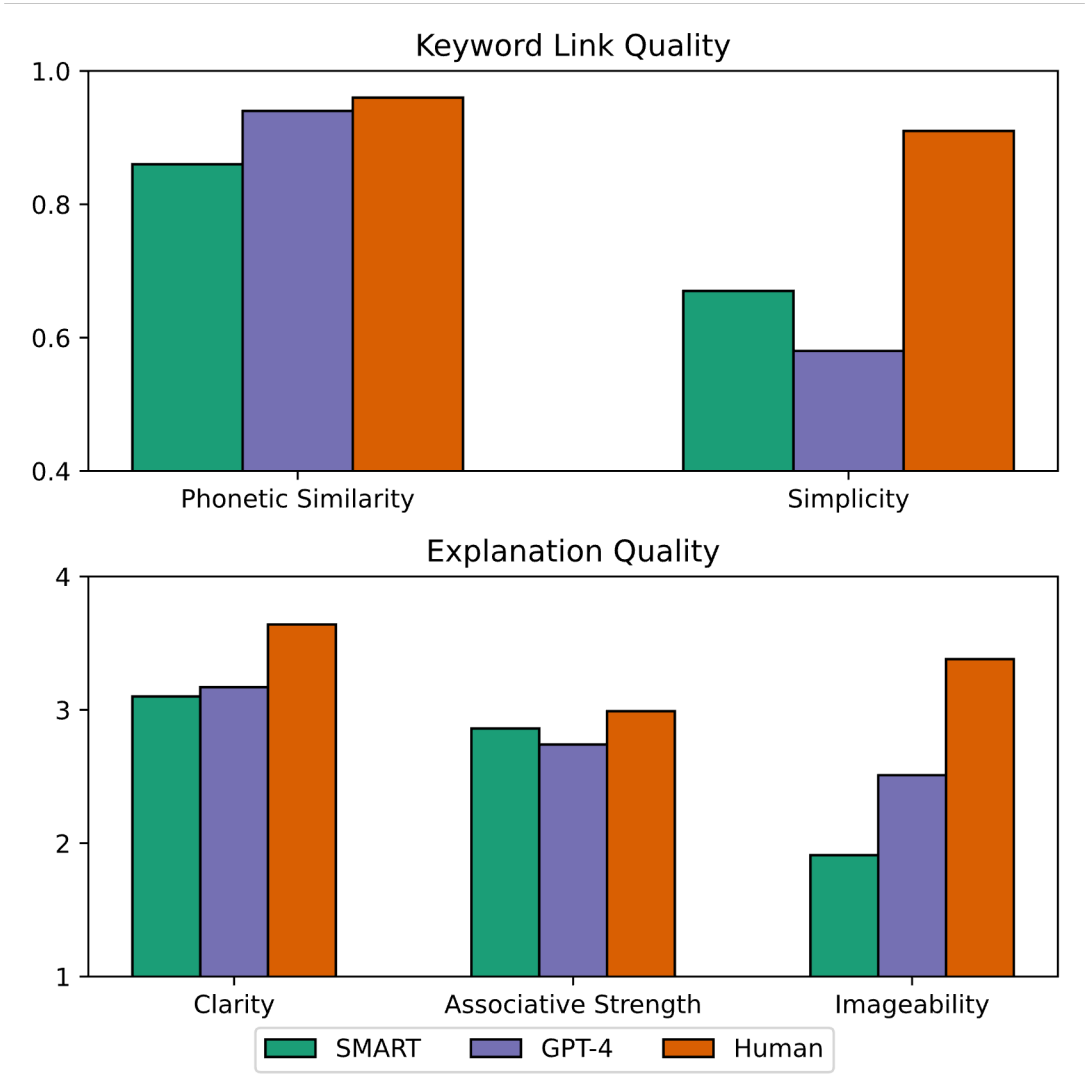
What do two experts (one post-doc, one professor) think?



1) **SMART** matches **GPT-4**!

SMART can match the helpfulness of a much larger system!

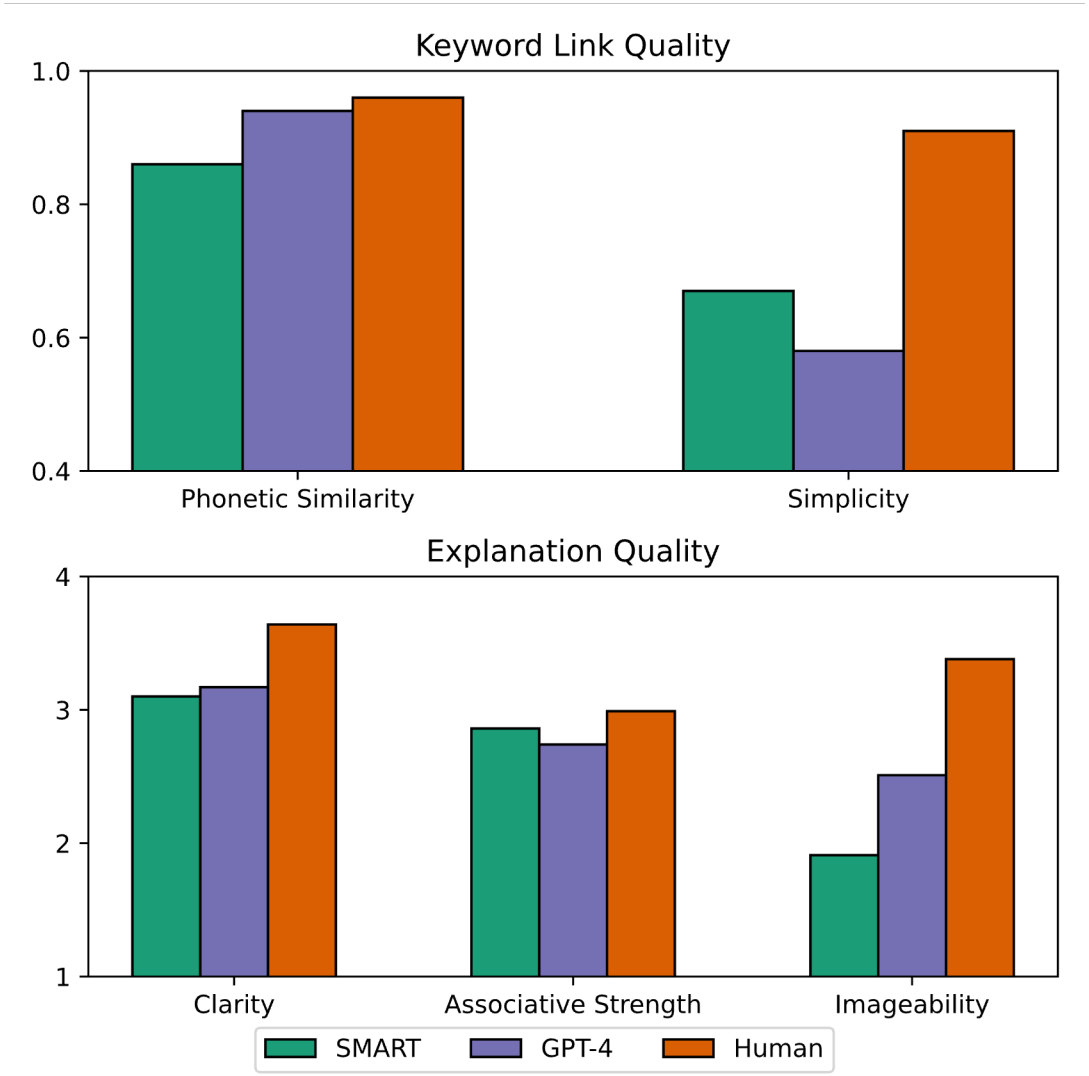
What do two experts (one post-doc, one professor) think?



1) **SMART** matches **GPT-4**!

SMART can match the helpfulness of a much larger system!

What do two experts (one post-doc, one professor) think?



1) **SMART** matches **GPT-4**!

SMART can match the helpfulness of a much larger system!

2) Our **human writer** is much better than **SMART** and **GPT-4**!

We have a long way to go!

Personalization?
Multi-Step Errors?

Conclusion: Answering Questions w/ Reasoning that **Truly Helps**

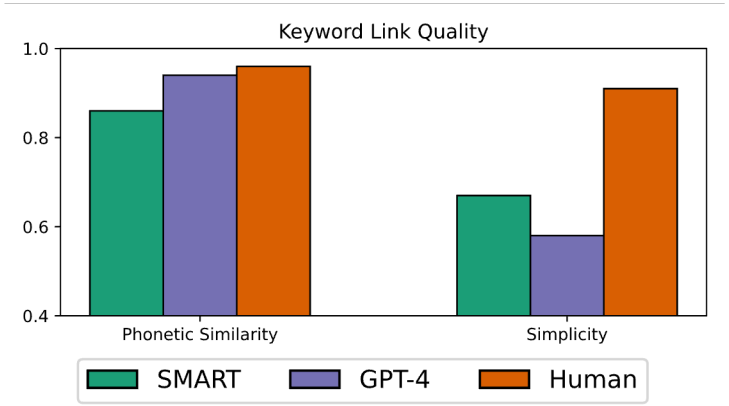
Typical preference training only teaches models what looks helpful

Perceived Helpfulness (Standard)

Mnemonic
Benevolent sounds like “benefit.” A boss giving workers benefits is kind

True Helpfulness (Ours)

But we can learn what is truly helpful



We match GPT-4, but we still need to study this further!

RQA informed us of this issue early!

Response 1
Question: What is the sum of the numbers from 1 to 27, inclusive?

Response 2
Question: What is 755 + 1?

Is LLM development aligned with helping users? **No**

Part I: Correctness Evaluation

Multiple-Choice Question Answering

Question: What's the capital of France?
(A) Paris
(B) London
(C) The Moon
Answer: (A) ✓

Factual Question Answering

Question: What's the capital of France?
Answer: Paris ✓

*Optimized to give responses matching defined answers, **hiding reasoning***

Part II: Preference Training

Question

What's the capital of France?

Chosen

Paris is the current capital of France, known for its food…

>

Rejected

Starts with a "P" and rhymes with "Maris"—it's Paris!

*Trained on responses most **users** pick and **perceive to be helpful***

[1] Process of Elimination (ACL 2024, Findings)
[2] Reverse Question Answering (NAACL 2025)

[3] Mnemonic Generation (EMNLP 2024)
[4] Plan Helpfulness for Multi-Step QA (Proposed)
[5] Personalized Preferences in QA (Proposed)

Is LLM development aligned with helping users? **No**

Part I: Correctness Evaluation

Multiple-Choice Question Answering

Question: What's the capital of France?
(A) Paris
(B) London
(C) The Moon
Answer: (A) ✓

Factual Question Answering

Question: What's the capital of France?
Answer: Paris ✓

*Optimized to give responses matching defined answers, **hiding reasoning***

Part II: Preference Training

Question

What's the capital of France?

Chosen

Paris is the current capital of France, known for its food…

>

Rejected

Starts with a "P" and rhymes with "Maris"—it's Paris!

*Trained on responses most **users** pick and **perceive to be helpful***

[1] Process of Elimination (ACL 2024, Findings)
[2] Reverse Question Answering (NAACL 2025)

[3] Mnemonic Generation (EMNLP 2024)
[4] Plan Helpfulness for Multi-Step QA (Proposed)
[5] Personalized Preferences in QA (Proposed)

Recall: How can QA systems help users achieve their goals?

Goal: Learn Something New

What does “LLM” mean?

Goal: Solve a Multi-Step Problem

How can I get my refund?

Goal: Receive Tailored Advice

How do I get 5 professors in the same room / time?

Goal: Recall Forgotten Information

Who gave that proposal talk with too many emojis?

Recall: How can QA systems help users achieve their goals?

Goal: Learn Something New

What does “LLM” mean?

Goal: Solve a Multi-Step Problem

How can I get my refund?

Problem-Solving

Goal: Receive Tailored Advice

How do I get 5 professors in the same room / time?

Goal: Recall Forgotten Information

Who gave that proposal talk with too many emojis?

Proposed: *Plan Helpfulness for Multi-Step Question Answering*

Supporting problem-solving

Multi-Step Question
You have $32 to spend on groceries. You buy a loaf of bread for $3, a candy bar for $2, and 1/3 of what's left on a turkey. How much money do you have left?

Supporting problem-solving with ***plans***

Multi-Step Question
You have $32 to spend on groceries. You buy a loaf of bread for $3, a candy bar for $2, and 1/3 of what's left on a turkey. How much money do you have left?

Plan Leading to the Answer (Reasoning Chain)
Step 1: Find how much money is left after buying bread
Step 2: Find the remaining money after buying candy
Step 3: Calculate the amount spent on the Turkey
Step 4: Subtract the cost in (3) from the amount in (4) to get the final answer!

Two benefits:

1. Guidance to the answer in real time (AI-assisted decision-making)
2. Study aid for learning to solve similar questions in the future (Scaffolding)

How can we generate more helpful plans?

How can we generate more helpful plans?

Question
You have $32 to spend on groceries…

Plan 1
1. Find money after buying bread 2. Find money after buying candy 3. Calculate money spent on the Turkey 4. Subtract (3) from the amount in (4)

?

Plan 2
1. Subtract the amount spent on food 2. Multiply the amount from (1) by one minus the fraction spent on turkey

A helpful plan allows users to answer questions:

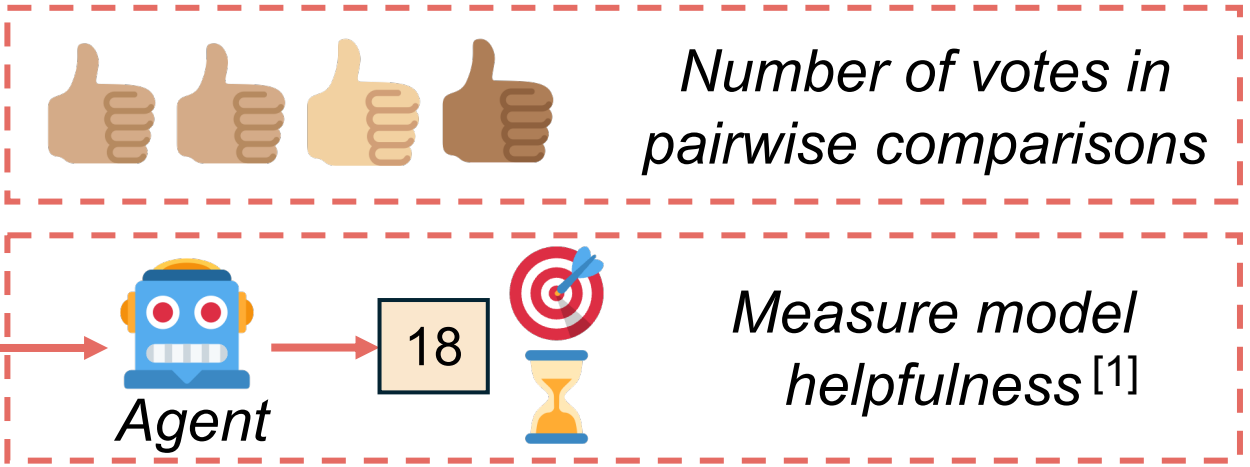
 Correctly

 Quickly

How do researchers try to measure plan helpfulness?

Question
You have $32 to spend on groceries…

Plan 1
1. Find money after buying bread
2. Find money after buying candy
3. Calculate money spent on the Turkey
4. Subtract (3) from the amount in (4)



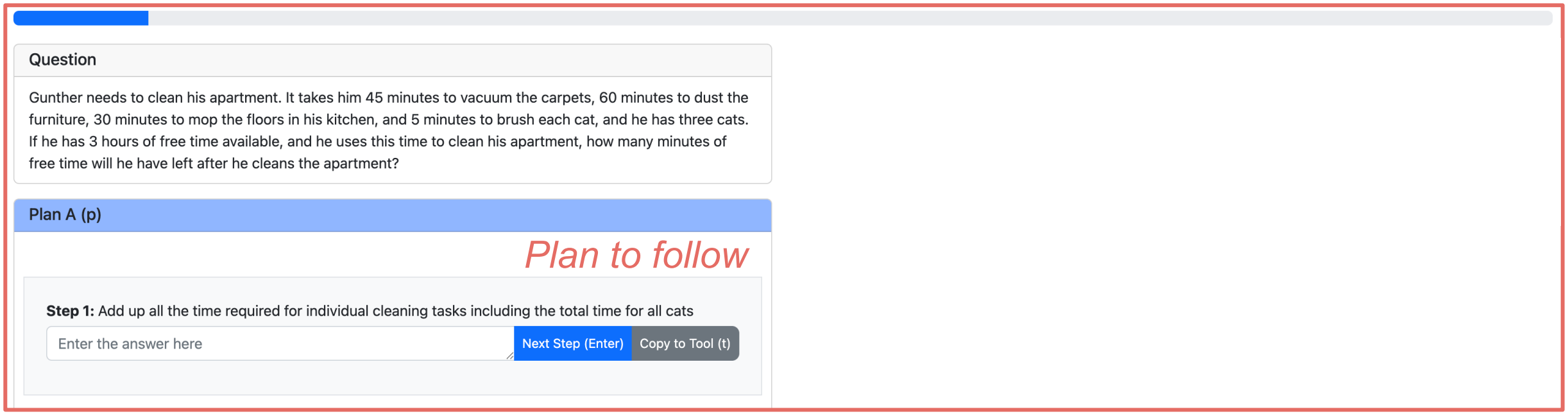
But we don't know:

1. Can users predict which plans are helpful?
2. Does helpful for models => helpful for humans?

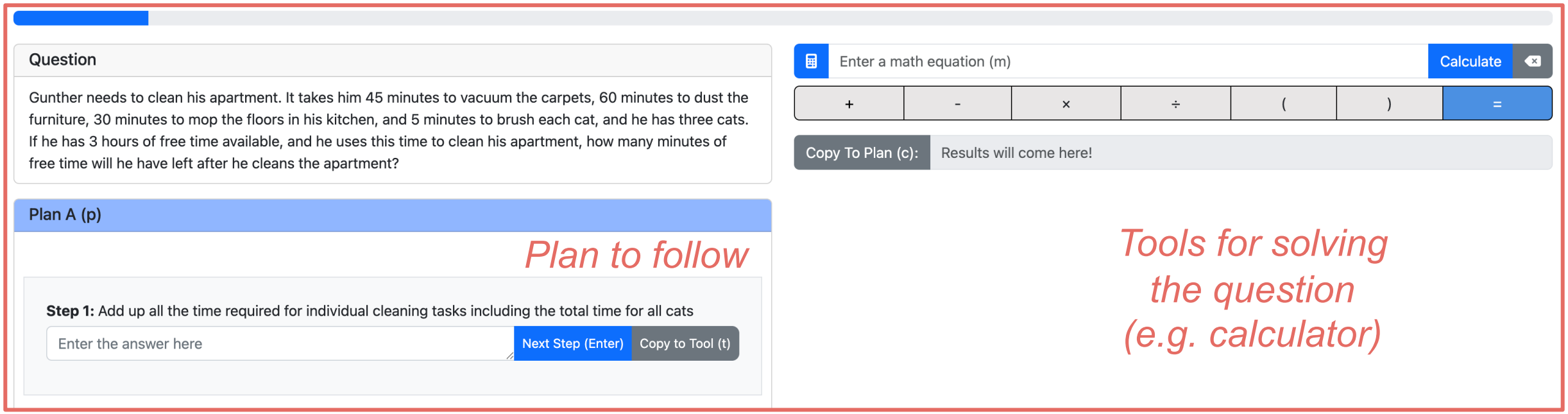
[1] Trial and Error: Exploration-Based Trajectory Optimization for LLM Agents

Proposal: An interface to measure how well plans help users

Proposal: An interface to measure how well plans help users



Proposal: An interface to measure how well plans help users

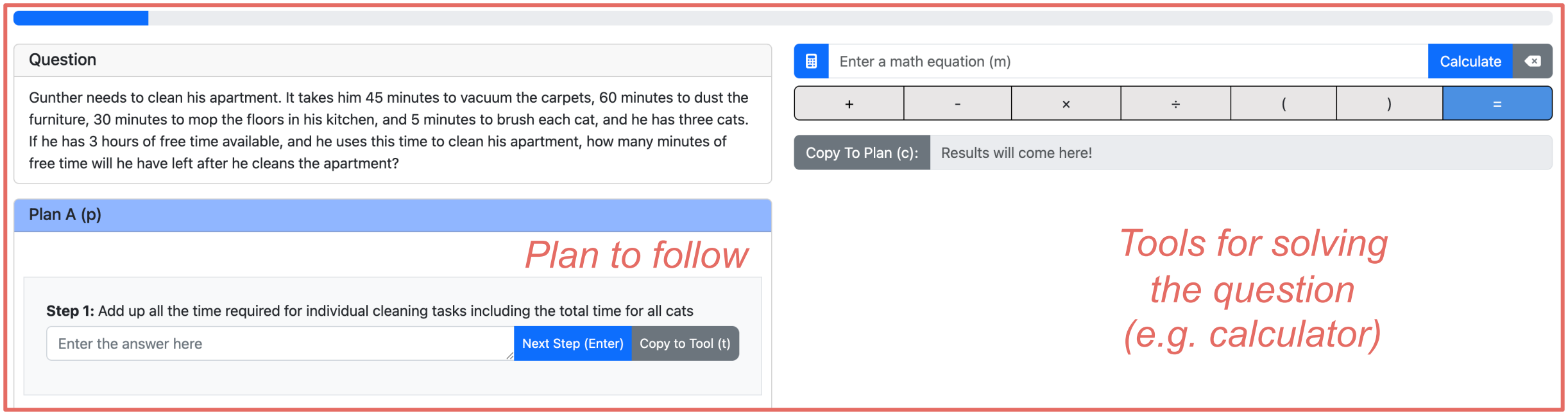


With this data, we plan to compare:

What *actually* helps humans

What humans *think* is helpful (adapt mnemonic UI)

Proposal: An interface to measure how well plans help users



With this data, we plan to compare:

What *actually* helps humans + models

What humans + models *think* is helpful (adapt mnemonic UI)

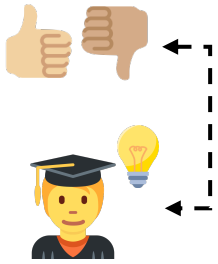
Conclusion: Answering Questions w/ Reasoning that **Truly** Helps

Does standard preference training *consistently* teach models what **looks** helpful?

Mnemonic
Benevolent sounds like “benefits”. A boss giving employees benefits is kind

Perceived Helpfulness

True helpfulness



Not the same!

Conclusion: Answering Questions w/ Reasoning that **Truly** Helps

Does standard preference training *consistently* teach models what **looks** helpful?

Plan: What's the 10th oldest U.S. state?

1. Find a list of states in the U.S.
2. Sort them by their founding date
3. Pick the 10th state on this list!

Perceived Helpfulness

True helpfulness

Are they the same?

Building an interface to measure true plan helpfulness for math/trivia

Do agents simulate user helpfulness?

Plan

…

Agent

Answer: Virginia

Agents do the same + study what helped them

How can we make plans truly helpful?

Perceived Helpful but Unhelpful Plan

…

- Qualitative biases that trick users (complexity)
- Custom preference training for helpful plans

Is LLM development aligned with helping users? **No**

Part I: Correctness Evaluation

Multiple-Choice Question Answering

Question: What's the capital of France?
(A) Paris
(B) London
(C) The Moon
Answer: (A) ✓

Factual Question Answering

Question: What's the capital of France?
Answer: Paris ✓

*Optimized to give responses matching defined answers, **hiding reasoning***

Part II: Preference Training

Question

What's the capital of France?

Chosen

Paris is the current capital of France, known for its food…

>

Rejected

Starts with a "P" and rhymes with "Maris"—it's Paris!

*Trained on responses most **users** pick and **perceive to be helpful***

[1] Process of Elimination (ACL 2024, Findings)
[2] Reverse Question Answering (NAACL 2025)

[3] Mnemonic Generation (EMNLP 2024)
[4] Plan Helpfulness for Multi-Step QA (Proposed)
[5] Personalized Preferences in QA (Proposed)

Is LLM development aligned with helping users? **No**

Part I: Correctness Evaluation

Multiple-Choice Question Answering

Question: What's the capital of France?
(A) Paris
(B) London
(C) The Moon
Answer: (A) ✓

Factual Question Answering

Question: What's the capital of France?
Answer: Paris ✓

*Optimized to give responses matching defined answers, **hiding reasoning***

Part II: Preference Training

Question

What's the capital of France?

Chosen

Paris is the current capital of France, known for its food…

>

Rejected

Starts with a "P" and rhymes with "Maris"—it's Paris!

*Trained on responses **most users** pick and **perceive to be helpful***

[1] Process of Elimination (ACL 2024, Findings)
[2] Reverse Question Answering (NAACL 2025)

[3] Mnemonic Generation (EMNLP 2024)
[4] Plan Helpfulness for Multi-Step QA (Proposed)
[5] Personalized Preferences in QA (Proposed) ←

Recall: How can QA systems help users achieve their goals?

Goal: Learn Something New

What does “LLM” mean?

Goal: Solve a Multi-Step Problem

How can I get my refund?

Goal: Receive Tailored Advice

How do I get 5 professors in the same room / time?

Goal: Recall Forgotten Information

Who gave that proposal talk with too many emojis?

Recall: How can QA systems help users achieve their goals?

Goal: Learn Something New

What does “LLM” mean?

Goal: Solve a Multi-Step Problem

How can I get my refund?

Goal: Receive Tailored Advice

How do I get 5 professors in the same room / time?

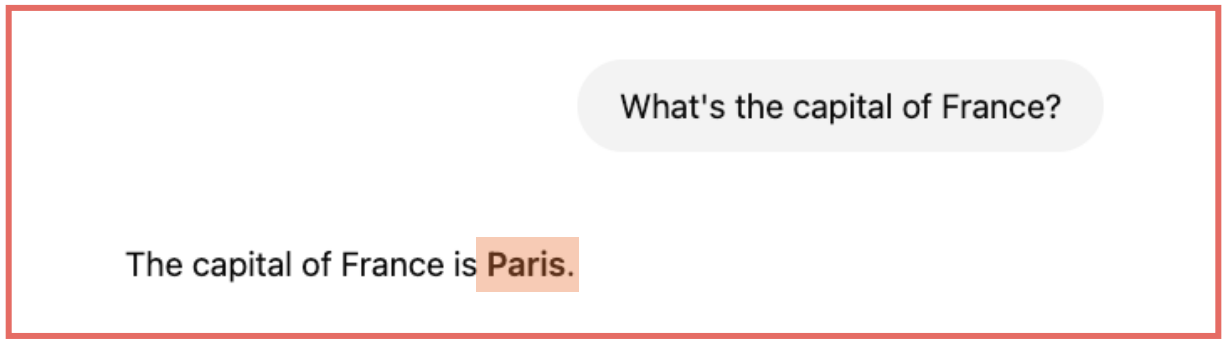
Goal: Recall Forgotten Information

Who gave that proposal talk with too many emojis?

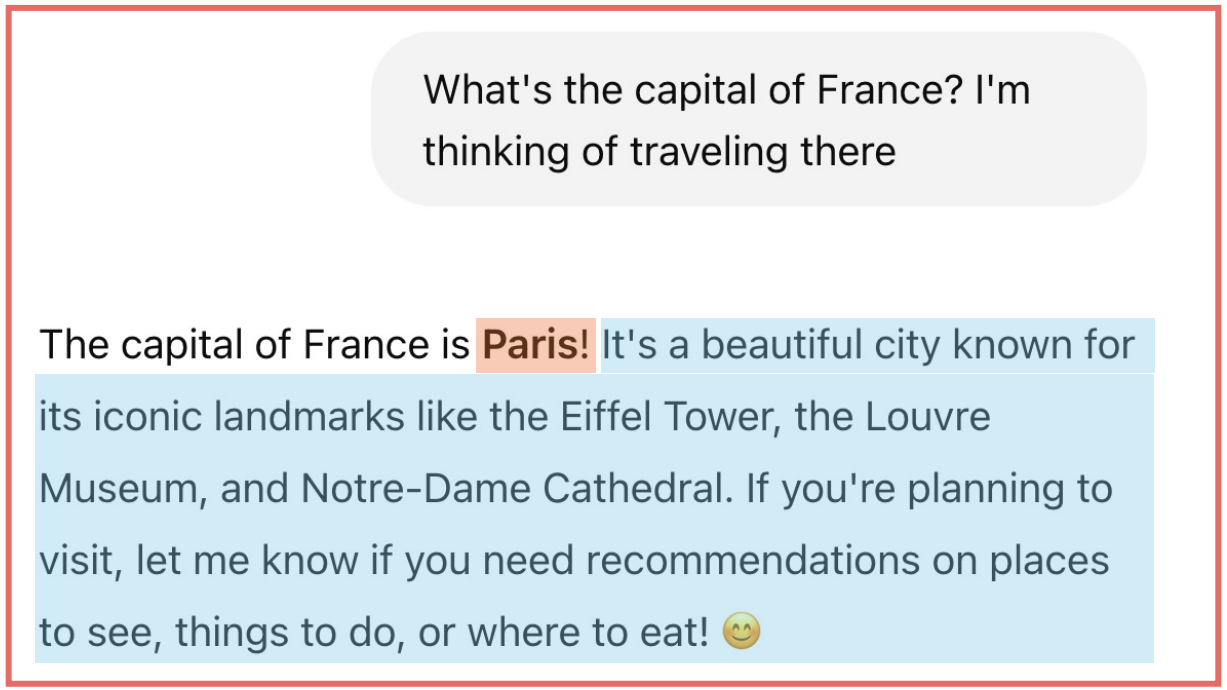
Personalization

Users often have personalized needs in QA!

Standard QA



QA With Personalization



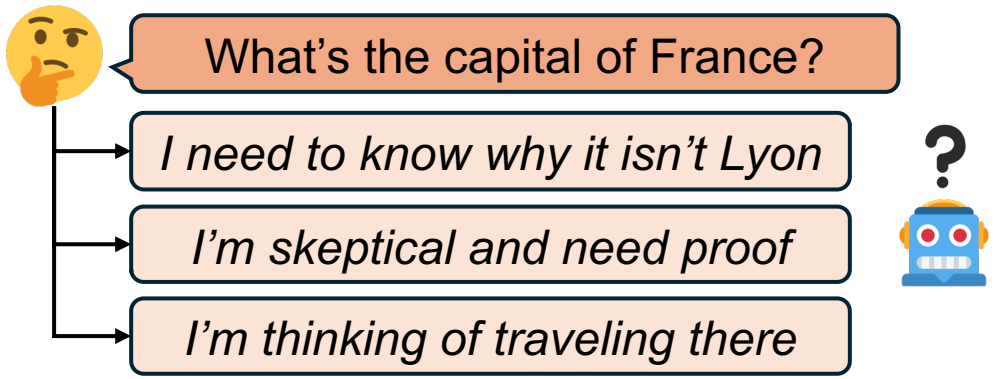
Both responses have correct answers…

…but reasoning chains must be personalized to any specified needs

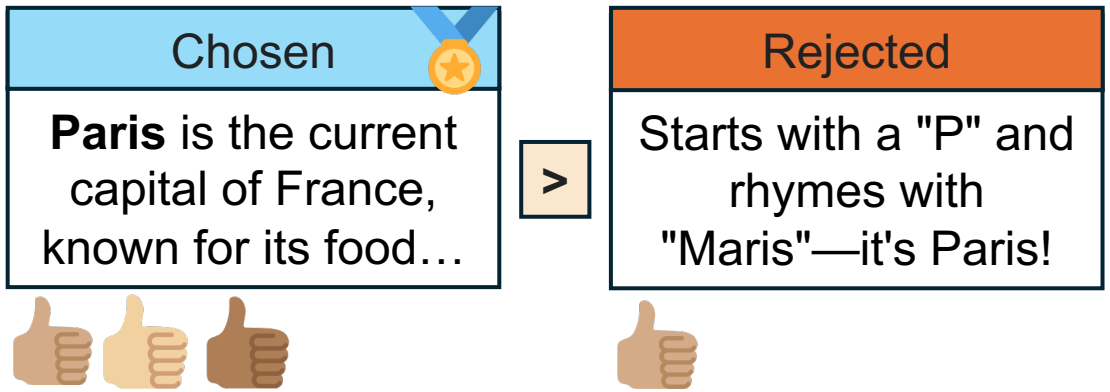
But our results indicate LLMs struggle with personalization![1, 2]

Process of Elimination: Adaptable Reasoning Flaws

Which molecule does not have a carbon-nitrogen bond?
Choices: (A) nucleic acid (B) carbohydrate
Answer: … There is no nitrogen in carbohydrates. So the incorrect answer is “carbohydrate” which is choice (B)



*Can LLMs **personalize** to user needs?*



Hypothesis: Preference training is a culprit

> Teaching *which* responses are preferred without considering *who* prefers them[3]

[1] Aligning to Thousands of Preferences via System Message Generalization
[2] Improving Context-Aware Preference Modeling for Language Models
[3] A Roadmap to Pluralistic Alignment

Proposal: Inferring user needs for training personalized QA systems

Proposal: Inferring user needs for training personalized QA systems

- LLMs can't personalize in inference, as they have not trained on user-specific needs

Typical Preference Dataset

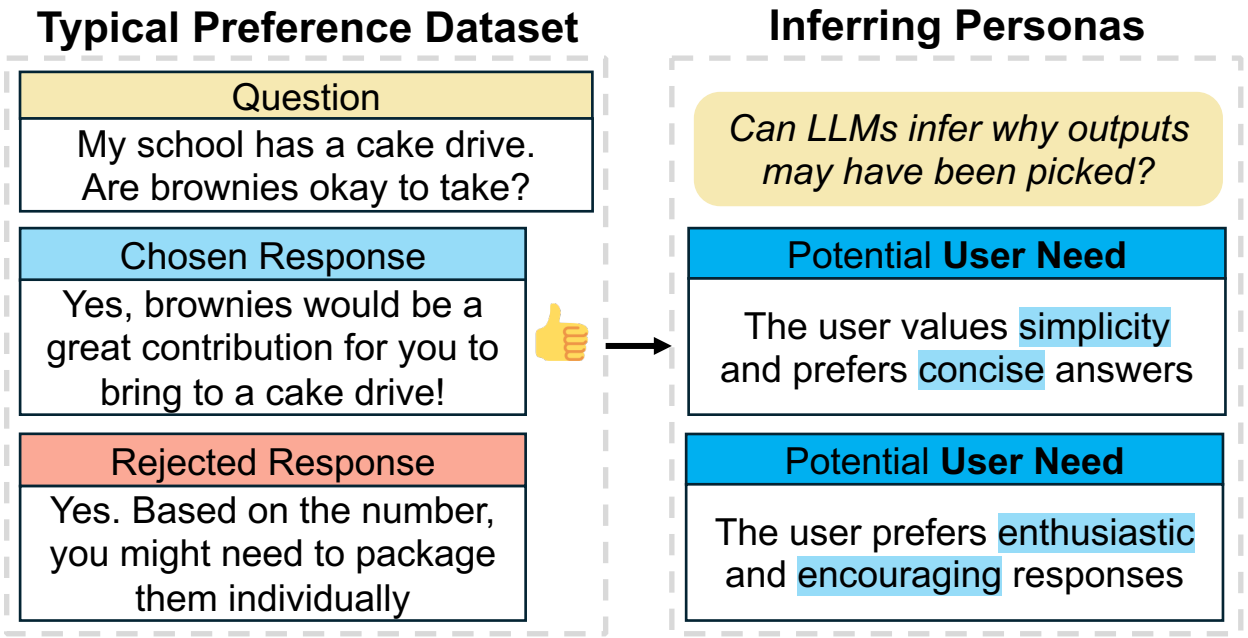
Question
My school has a cake drive. Are brownies okay to take?

Chosen Response
Yes, brownies would be a great contribution for you to bring to a cake drive!

Rejected Response
Yes. Based on the number, you might need to package them individually

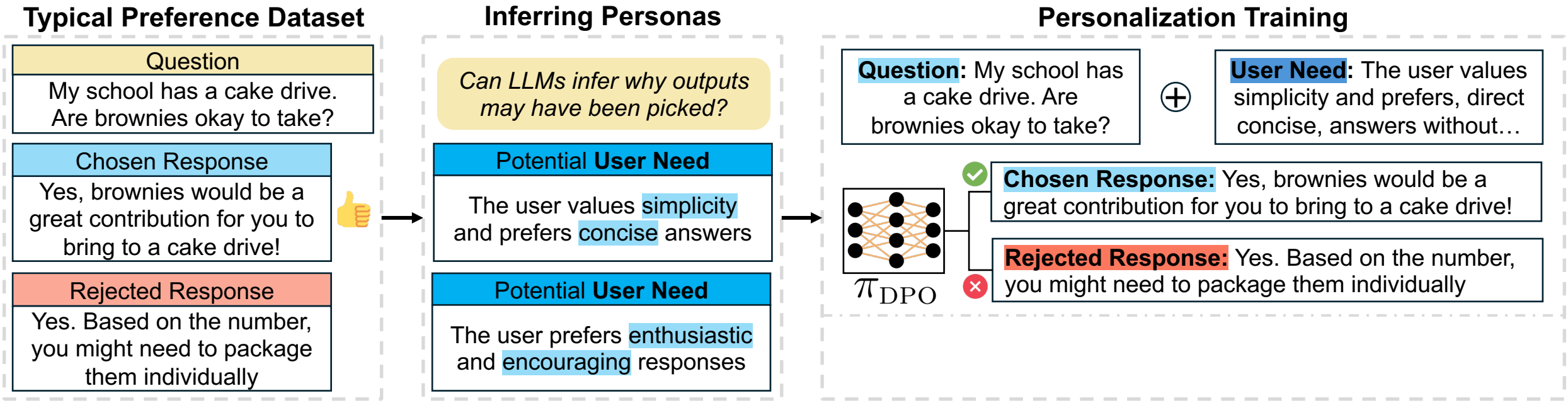
Proposal: Inferring user needs for training personalized QA systems

- LLMs can't personalize in inference, as they have not trained on user-specific needs
- Can LLMs infer these user needs in preference data to reconstruct missing training data?



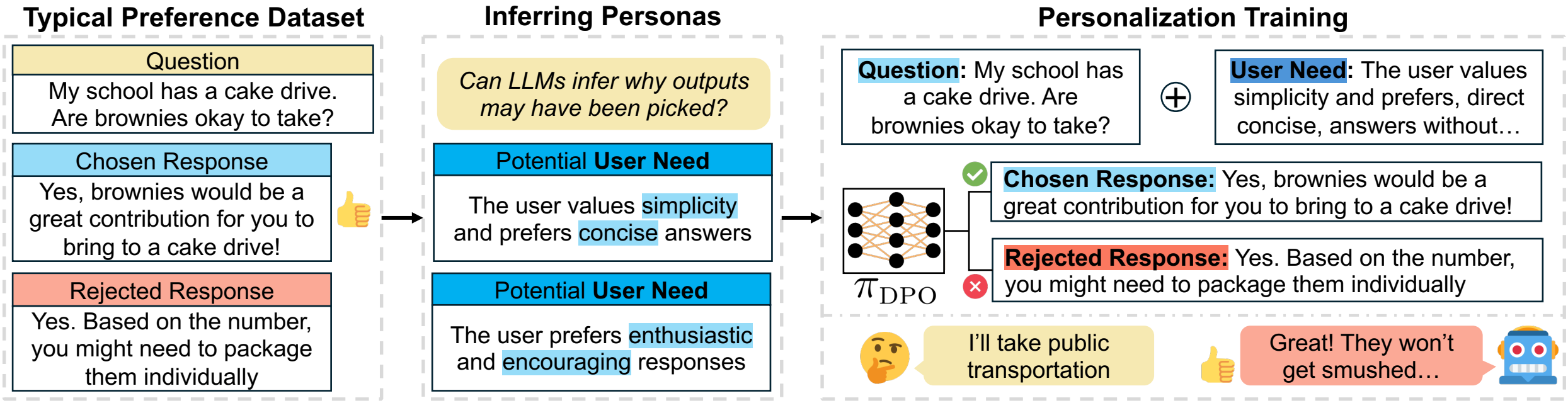
Proposal: Inferring user needs for training personalized QA systems

- LLMs can't personalize in inference, as they have not trained on user-specific needs
- Can LLMs infer these user needs in preference data to reconstruct missing training data?
- Then, we can train QA systems on these personalized needs



Proposal: Inferring user needs for training personalized QA systems

- LLMs can't personalize in inference, as they have not trained on user-specific needs
- Can LLMs infer these user needs in preference data to reconstruct missing training data?
- Then, we can train QA systems on these personalized needs



Evaluation: When a user gives a specific need, how satisfied is this same user?[1]

[1] Personalized content recommendation and user satisfaction: Theoretical synthesis and empirical findings

Proposal (follow-up): Can we also prevent over-generalization?

How do I code FizzBuzz? I really like Pizza. I only want one program

Nice! Let's combine your love for **Pizza** with the classic **FizzBuzz** in a single, fun Python program.

python Copy Edit

```python
def fizzbuzz_with_pizza():
    for i in range(1, 101):
        if i % 3 == 0 and i % 5 == 0:
            print("FizzBuzz")
        elif i % 3 == 0:
            print("Fizz")
        elif i % 5 == 0:
            print("Buzz")
        else:
            print(i)

    print("\nYou made it through FizzBuzz! Time for a slice of 🍕 PIZZA 🍕!")

fizzbuzz_with_pizza()
```

We might automatically infer multiple needs to tailor to

But it doesn't mean all of them are relevant!

Question: Can you code FizzBuzz?

1. The user codes in Python ✅
2. The user likes Pizza ❌
3. The user values conciseness ❓

Can we teach QA systems *when* to personalize?[1]

[1] RLVF: Learning from Verbal Feedback without Overgeneralization

Conclusion: Answering Questions w/ Reasoning that **Truly Helps**

Can LLMs infer *why* users may prefer responses for personalization?

Chosen
Paris is the current capital of France…

Rejected
It starts with a "P" and rhymes with "Maris"

The user is boring and hates jokes

Plausible needs for training data?

Conclusion: Answering Questions w/ Reasoning that **Truly Helps**

Can LLMs infer *why* users may prefer responses for personalization? And *when*?

Chosen
Paris is the current capital of France…

Rejected
It starts with a "P" and rhymes with "Maris"

The user is boring and hates jokes

Plausible needs for training data?

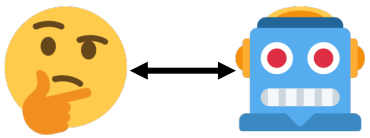
The user is boring + hates jokes

Tell me a joke!

No since you hate them

Learning from more helpfulness signals

Is our personalized QA system helpful?



First step: satisfaction with the user-QA interaction

But can we also apply or design true helpfulness measures?

Did PoE signal this weakness early?

What's the capital of France?

- *I need to know why it isn't Lyon*
- *I'm skeptical and need proof*
- *I'm thinking of traveling there*

PoE: Can LLMs personalize to user needs?

Is LLM development aligned with helping users? **No**

Part I: Correctness Evaluation

Multiple Choice Question Answering

Question: What's the capital of France?
(A) Paris
(B) London
(C) The Moon
Answer: (B) ✅

Factual Question Answering

Task: What's the capital of France?
Answer: Paris ✅

Optimized to give responses matching defined answers, ***hiding reasoning***

Part II: Preference Training

Question

What's the capital of France?

Chosen

Paris is the current capital of France, known for its food…

>

Rejected

Starts with a "P" and rhymes with "Maris"—it's Paris!

Trained on responses ***most users*** *pick and* ***perceive to be helpful***

[1] Process of Elimination (ACL 2024, Findings)
[2] Reverse Question Answering (NAACL 2025)

[3] Mnemonic Generation (EMNLP 2024)
[4] Plan Helpfulness for Multi-Step QA (Proposed)
[5] Personalized Preferences in QA (Proposed)

But we can develop LLMs to help users in QA!

Reasoning Part I: ~~Correctness~~ Evaluation

Process of Elimination

Question: What's the capital of France?
(A) Paris
(B) London
(C) The Moon
Answer: The Moon isn't a country, so not (C) ✅

Reverse Question Answering

Task: Give me a question with the answer "Paris"
Answer: What city has the Eiffel Tower? ✅

Evaluate reasoning *beyond answer correctness to expose model QA failures*

Improved Part II: Preference Training

Question

What's the capital of France?

Chosen	>	Rejected
Paris is the current capital of France, known for its food…		Starts with a "P" and rhymes with "Maris"—it's Paris!

I don't like jokes! 😠

Capture what ***actually helps*** *users and* ***user needs*** *behind preferences*

[1] Process of Elimination (ACL 2024, Findings)
[2] Reverse Question Answering (NAACL 2025)

[3] Mnemonic Generation (EMNLP 2024)
[4] Plan Helpfulness for Multi-Step QA (Proposed)
[5] Personalized Preferences in QA (Proposed)

Thank you!

:)

Nishant Balepur

Jordan Boyd-Graber (Chair)
Rachel Rudinger (Co-chair)
Fumeng Yang (Department Representative)
David Weintrop (Dean's Representative)
Shi Feng (External Member, George Washington University)

UNIVERSITY OF MARYLAND