PAUL G. ALLEN SCHOOL

UNIVERSITY of WASHINGTON

Natural Language Processing

CSE 447 @ UW

In-Context Learning, Prompting, and Basics of Reasoning

Guest Lecturer: Chan Young Park

Some slides adapted from: Charlie Dickens

- ★ **Basics of Prompting**
  - In-Context Learning
- ★ **More Strategic Prompting**
  - Chain-of-Thought Reasoning (and More)
- ★ **Advanced Prompting & Basics of Reasoning**
  - Knowledge Enhanced Reasoning & Dialog
  - Think-Before-Speaking
  - Agent & Tool Use
  - Preference Elicitation with Clarification Questions

Advanced Prompting & Basics of Reasoning:

Knowledge Enhanced Reasoning & Dialog

Think-Before-Speaking

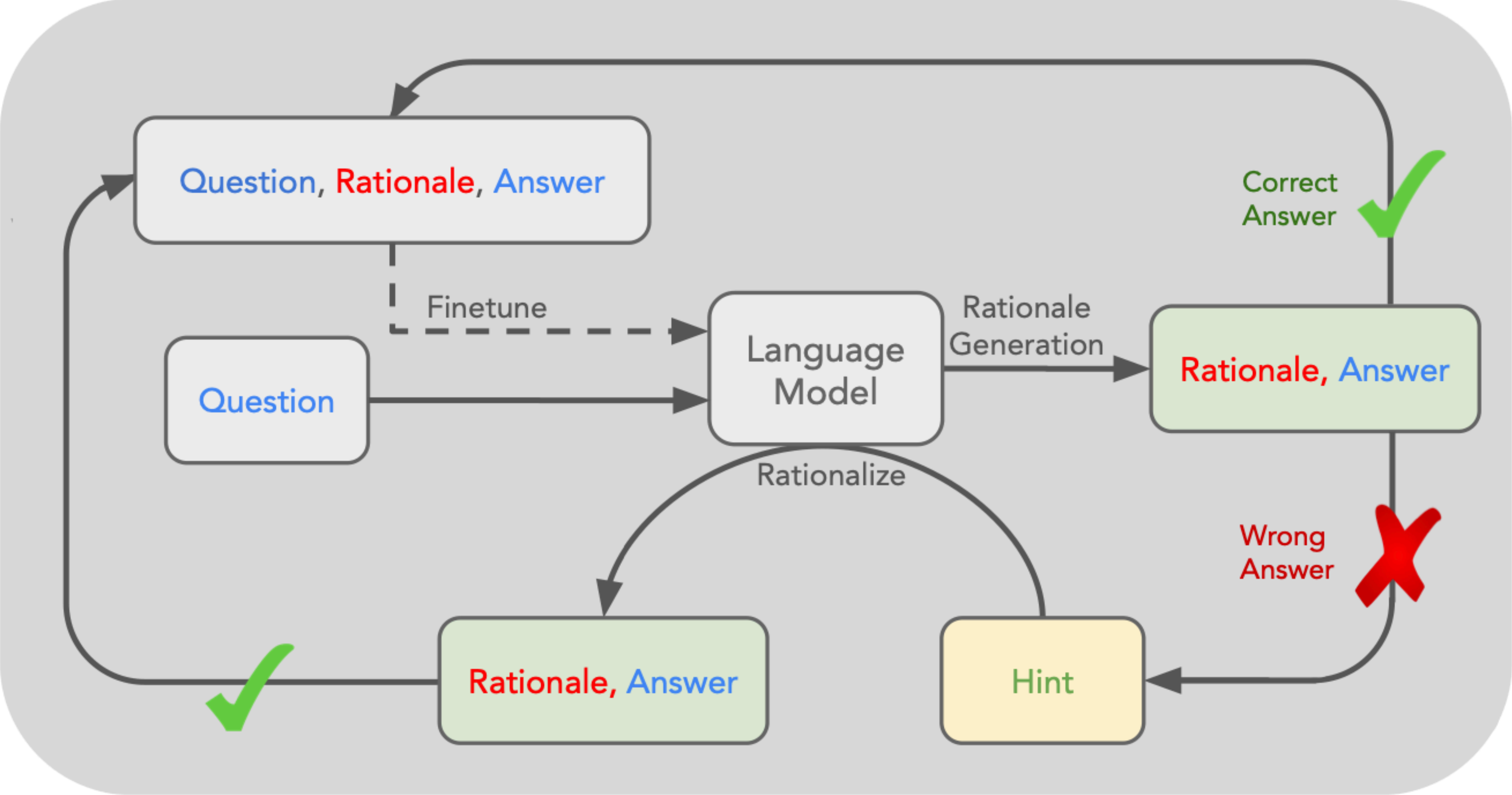
Agent & Tool Use

Preference Elicitation with Clarification Questions

STaR: Self-Taught Reasoner

(STaR, Zelikman et al. 2022)

Bootstrapping Reasoning With Reasoning




```
Q: What can be used
to carry a small dog?
Answer Choices:
(a) swimming pool
(b) basket
(c) dog show
(d) backyard
(e) own home
A: The answer must be
something that can be
used to carry a small
dog. Baskets are
designed to hold things.
Therefore, the answer
is basket (b).
```

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.

STaR: Self-Taught Reasoner

(STaR, Zelikman et al. 2022)

Bootstrapping Reasoning With Reasoning

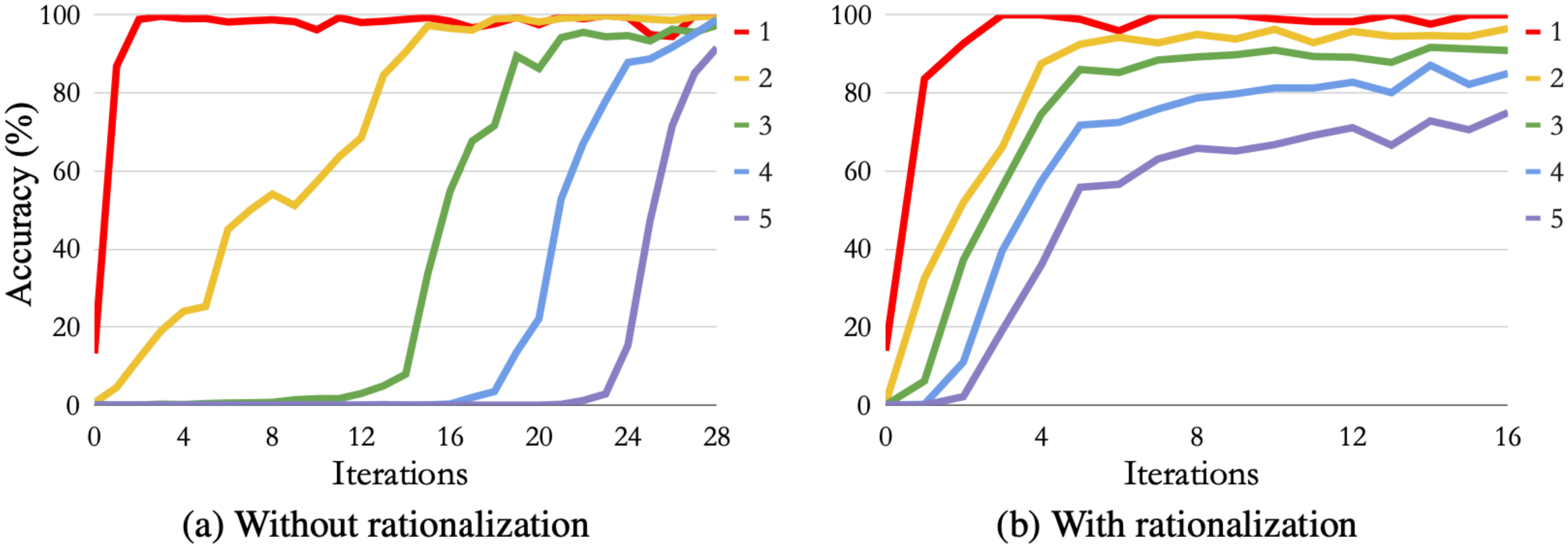


Figure 4: A visualization of the accuracy of n-digit summation with each iteration of STaR with and without rationalization for arithmetic. Each series corresponds to the accuracy of summing two n-digit numbers.

Quiet-STaR:
Language Models Can Teach Themselves to Think Before Speaking

(Quiet-STaR, Zelikman et al. 2024)

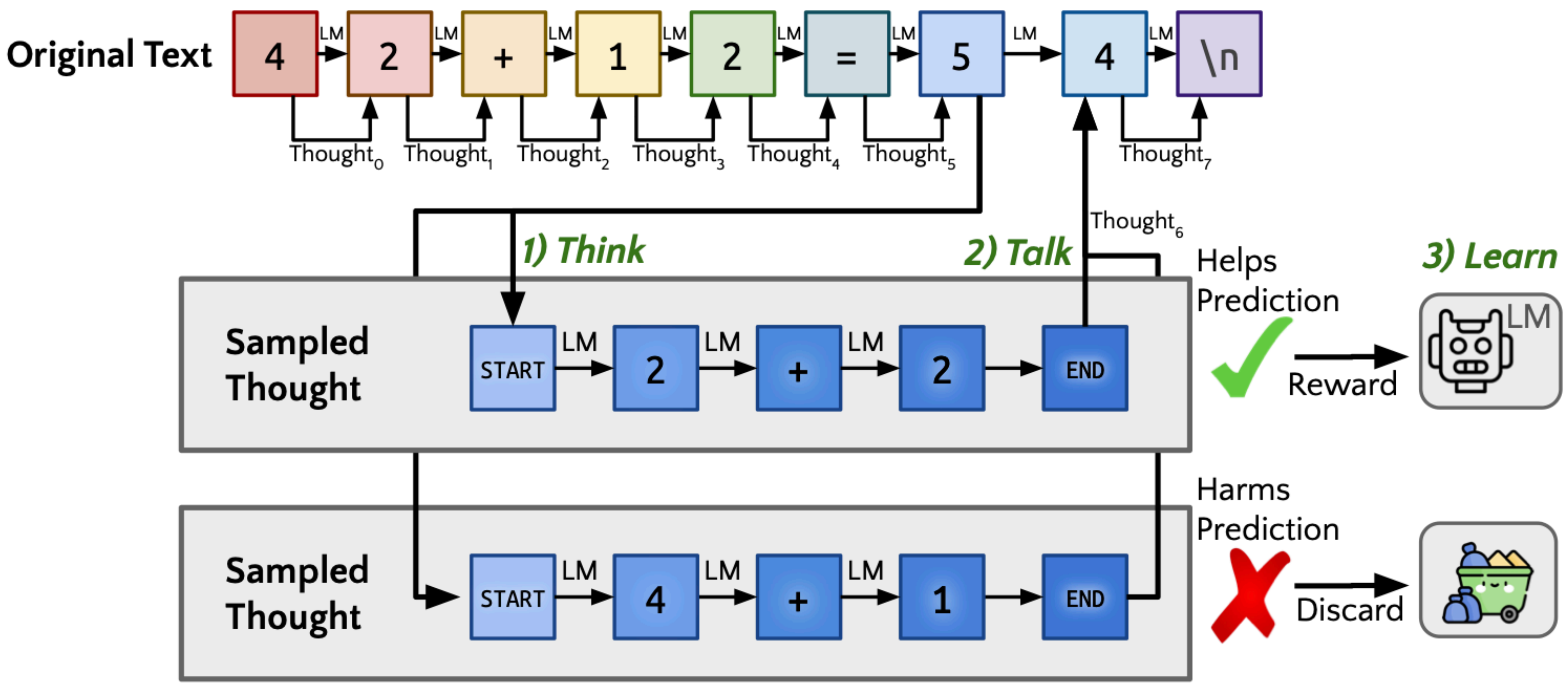
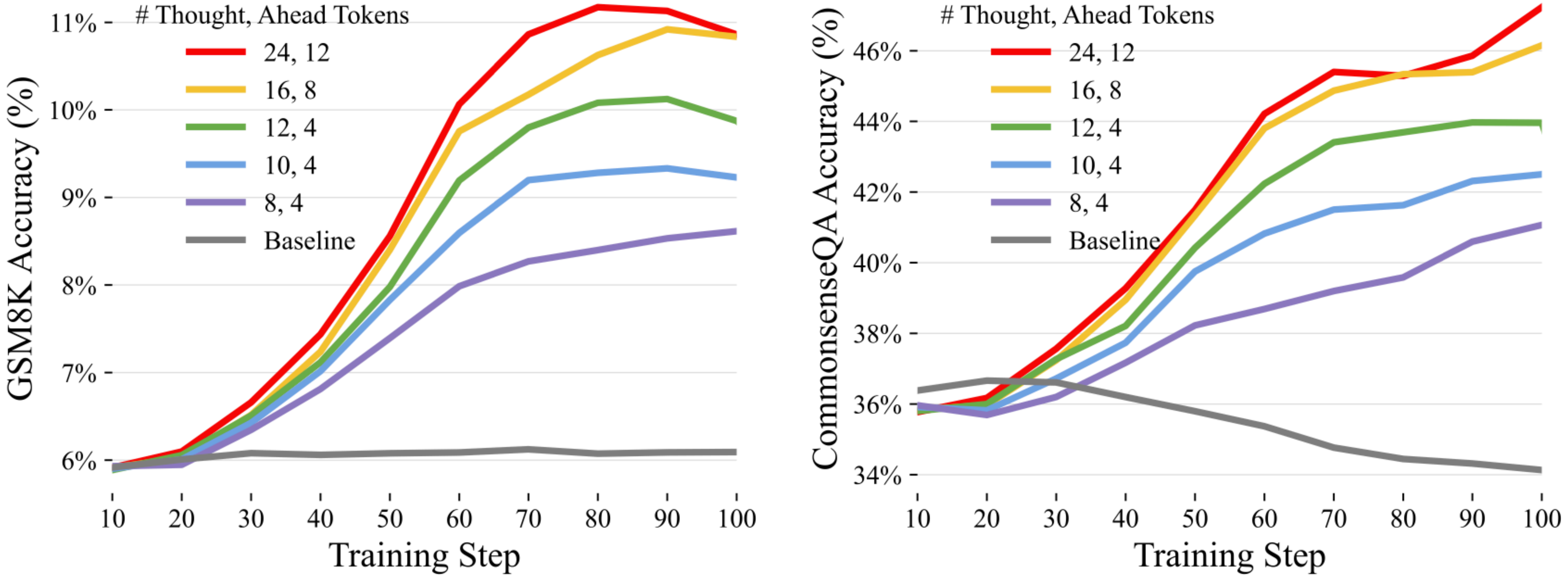


Figure 1: **Quiet-STaR**. We visualize the algorithm as applied during training to a single thought. We generate thoughts, in parallel, following all tokens in the text (think). The model produces a mixture of its next-token predictions with and without a thought (talk). We apply REINFORCE, as in STaR, to increase the likelihood of thoughts that help the model predict future text while discarding thoughts that make the future text less likely (learn).

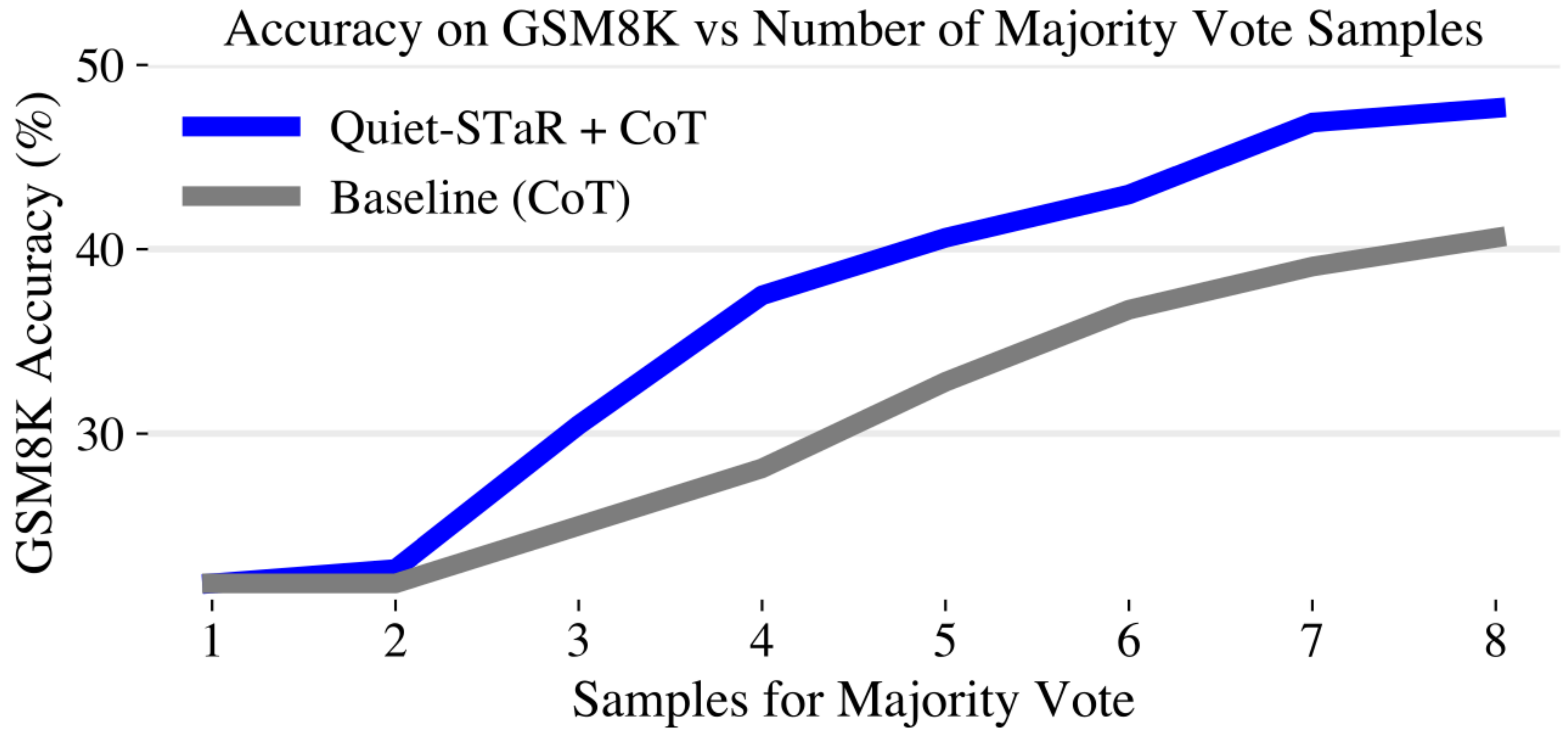
Quiet-STaR:

Language Models Can Teach Themselves to Think Before Speaking

[(Quiet-STaR, Zelikman et al. 2024)]



(a) GSM8K (b) CommonsenseQA

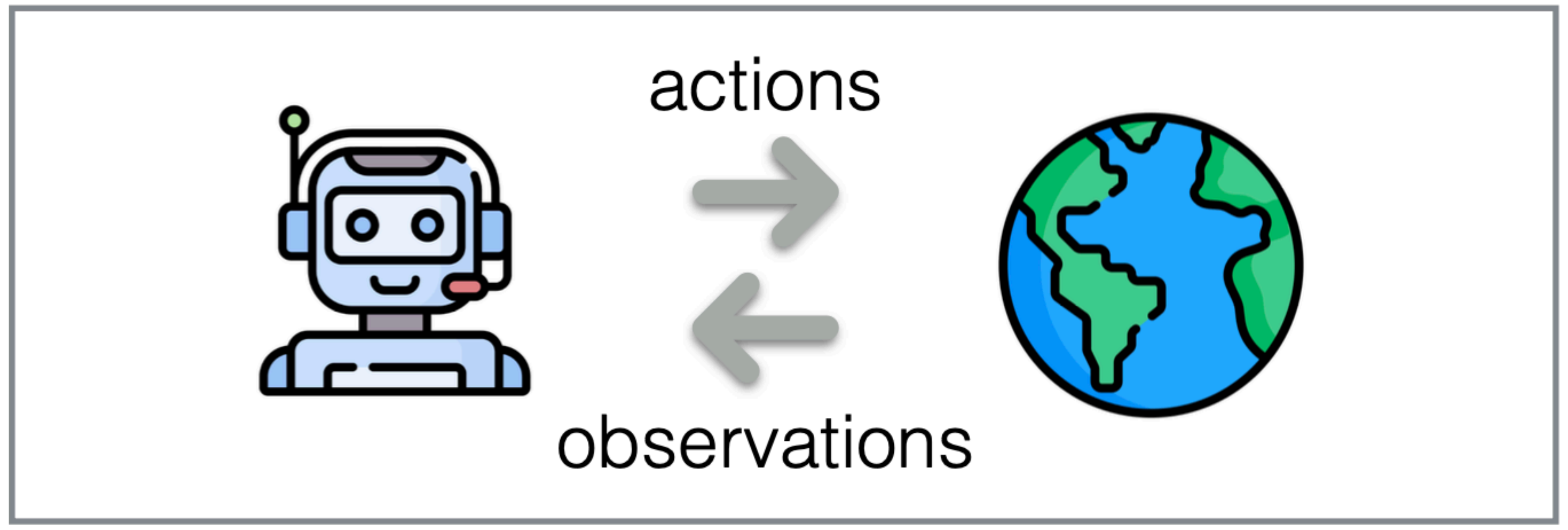


What are LLM-Powered Agents?

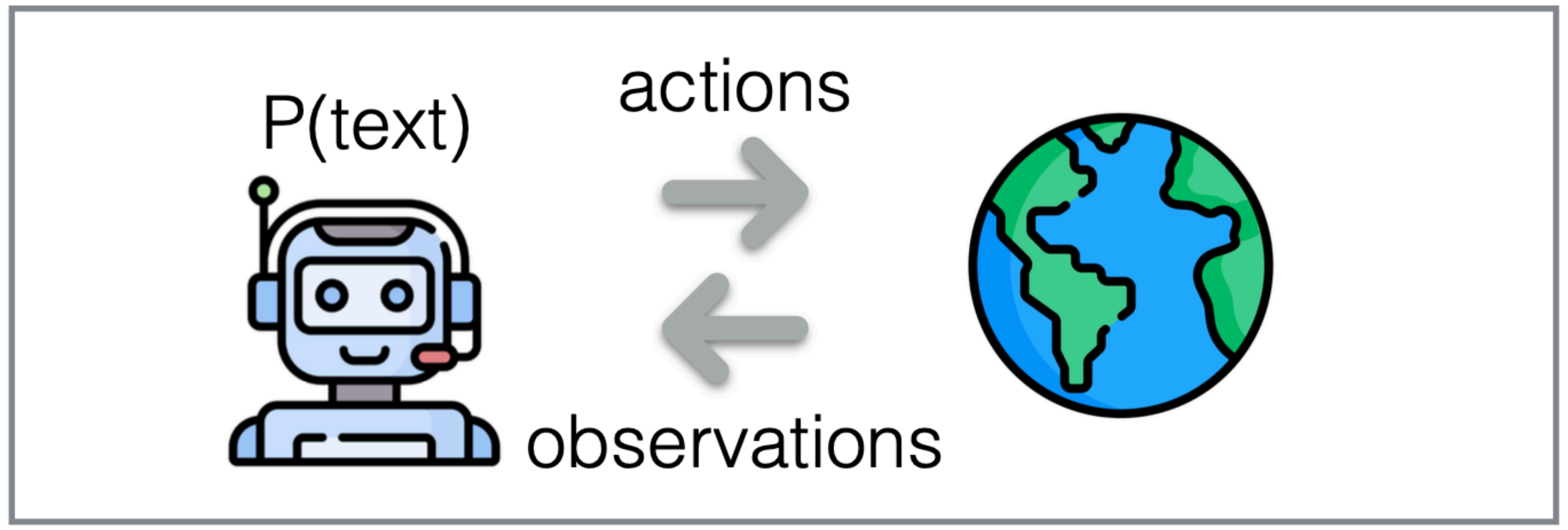
Language models predict text

P(text)

AI agents iteratively perform actions in the world



LM agents are an agent with a an LM backbone



Minimal Components of LLM Agents:

- Underlying LLM
- Prompt
- Action/Observation Space

Things that LLMs Are Bad At...

Numerical/symbolic operations

1. Calculation
2. Logic deduction
3. Exact operations

Knowledge not in their pre-training corpus

1. Tail factual knowledge
2. New information
3. Private information

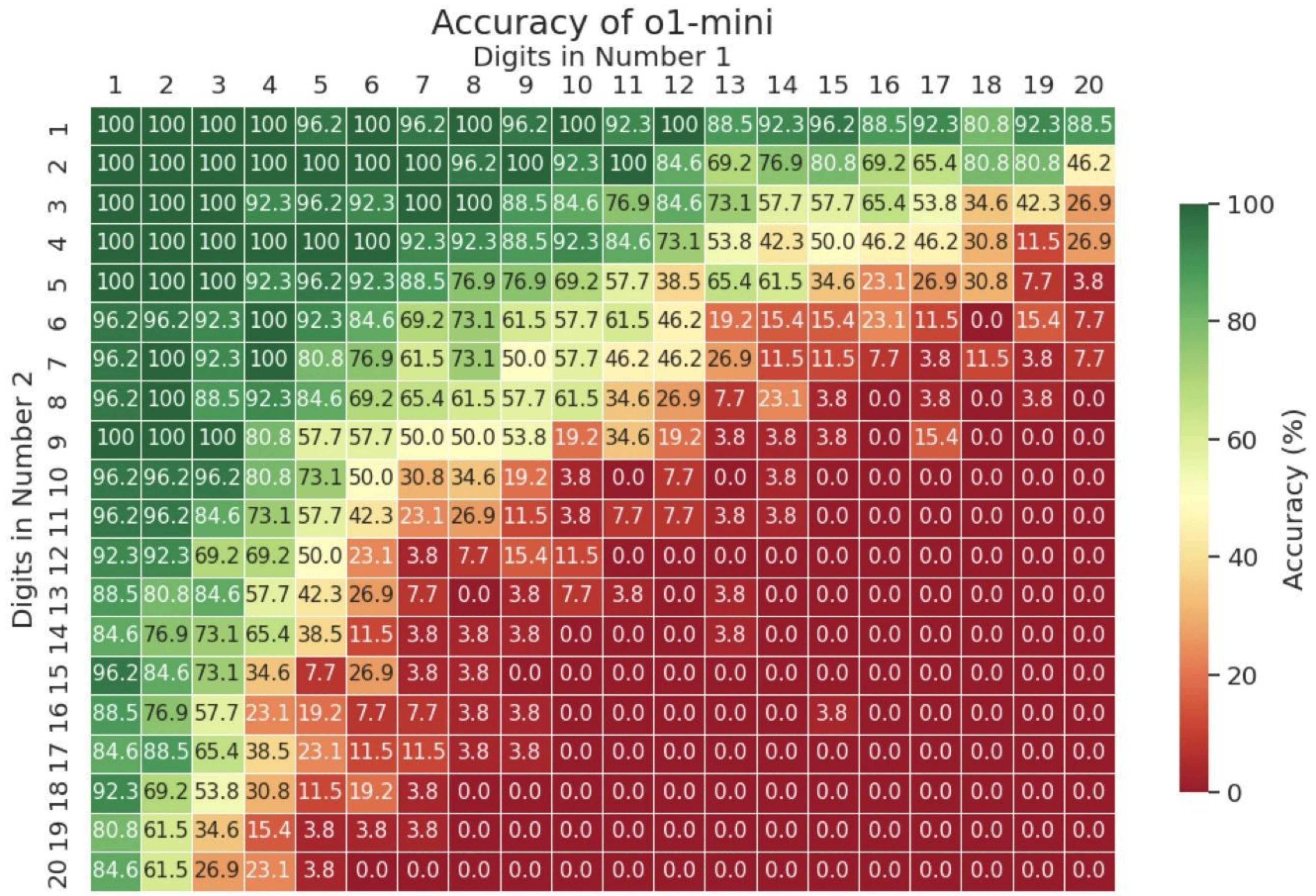
Interaction with the external world

1. Non natural language interfaces
2. Physical world
3. Environmental information (e.g., time)

To aid LLMs on tasks beyond their ability: **knowledge, symbolic, and external environment operations**

Why Tools?

LLMs are not the solution for everything. (Not AGI yet. Surprise?)



Multiplication Accuracy of OpenAI O1 (Yuantian Deng, X)

O1 cannot solve multiplications of 10+ digits…

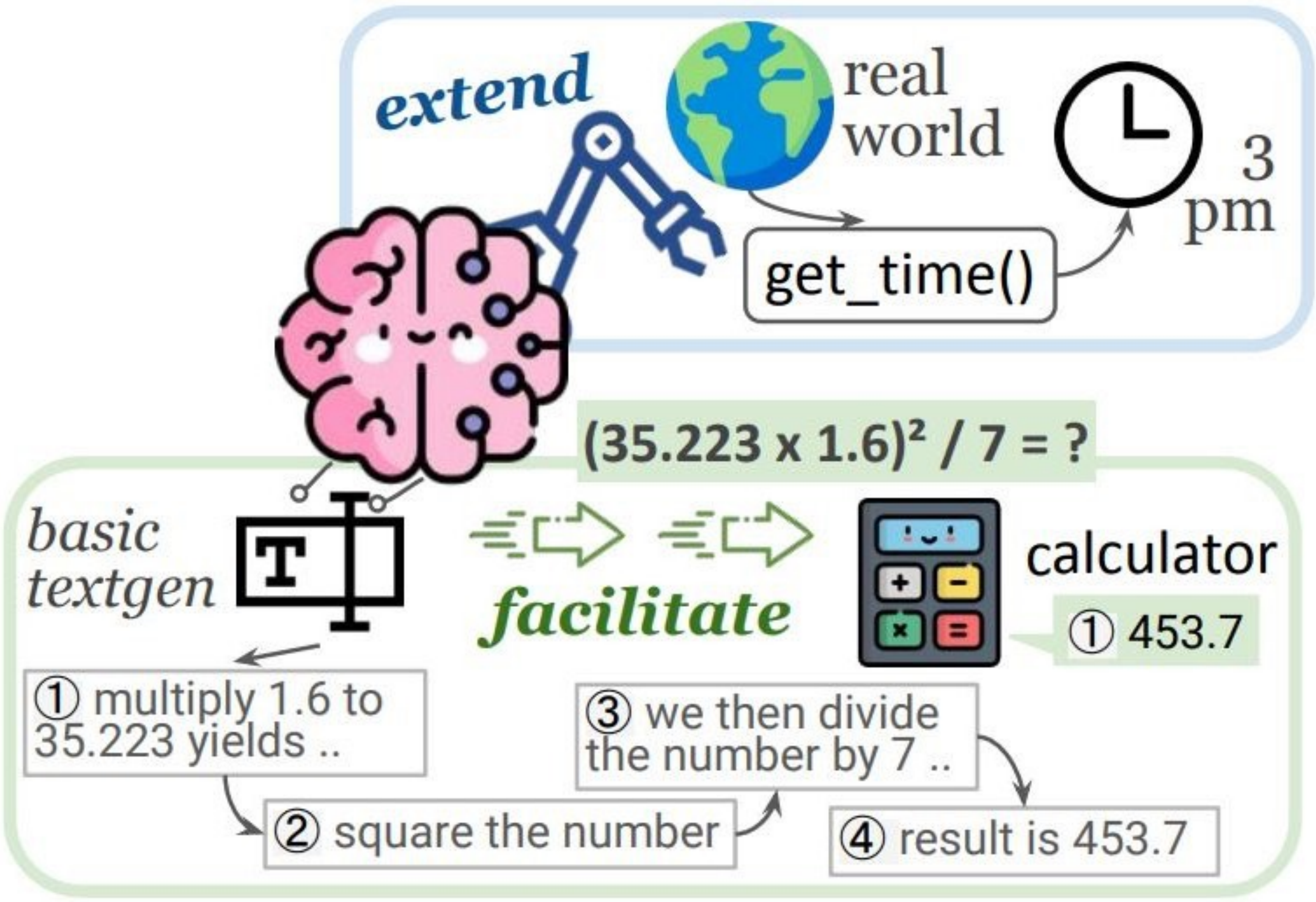
But why should we expect LLMs to do so?

Humans cannot do this on-the-fly either… but we can use calculator to solve it easily.

Can LLMs use tools too?

What are Tools? (Wang et al. 2024.)

Definition: An LM-used tool is a function interface to a computer program that runs externally to the LM, where the LM generates the function calls and input arguments in order to use the tool.



A tool is:

- A Computer Program
- External to the LM
- Used through generated function calls

What are Tools?

(Wang et al. 2024.)

Category	Example Tools

Knowledge access	`sql_executor(query: str) -> answer: any`
	`search_engine(query: str) -> document: str`
	`retriever(query: str) -> document: str`

Computation activities	`calculator(formula: str) -> value: int | float`
	`python_interpreter(program: str) -> result: any`
	`worksheet.insert_row(row: list, index: int) -> None`

Interaction w/ the world	`get_weather(city_name: str) -> weather: str`
	`get_location(ip: str) -> location: str`
	`calendar.fetch_events(date: str) -> events: list`
	`email.verify(address: str) -> result: bool`

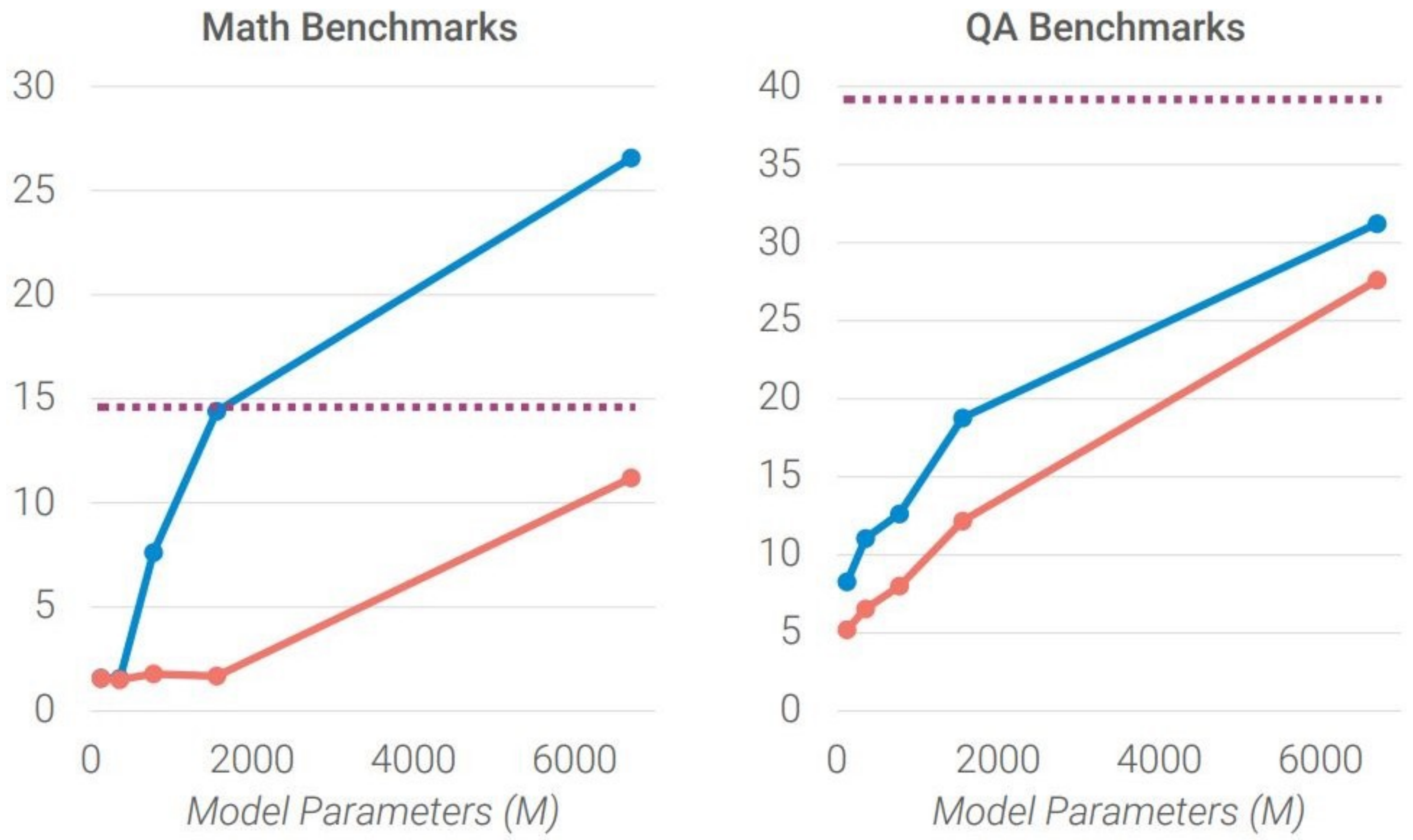
Non-textual modalities	`cat_image.delete(image_id: str) -> None`
	`spotify.play_music(name: str) -> None`
	`visual_qa(query: str, image: Image) -> answer: str`

Special-skilled LMs	`QA(question: str) -> answer: str` `translation(text: str, language: str) -> text: str`

Tool Use & Agent

- **Agent Definition**
  - Disagreement on what “agent” or “agentic” means
- **Requirements:**
  - *Probably:* Proactive use of tools
  - *Probably:* An iterative, multi-step process
  - *Maybe:* Interaction with the outside world

Tool Usage Performance

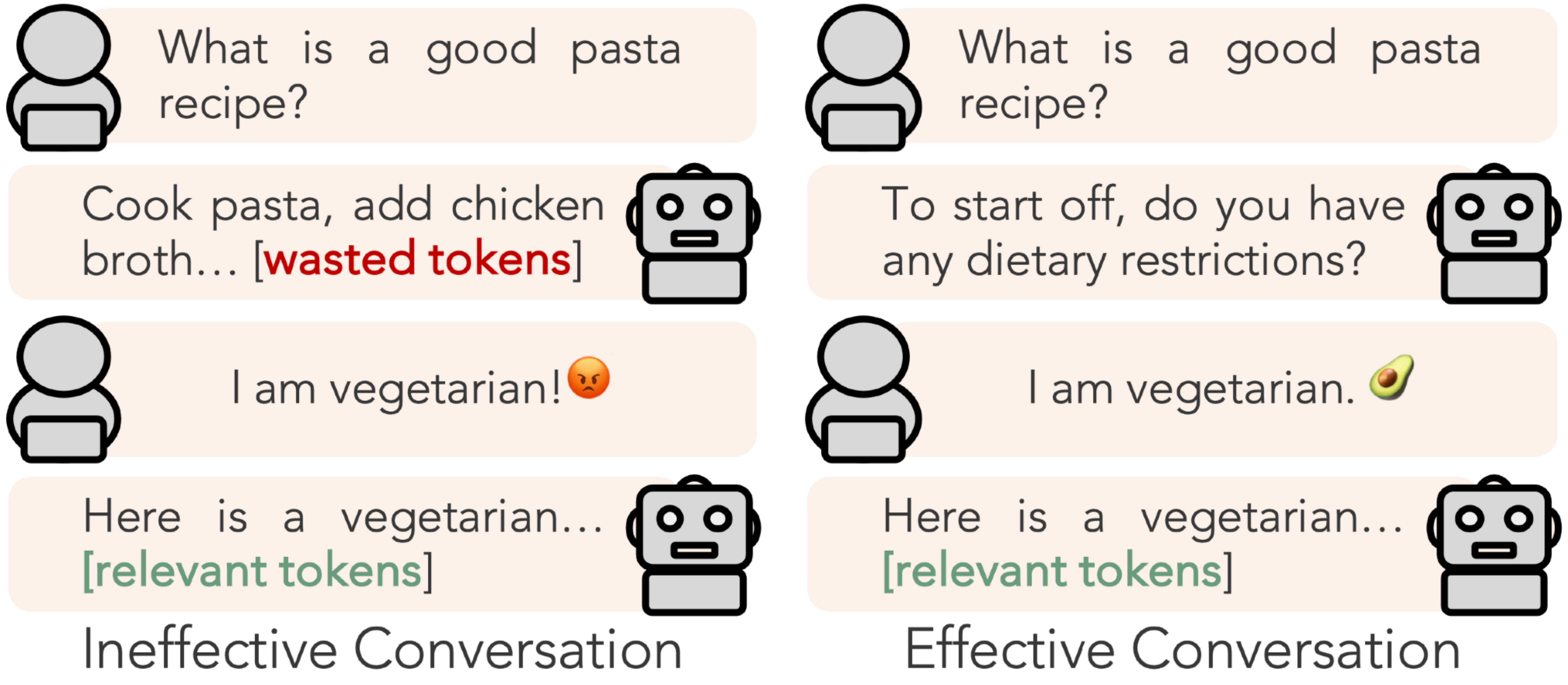


(Toolformer, Snihck et al. 2023)

Significantly Improving GPT's Performances

Can Models Ask Clarification Questions?

Similar to humans, but LMs (as-is) don't complain when the instructions are unclear



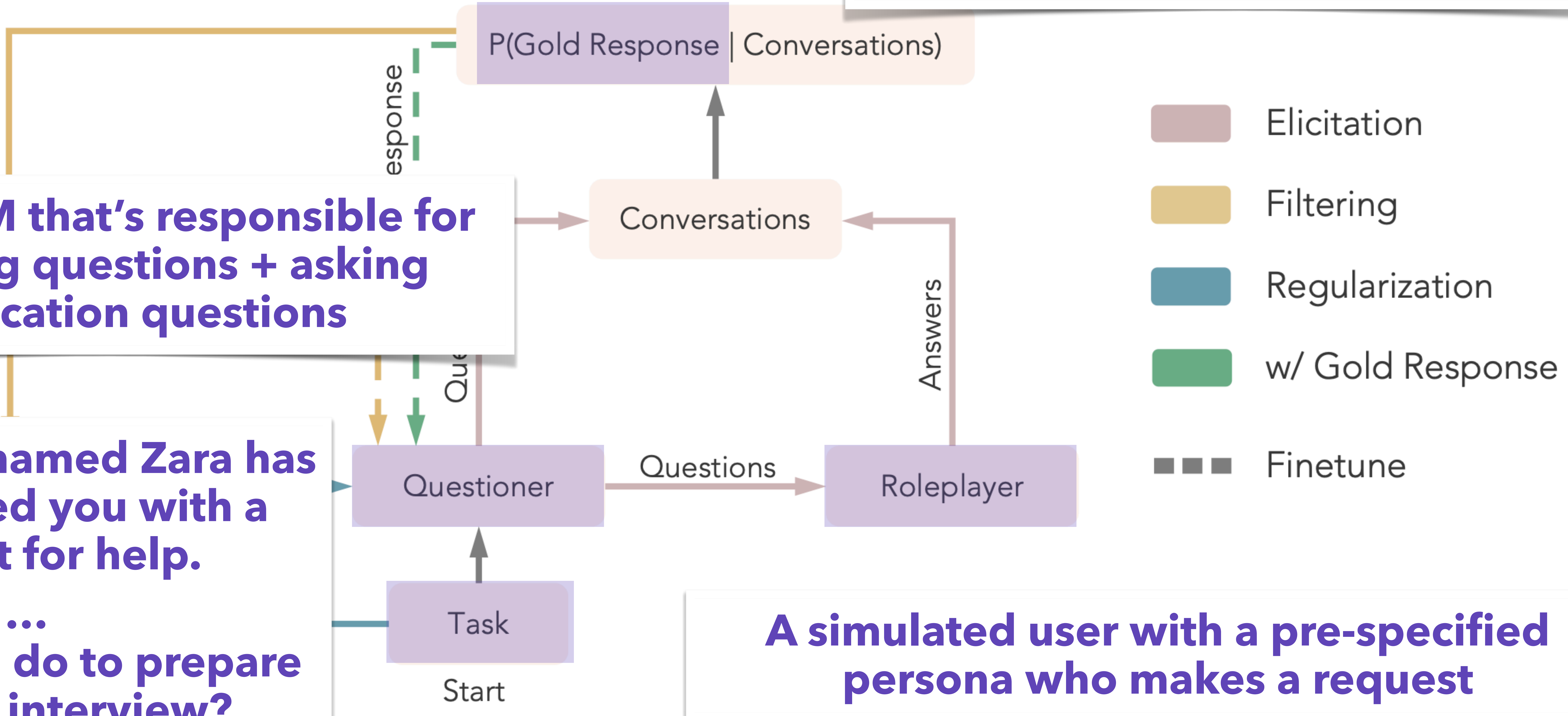
- **Task ambiguity**
- Teaching the model to ask questions that best **elicit a particular user's preferences**

(STaR-GATE, Andukuri et al. 2024)

STaR-GATE

(STaR-GATE, Andukuri et al. 2024)

Response generated by knowing both (Task + Persona)



The base LM that's responsible for answering questions + asking clarification questions

E.g., A user named Zara has approached you with a request for help.

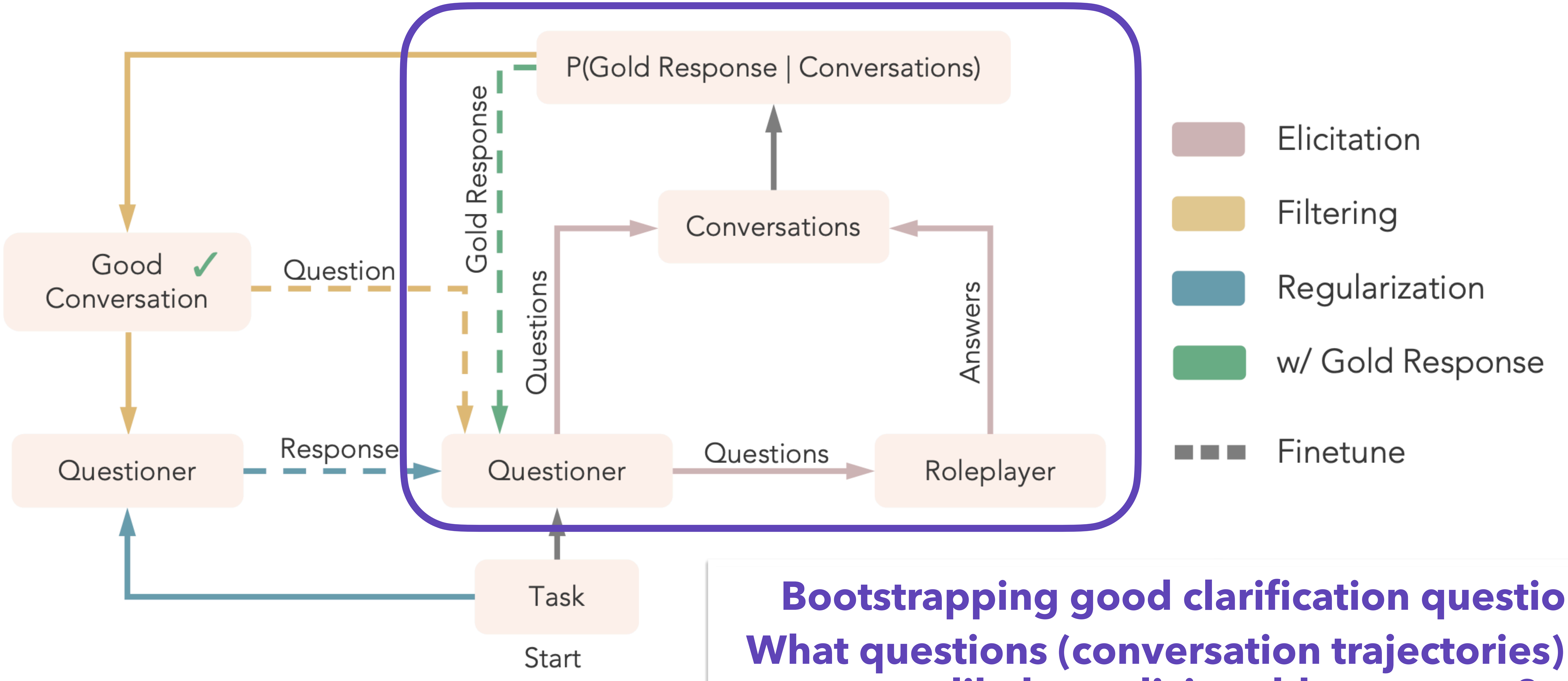
...

What must I do to prepare for a job interview?

A simulated user with a pre-specified persona who makes a request

STaR-GATE

(STaR-GATE, Andukuri et al. 2024)

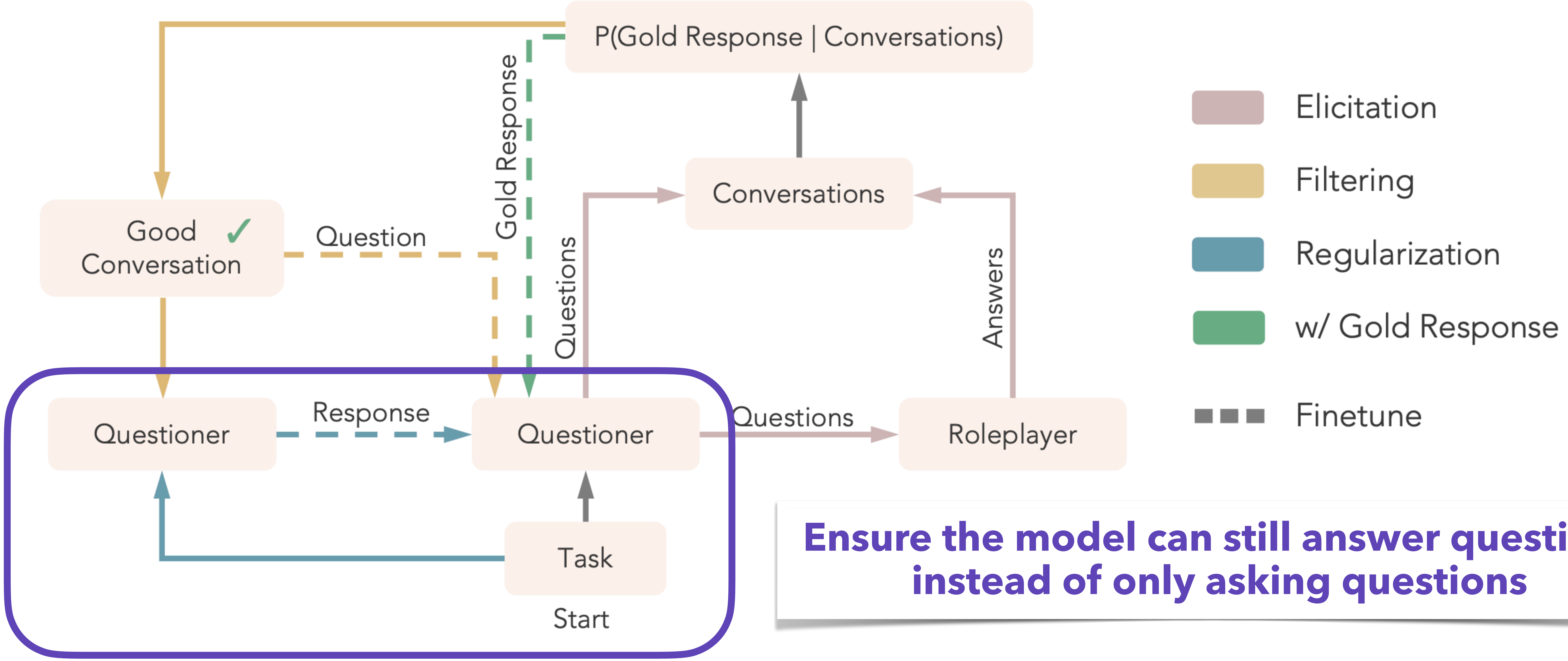


Bootstrapping good clarification question.

What questions (conversation trajectories) are most likely to elicit gold responses?

STaR-GATE

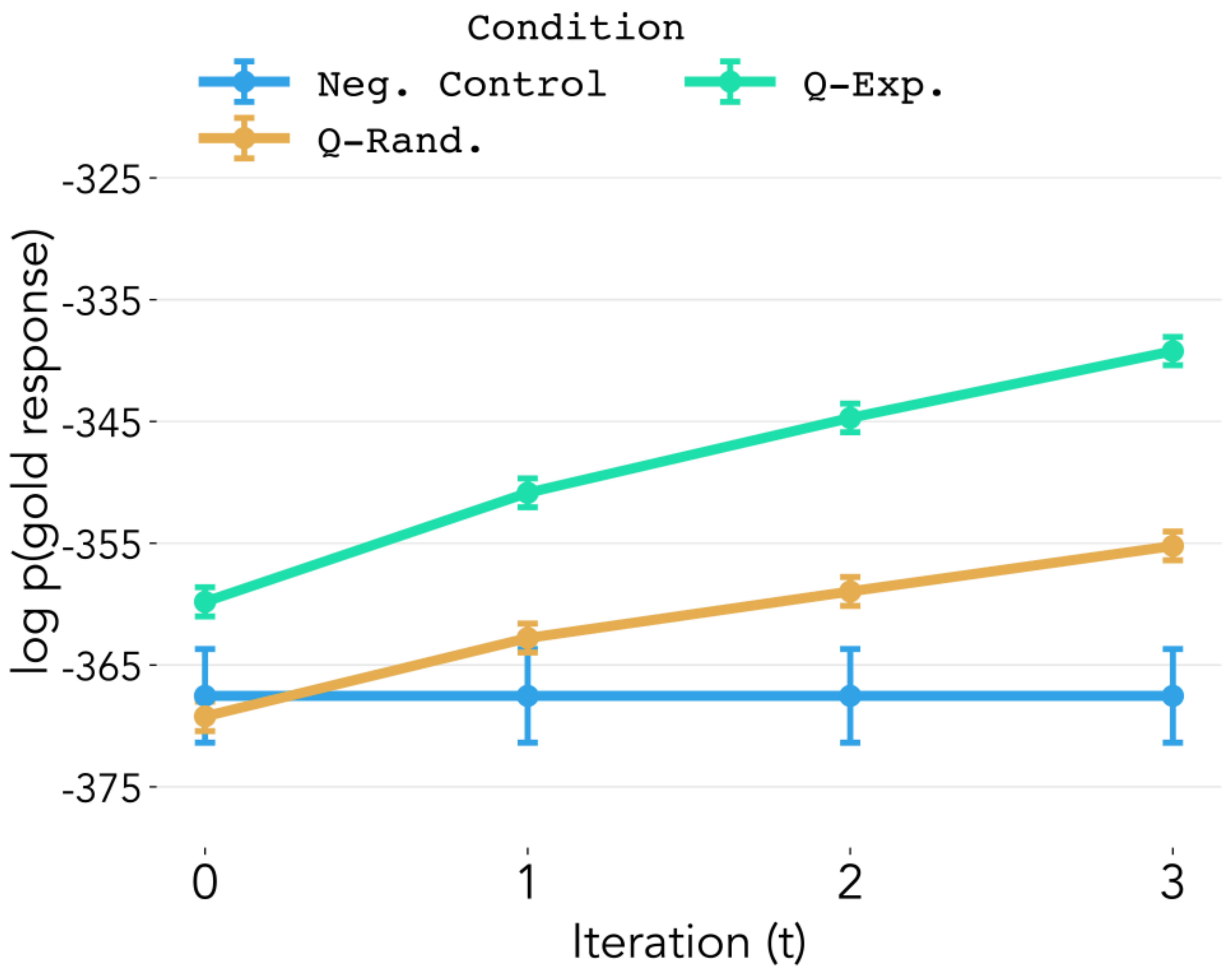
(STaR-GATE, Andukuri et al. 2024)



Ensure the model can still answer questions instead of only asking questions

Can Models Ask Clarification Questions?

(STaR-GATE, Andukuri et al. 2024)



[a] STaR-GATE [b] w/o Regularization

Natural Language Processing

CSE 447 @ UW

Knowledge Distillation

Guest Lecturer: Chan Young Park

Some slides adapted from: Charlie Dickens

★ **Basics of Knowledge Distillation**

Definition and Steps

★ **Types of Knowledge Distillation**

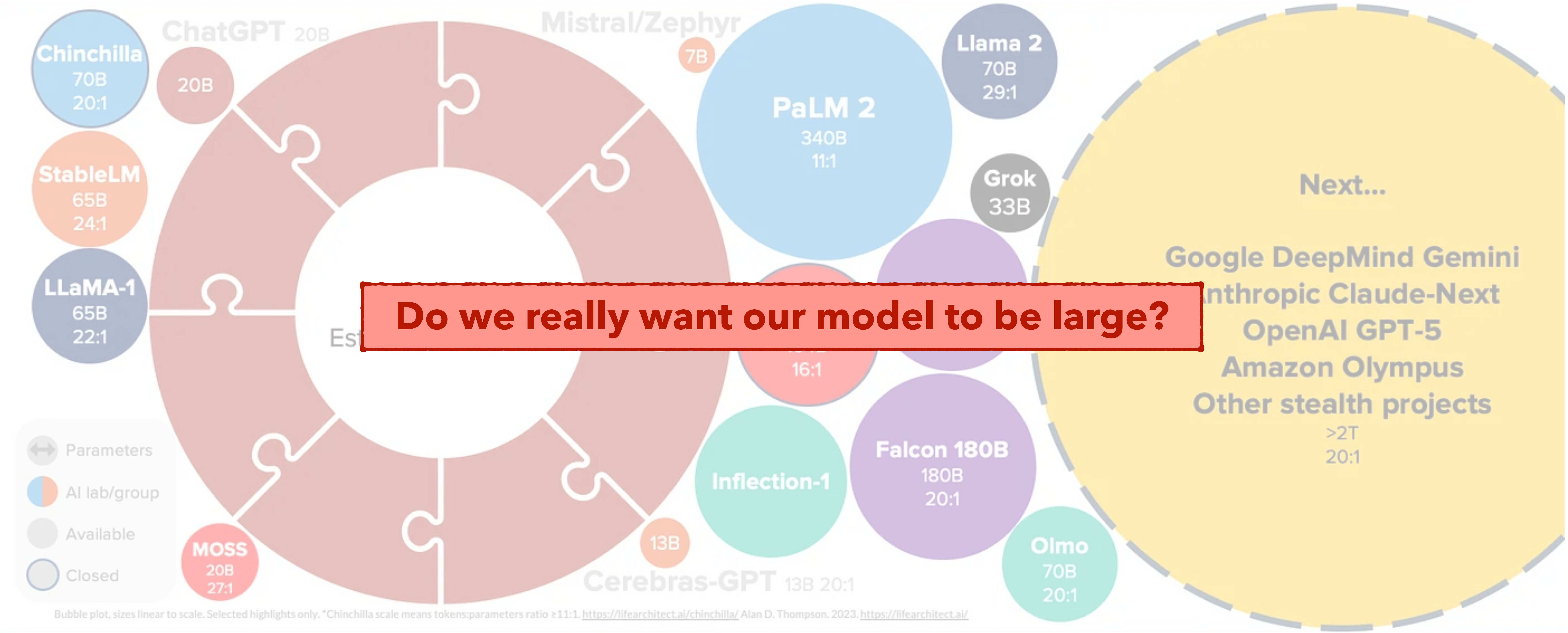
Labels, Representations, Synthetic Data, Feedback

★ **Advanced Knowledge Distillation**

Impossible Distillation

Basics of Knowledge Distillation:
Definition and Steps

Why Distillation?



Do we really want our model to be large?

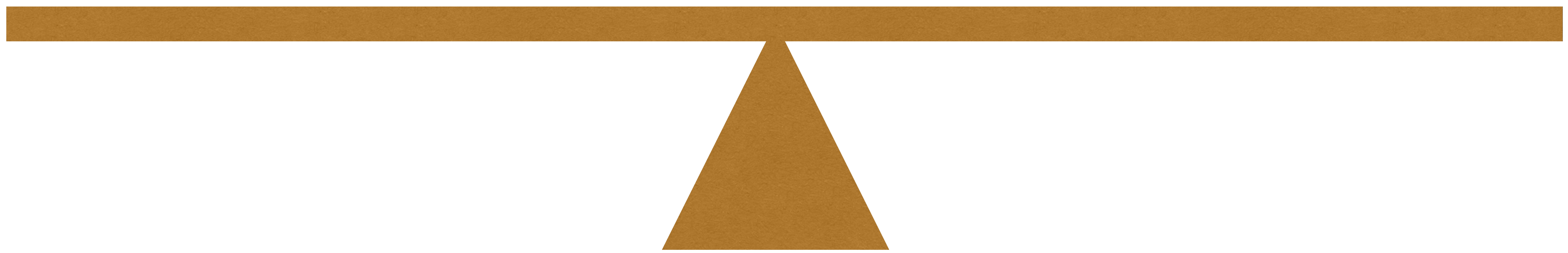
Size-Cost Trade-Off

Better Generalizability

Better Performance

Higher Latency

Higher Inference Cost



Bigger models are not always desirable

Ideally...

Fast response (low latency)

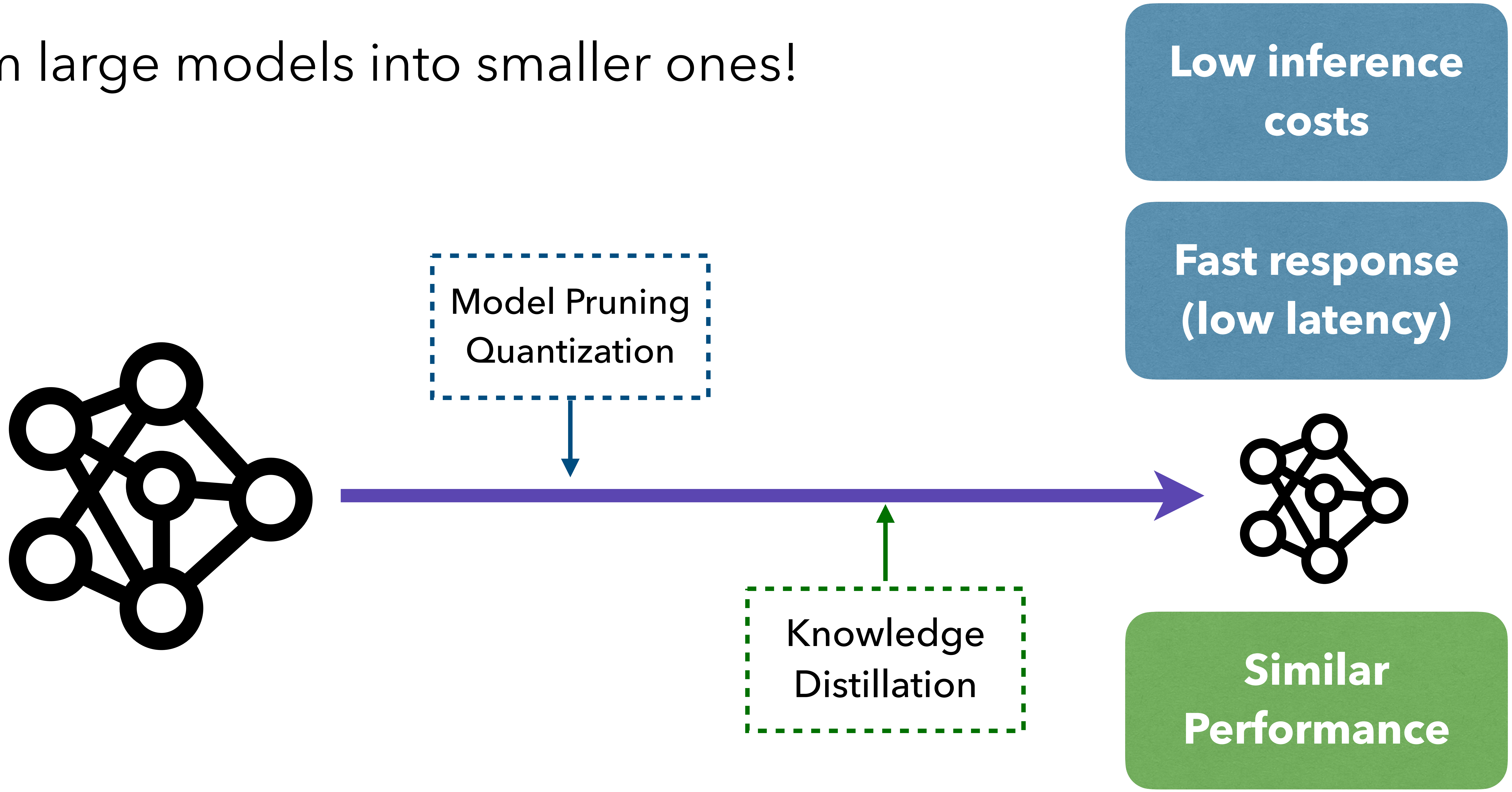
+

Low inference costs

While retaining similar performance as large models!

What we can do

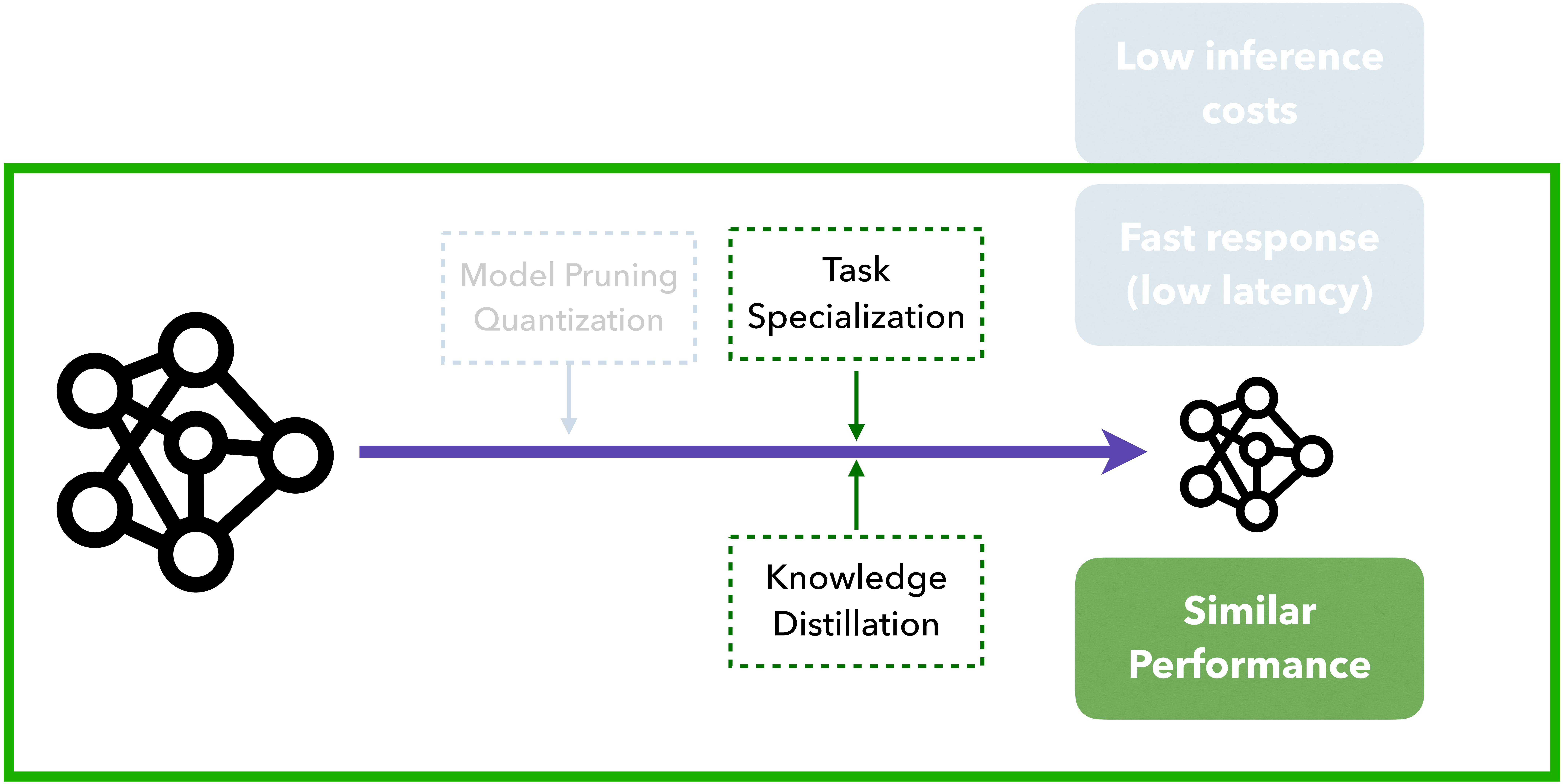
Transform large models into smaller ones!



What we can do

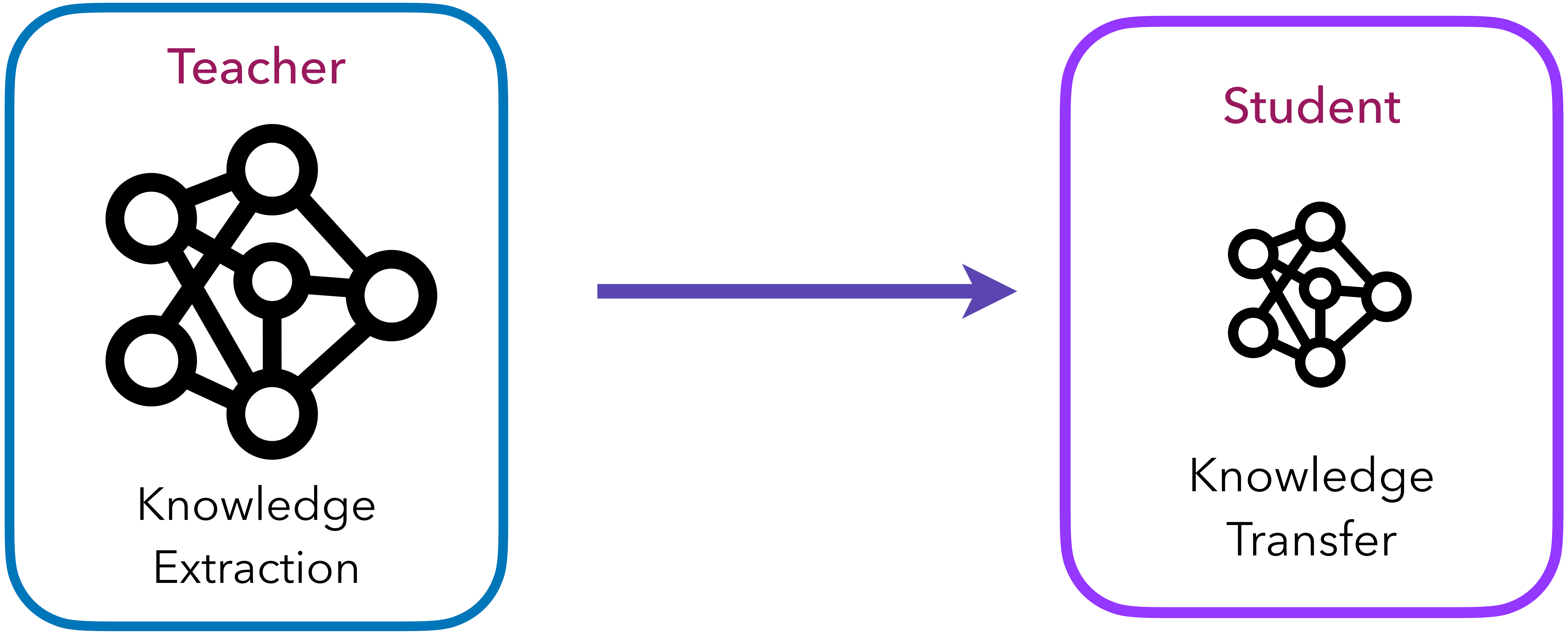
We often don't need to retrain everything!

Today's focus



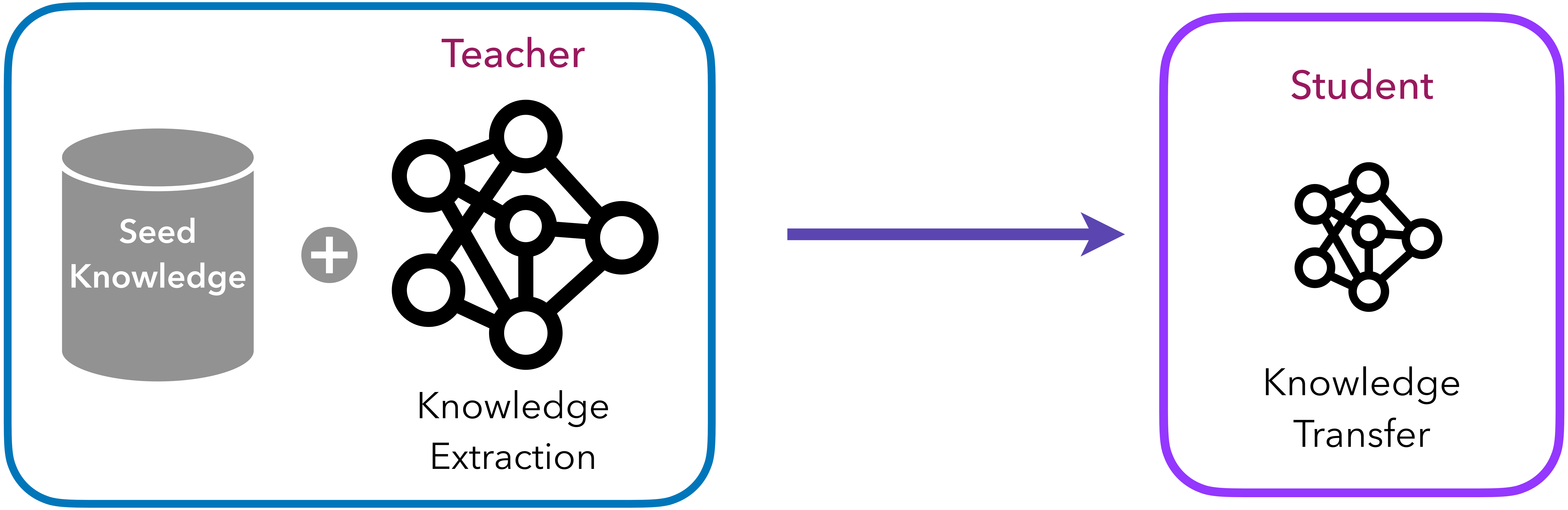
What is Knowledge Distillation?

1. **Knowledge Extraction** from a generalist model (the **teacher**)
2. **Transfer Knowledge** to a specialized model (the **student**)



What is Knowledge Distillation?

1. **Knowledge Extraction** from a generalist model (the **teacher**)
2. **Transfer Knowledge** to a specialized model (the **student**)

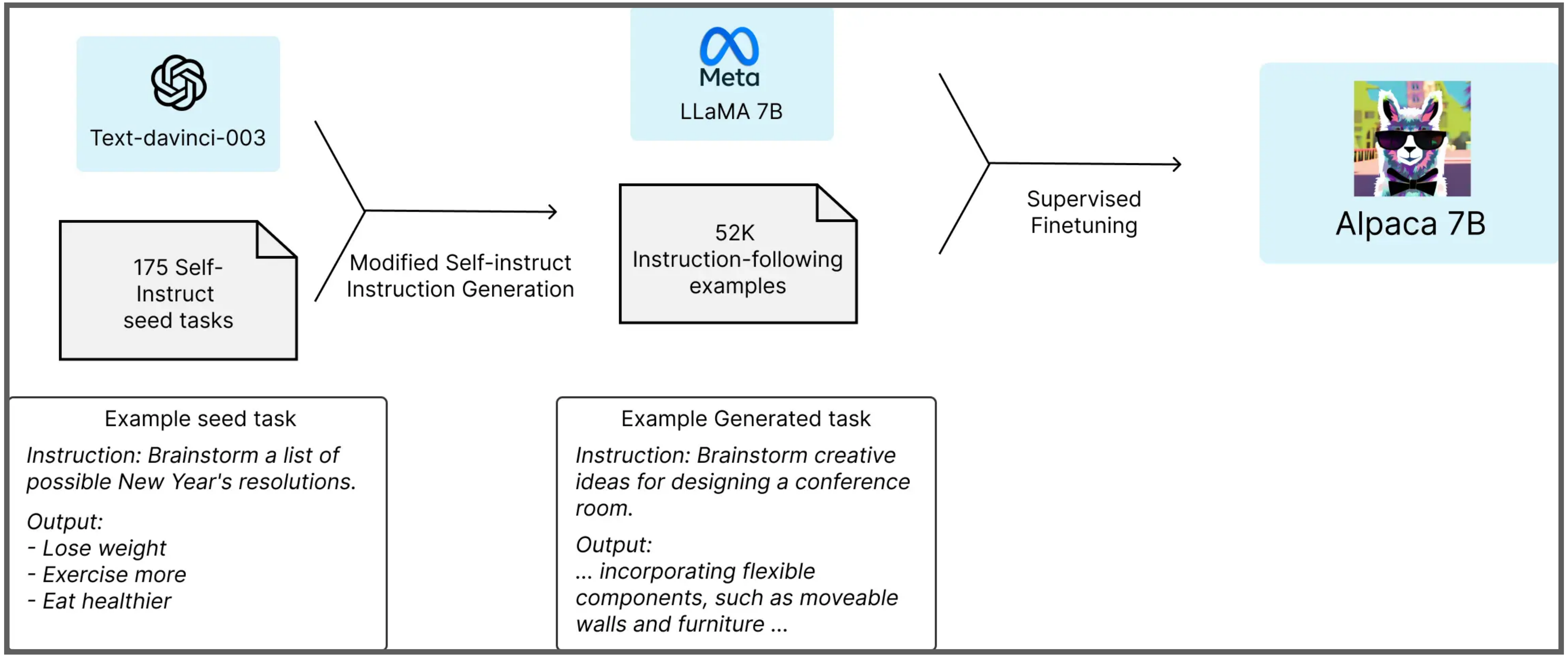


Examples of KD: Alpaca

Trained on 52K QA pairs generated by OpenAI's *text-davinci-003*

Took less than two months cost less than $600

Comparable to GPT 3.5



https://crfm.stanford.edu/2023/03/13/alpaca.html

Knowledge Extraction from LLMs

Identify target skills and domain	What to retain Classification, information extraction , Summarization, QA?
Curate seed knowledge	Select in-domain examples and create prompt templates
Generate teacher knowledge	What to Extract Labels, Synthetic data, hidden representations, feedback

Types of Knowledge Distillation:

Labels, Representations, Synthetic Data, and Feedback

The most basic KD: teacher labeling

Teacher provides supervision for student

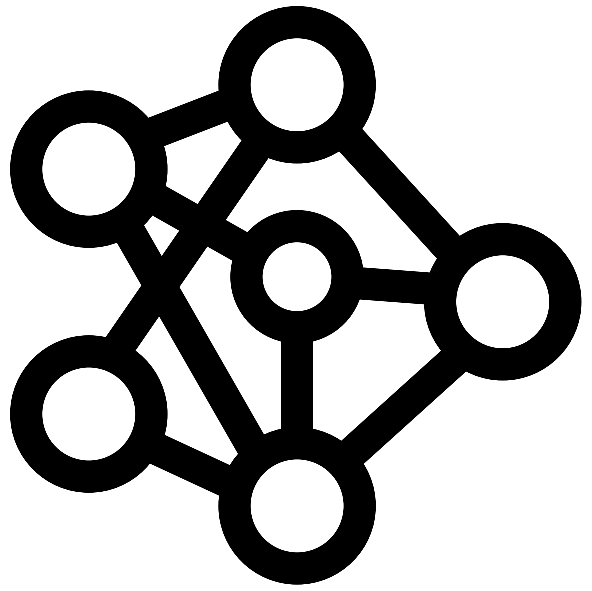
Target Skills/Domain
text classification
query categorization

Seed Knowledge
in-domain examples
input prompt for CLS

1. Knowledge Extraction

Help the user classify queries into 1 of 5 categories…
Query: “What is the capital of France?”

→ Teacher



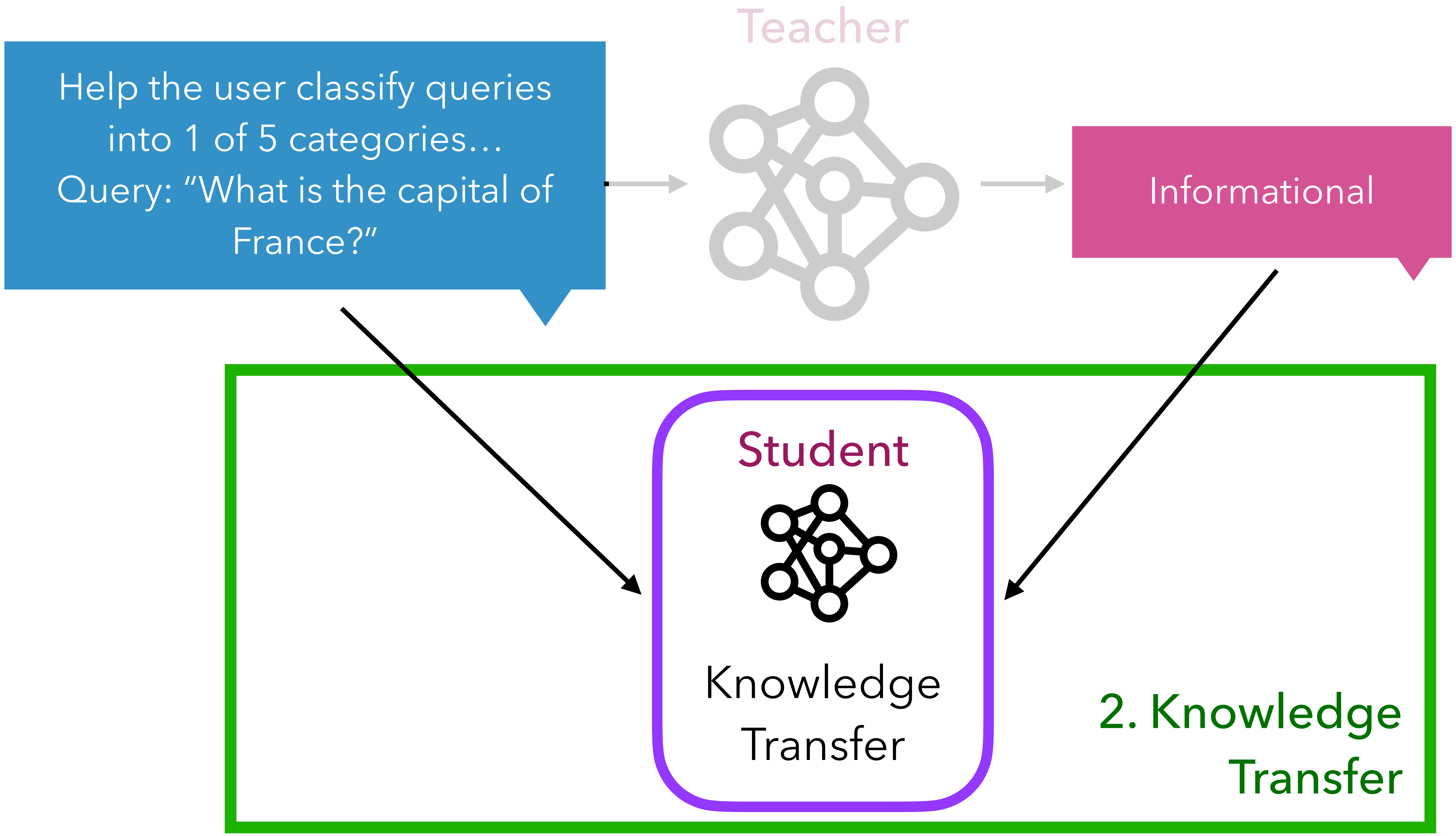
→ Informational

The most basic KD: teacher labeling

Teacher provides supervision for student

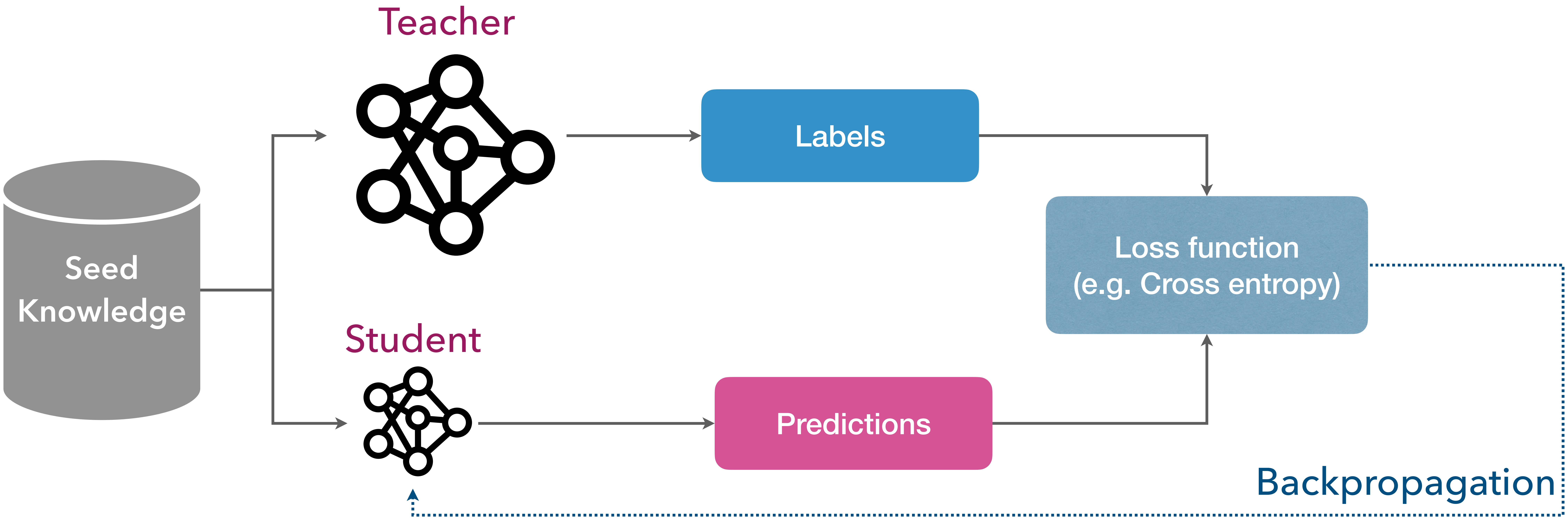
Target Skills/Domain
text classification
query categorization

Seed Knowledge
in-domain examples
input prompt for CLS



KD via hidden representations

Teacher provides supervision for student



Strengths: Soft-labels (logits) express uncertainty and teacher knowledge

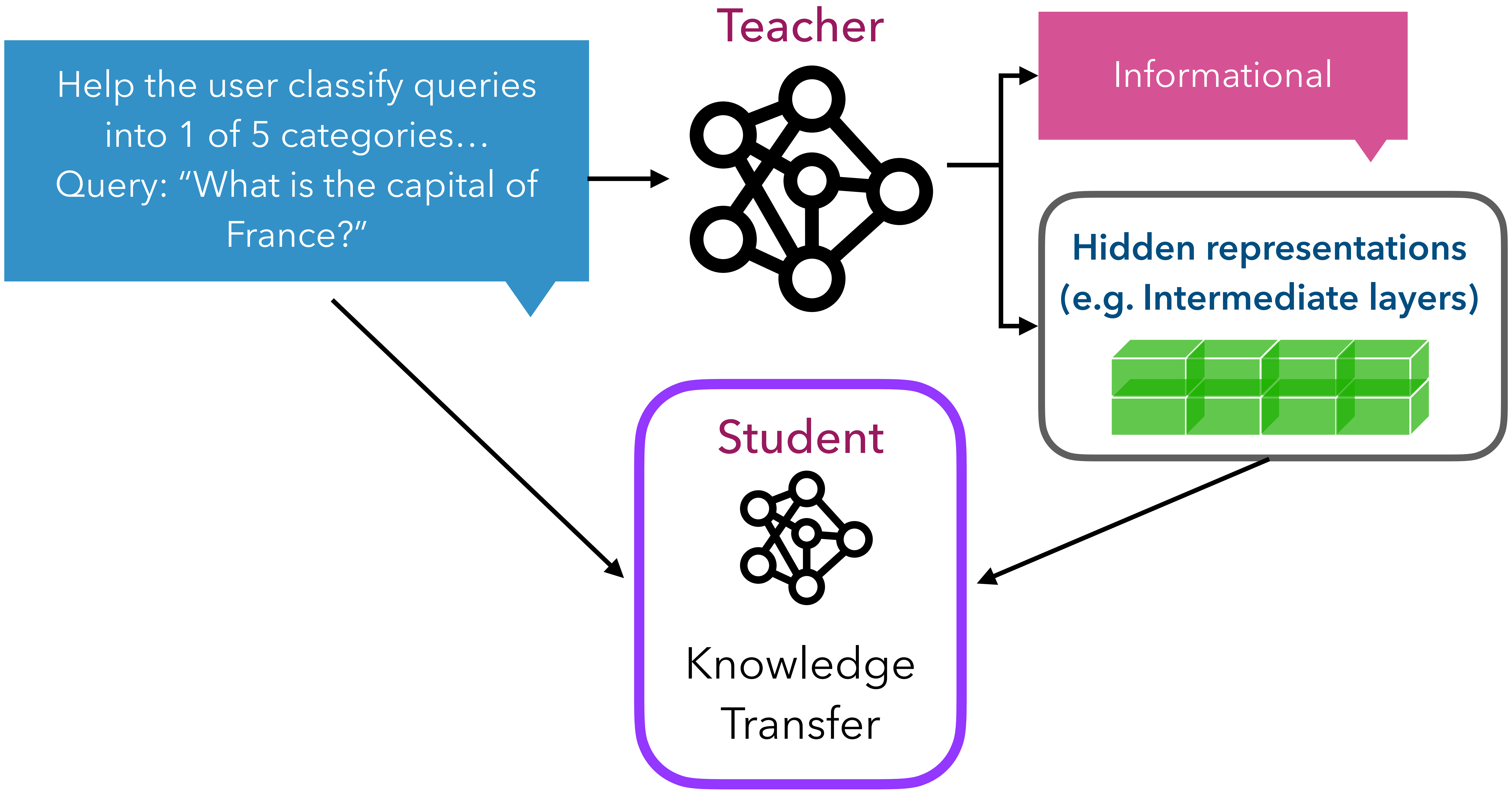
Weaknesses: Labels don't capture all of the rich knowledge of the teacher

KD via hidden representations

Teacher and student hidden representations are aligned

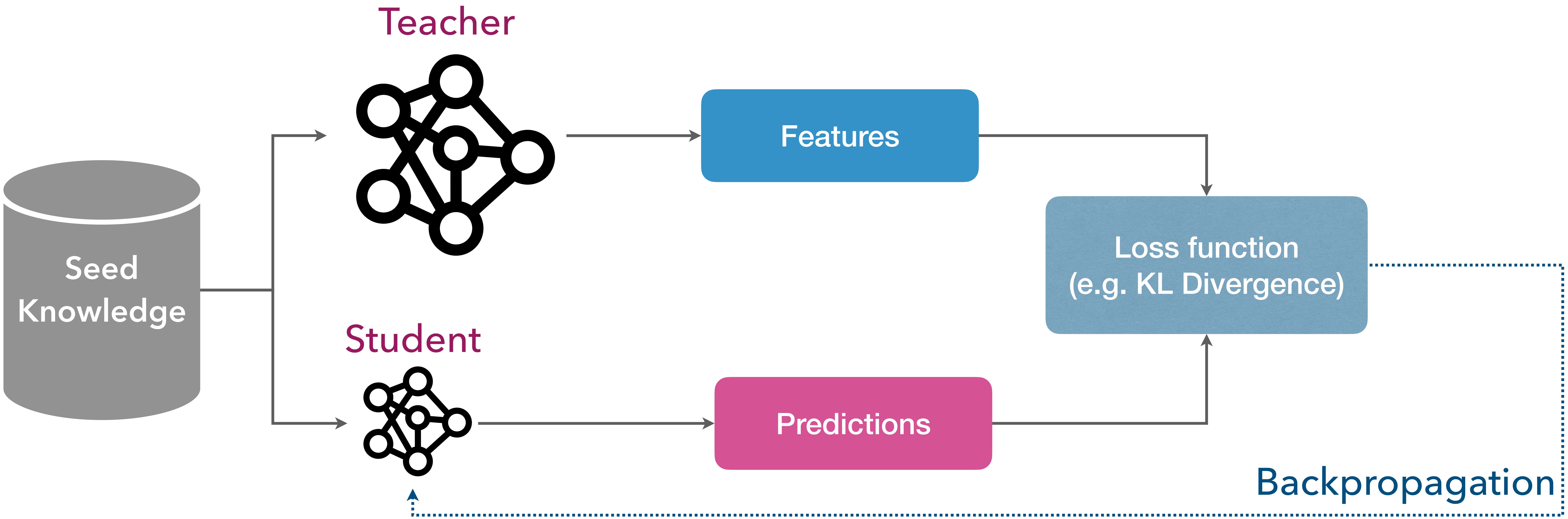
Target Skills/Domain
text classification
query categorization

Seed Knowledge
in-domain examples
input prompt for CLS



KD via hidden representations

Teacher and student hidden representations are aligned



Strengths: Hidden representations expressed nuanced understanding of task

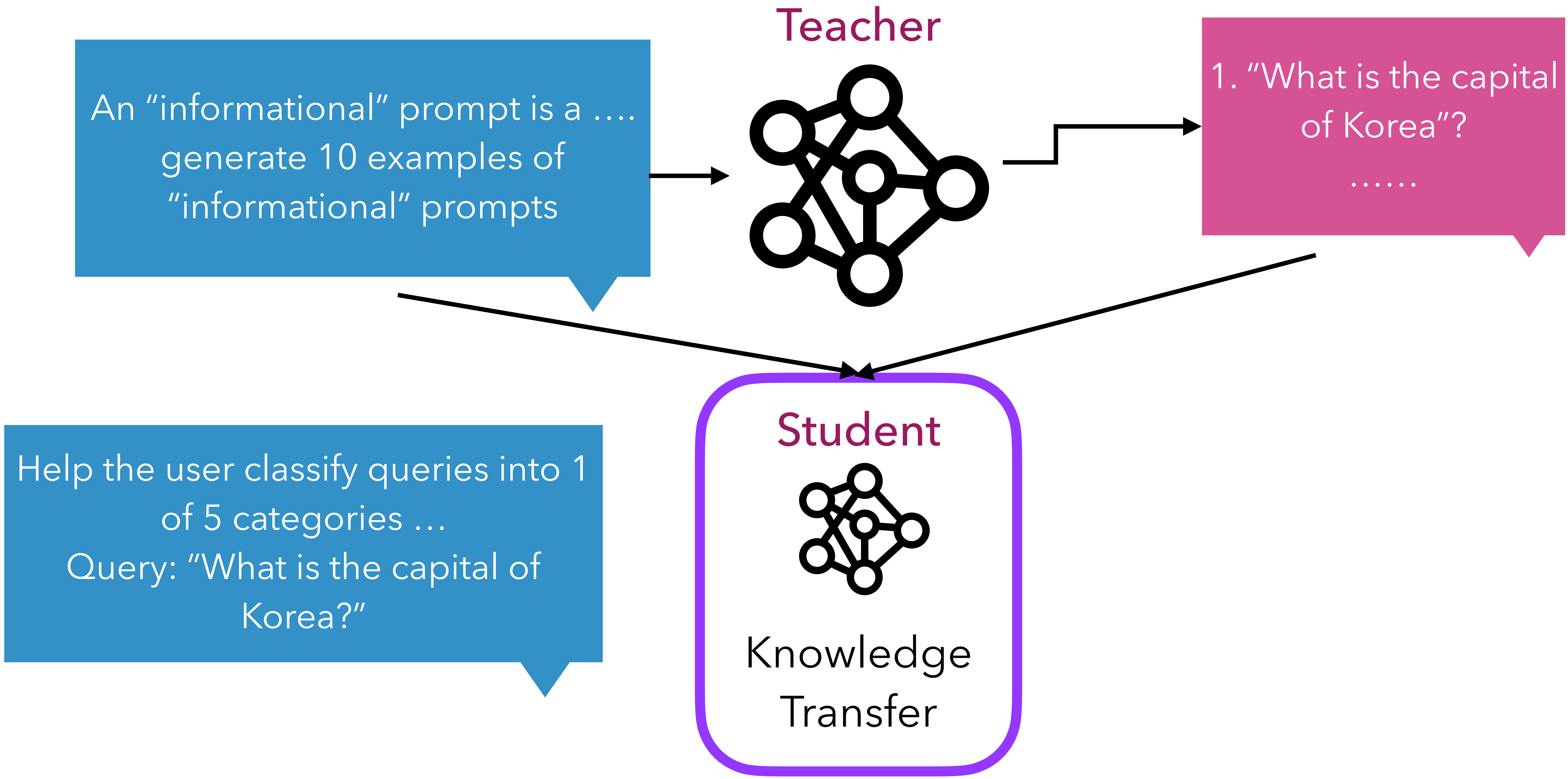
Weaknesses: Requires (un)labeled data source as seed knowledge

KD via synthetic data

Teacher expands the student training dataset

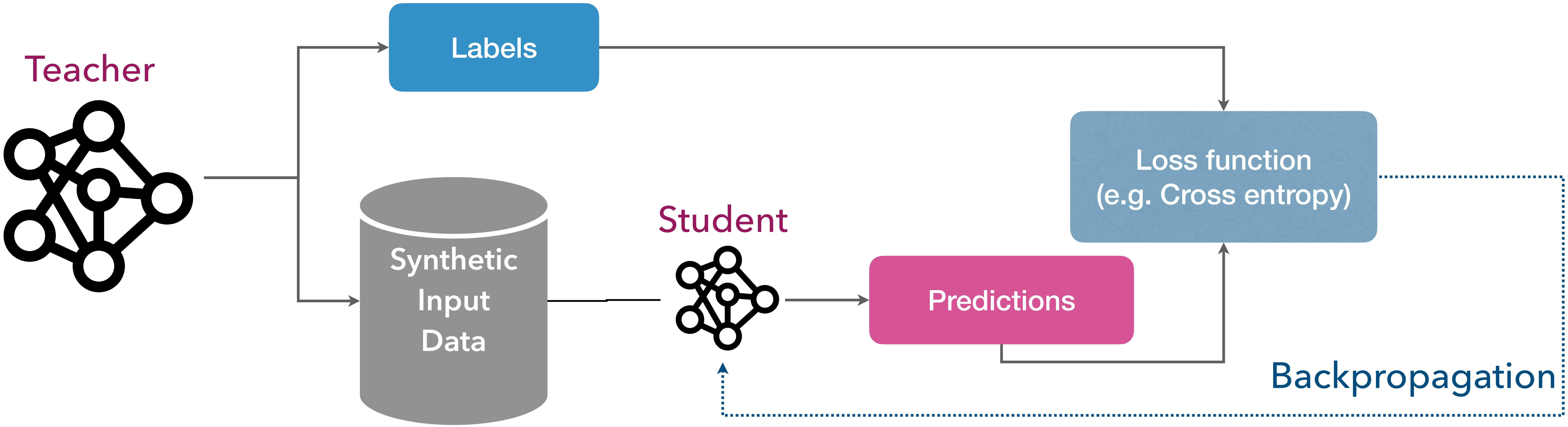
Target Skills/Domain
text classification
query categorization

Seed Knowledge
~~in-domain examples~~
data generation prompt
input prompt for CLS



KD via synthetic data

Teacher expands the student training dataset



Strengths: Leverage generation of teacher to overcome a lack of in-domain data

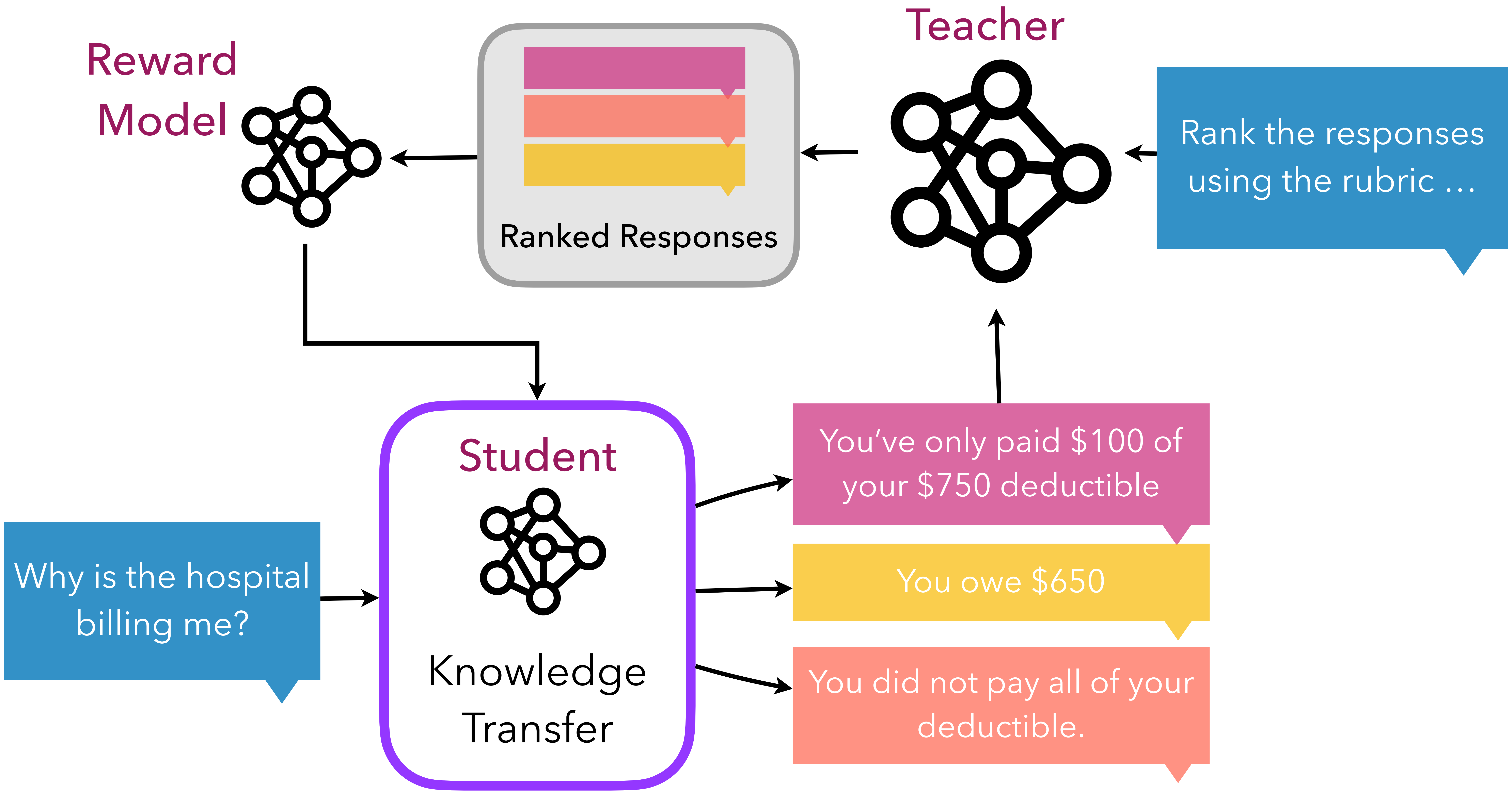
Weaknesses: Misalignment of synthetic and real-world data

KD via feedback

Teacher provides feedback on student generations

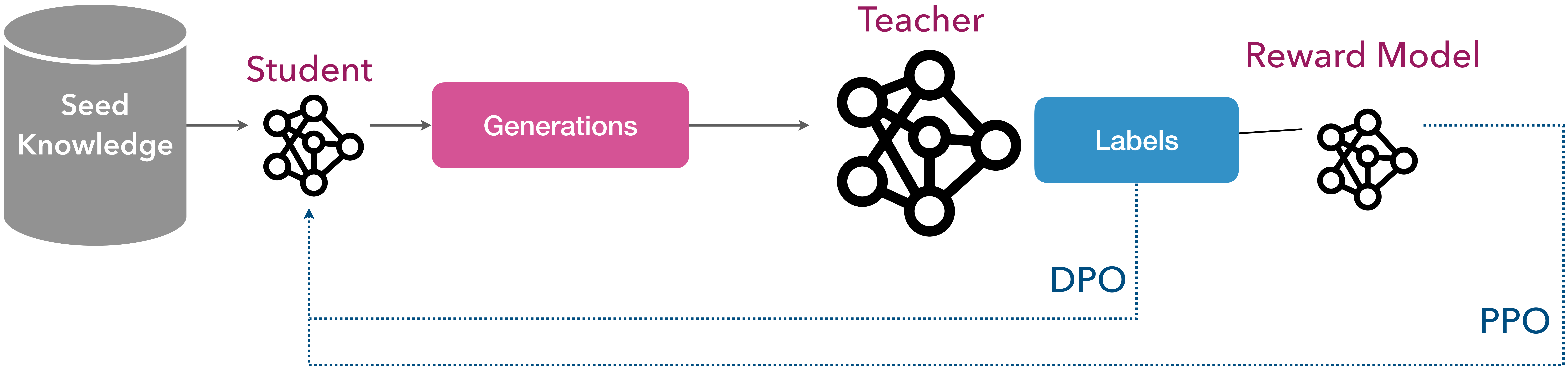
Target Skills/Domain
Health insurance QA

Seed Knowledge
in-domain examples
input prompt for CLS



KD via feedback

Teacher provides feedback on student generations



Strengths: Automate preference feedback process

Weaknesses: Risk of reinforcing teacher biases

Summary

What is knowledge distillation:
Extracting task specific knowledge from a generalist teacher model and transferring it to a specialized student model

Steps for knowledge extraction:
1) identify large skills, 2) curate seed knowledge, 3) generate knowledge

Types of knowledge extraction:
1) teacher labeling, 2) hidden representations, 3) synthetic data, and 4) feedback

Challenges and Best Practices

Teacher Quality

Performance is limited by the teacher

Need fine-grained **evaluations of potential teachers** to understand teacher capabilities

+ also open-source vs. closed limits the types of KD you can use

Data Quality

Data Quality is vital for success

Data curation for seed knowledge is important for effective transfer

If unlabeled data is scarce, try multi-task student learning

Advanced Knowledge Distillation:
Impossible Distillation

Impossible Distillation

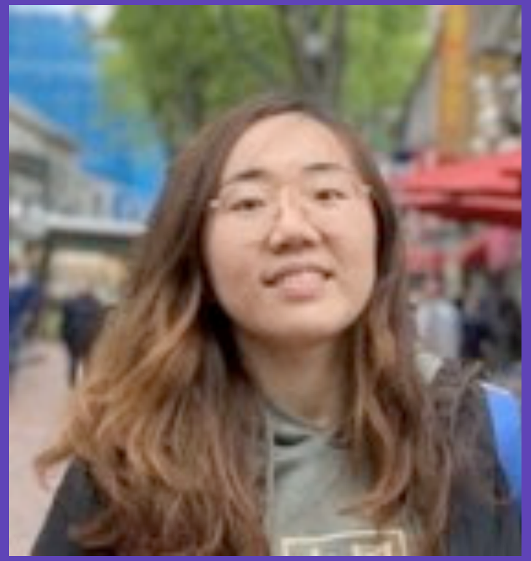
from Low-quality Model to High-Quality Dataset & Model
for Summarization and Paraphrasing

— NAACL 2024 —

Jaehun Jung

Peter West

Liwei Jiang



Faeze Brahman

Ximing Lu

Jillian Fisher

Taylor Sorensen

Yejin Choi

winning recipe = extreme-scale pre-training + RLHF at scale

GPT-2

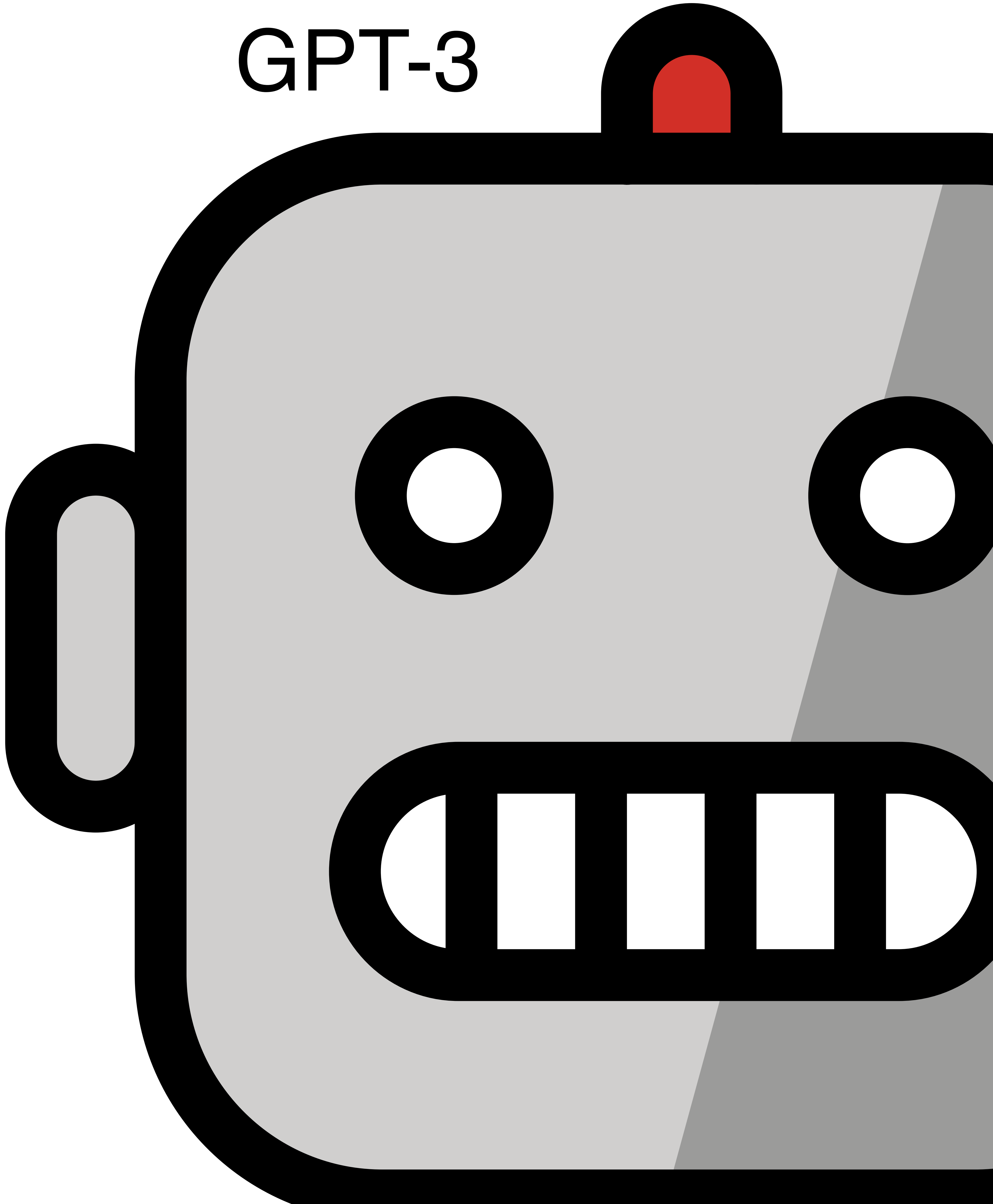
Low-quality, small model

???

High-quality, small model

VS

GPT-3



How is that even possible when imitating from proprietary LLMs are supposedly hopeless?

True for the particularity of their experimental settings, but one must not generalize beyond what the paper showed:

— factual QA is especially hard to distill
— generalist vs specialist

The False Promise of Imitating Proprietary LLMs

Arnav Gudibande* UC Berkeley arnavg@berkeley.edu
Eric Wallace* UC Berkeley ericwallace@berkeley.edu
Charlie Snell* UC Berkeley csnell22@berkeley.edu
Xinyang Geng UC Berkeley young.geng@berkeley.edu
Hao Liu UC Berkeley hao.liu@berkeley.edu
Pieter Abbeel UC Berkeley pabbeel@berkeley.edu
Sergey Levine UC Berkeley svlevine@berkeley.edu
Dawn Song UC Berkeley dawnsong@berkeley.edu

Are small LMs completely out of league?

https://twitter.com/EmojiMashupBot/status/1266262982406729730

Hope: Task-specific Symbolic Knowledge Distillation works!

Symbolic Knowledge Distillation: from General Language Models to Commonsense Models

Peter West[†‡*] **Chandra Bhagavatula**[‡] **Jack Hessel**[‡] **Jena D. Hwang**[‡]
Liwei Jiang[†‡] **Ronan Le Bras**[‡] **Ximing Lu**[†‡] **Sean Welleck**[†‡] **Yejin Choi**[†‡*]
[†]Paul G. Allen School of Computer Science & Engineering, University of Washington
[‡]Allen Institute for Artificial Intelligence

LLM-Planner: Few-Shot Grounded Planning for Embodied Agents with Large Language Models

Chan Hee Song, The Ohio State University, song.1855@osu.edu
Jiaman Wu, The Ohio State University, wu.5686@osu.edu
Clayton Washington, The Ohio State University, washington.534@osu.edu
Brian M. Sadler, DEVCOM ARL, brian.m.sadler6.civ@army.mil
Wei-Lun Chao, The Ohio State University, chao.209@osu.edu
Yu Su, The Ohio State University, su.806@osu.edu

Specializing Smaller Language Models towards Multi-Step Reasoning

Yao Fu[♠] **Hao Peng**[♣] **Litu Ou**[♠] **Ashish Sabharwal**[♣] **Tushar Khot**[♣]

Teaching Small Language Models to Reason

Lucie Charlotte Magister[*], University of Cambridge, lcm67@cam.ac.uk
Jonathan Mallinson, Google Research, jonmall@google.com
Jakub Adamek, Google Research, enkait@google.com
Eric Malmi, Google Research, emalmi@google.com
Aliaksei Severyn, Google Research, severyn@google.com

Textbooks Are All You Need

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, Yuanzhi Li
Microsoft Research

Orca: Progressive Learning from Complex Explanation Traces of GPT-4

Subhabrata Mukherjee[*†], **Arindam Mitra**[*]
Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, Ahmed Awadallah
Microsoft Research

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

Cheng-Yu Hsieh[1*], **Chun-Liang Li**[2], **Chih-Kuan Yeh**[3], **Hootan Nakhost**[2], **Yasuhisa Fujii**[3], **Alexander Ratner**[1], **Ranjay Krishna**[1], **Chen-Yu Lee**[2], **Tomas Pfister**[2]
[1]University of Washington, [2]Google Cloud AI Research, [3]Google Research
cydhsieh@cs.washington.edu

Our task in focus: learning to "abstract" in language

✨ In NLP: ~ "sentence summarization" ✨

Mission Impossible:🔥Learn to "summarize sentences" 🔥

- without extreme-scale pre-training
- without RL with human feedback at scale
- without supervised datasets at scale

AI is as good as the data it was trained on

We will build on …

Symbolic Knowledge Distillation

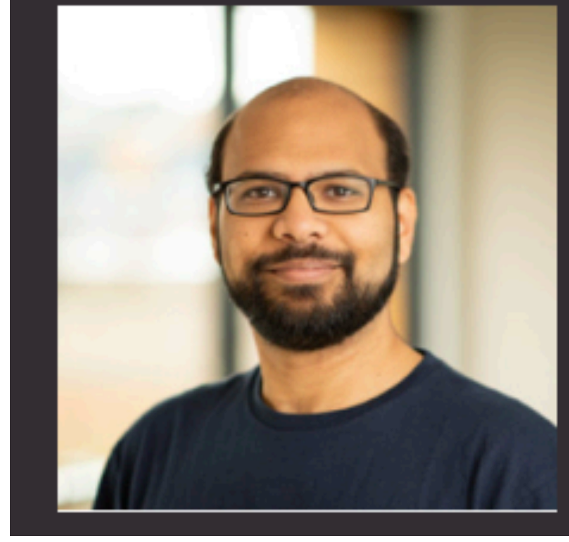
From General Language Models to Commonsense Models

— NAACL 2022 —

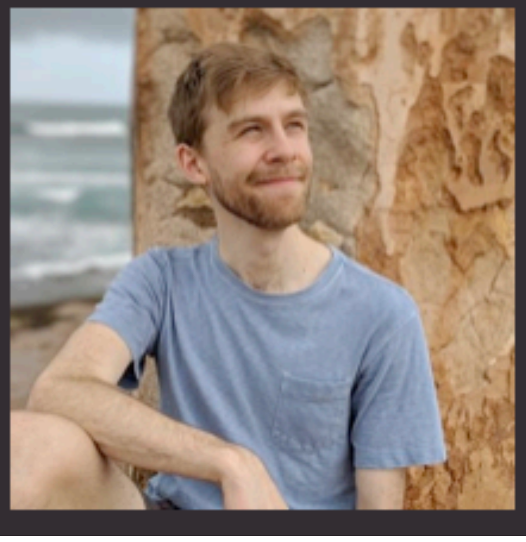
Peter West

New:
ATOMIC-10x
COMET-distill

Chandra Bhagavatula

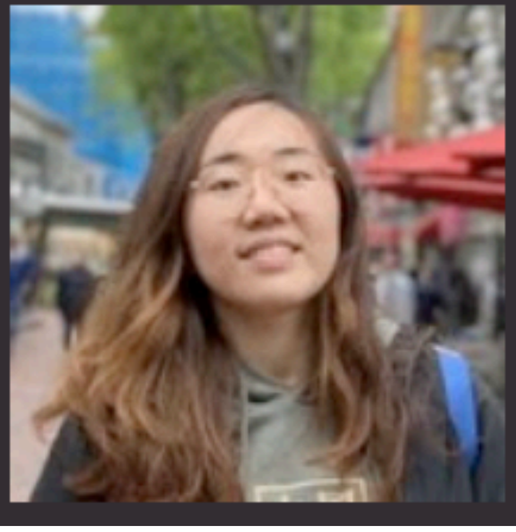


Jack Hessel

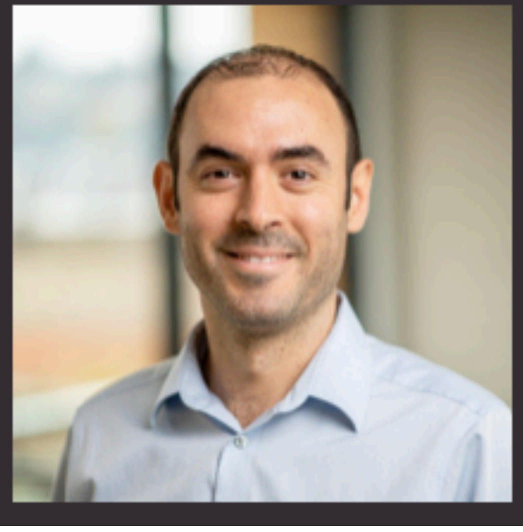


Jena Hwang

Liwei Jiang



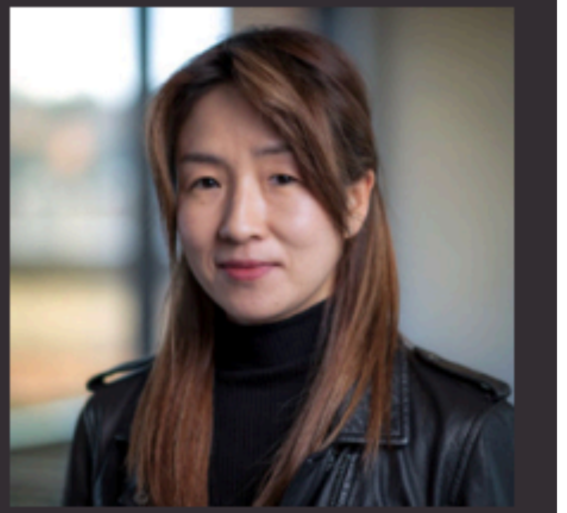
Ronan Le Bras



Ximing Lu

Sean Welleck

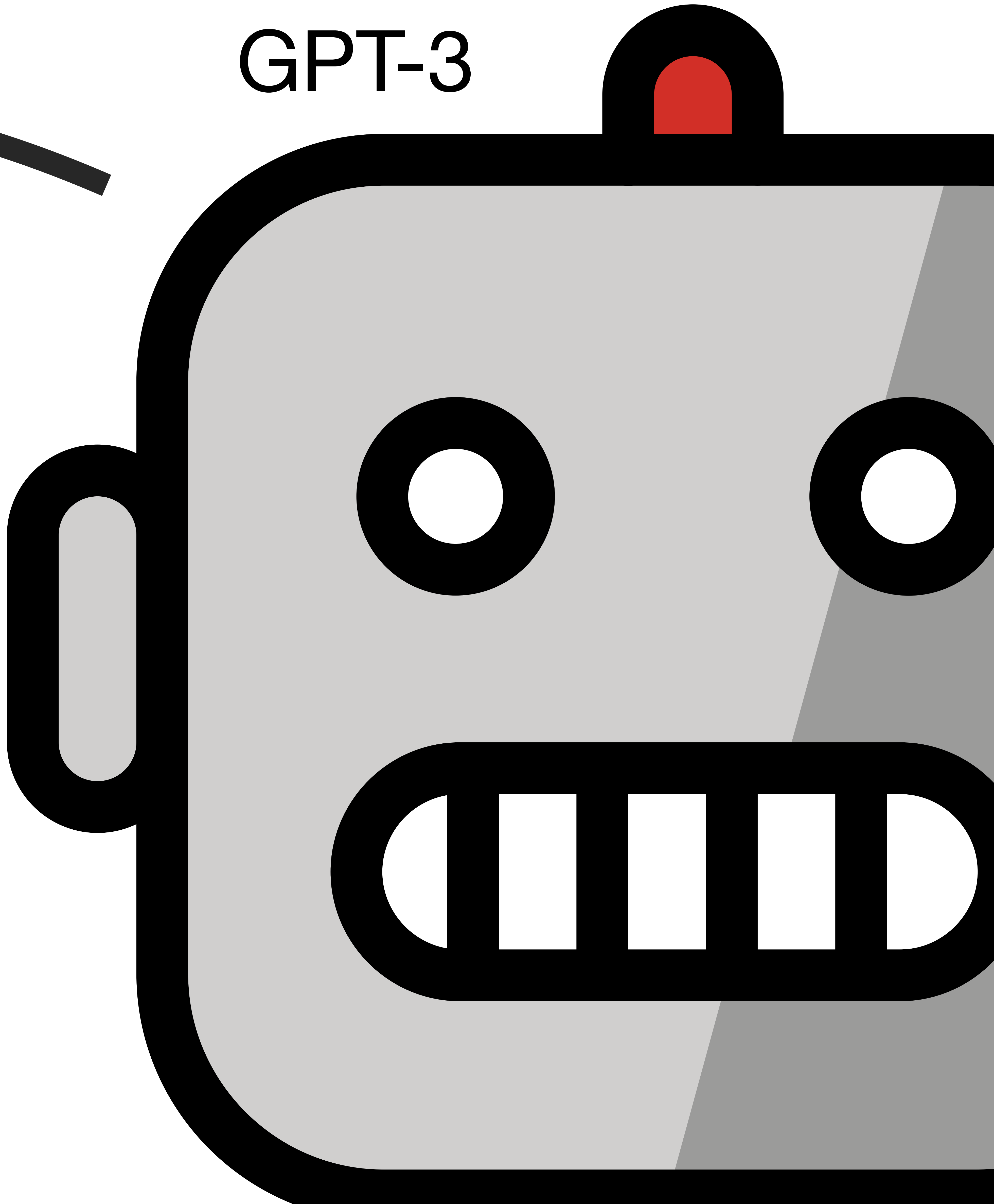
Yejin Choi



Symbolic Knowledge Distillation

Few-shot generate / Filter

GPT-3



ATOMIC^{10X}:
High-quality Commonsense KG

Fine-tune

COMET^{DISTILL}: High-quality, small commonsense model

Impossible Distillation

GPT-2

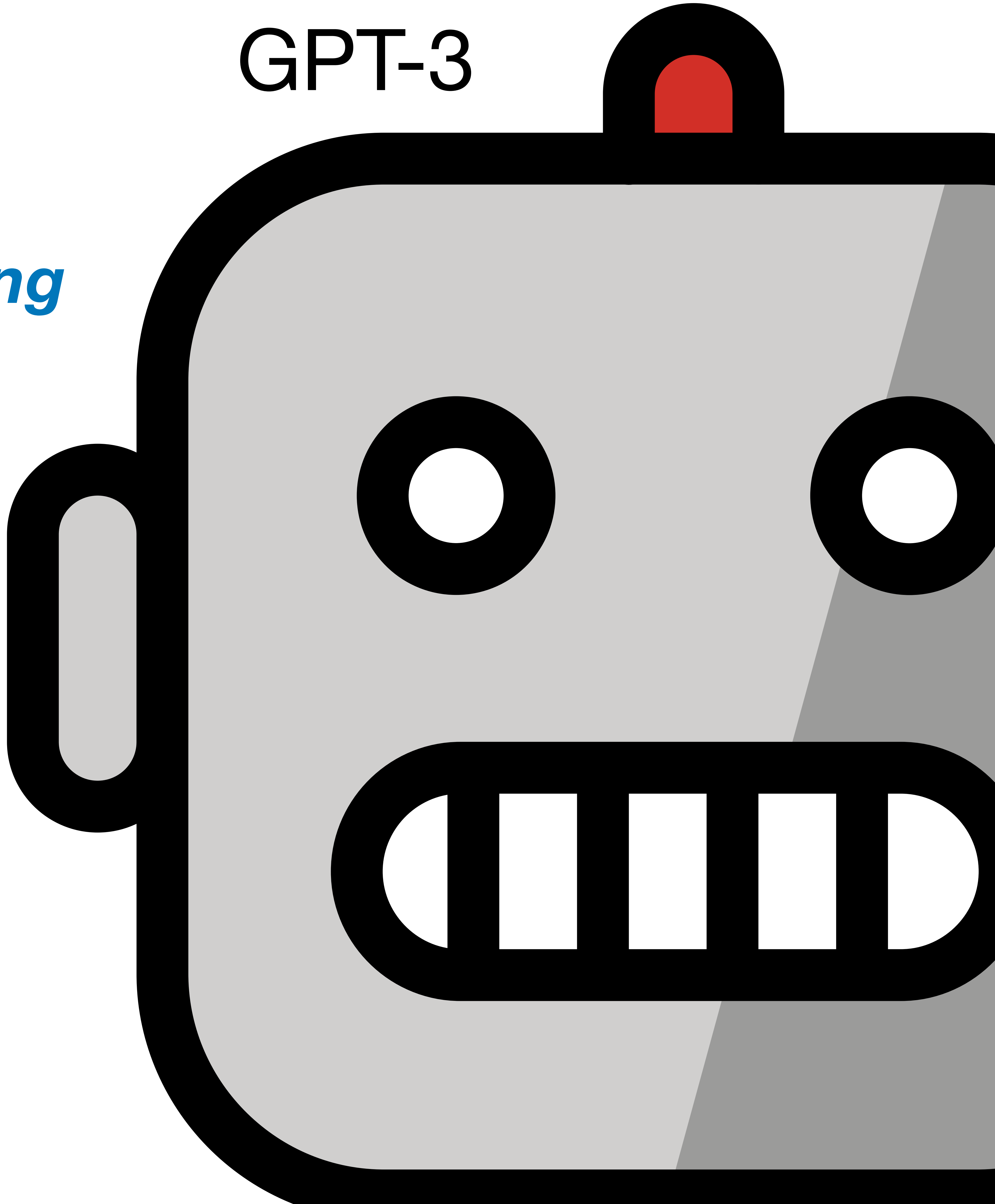
Low-quality, small model

+ Constrained Decoding
+ Off-the-shelf Filters

High-quality Task Dataset

High-quality, small model

GPT-3



When GPT-2 is prompted to summarize…
it generates *< 0.1% correct pairs*!

With Lexically-constrained Decoding,
now gener

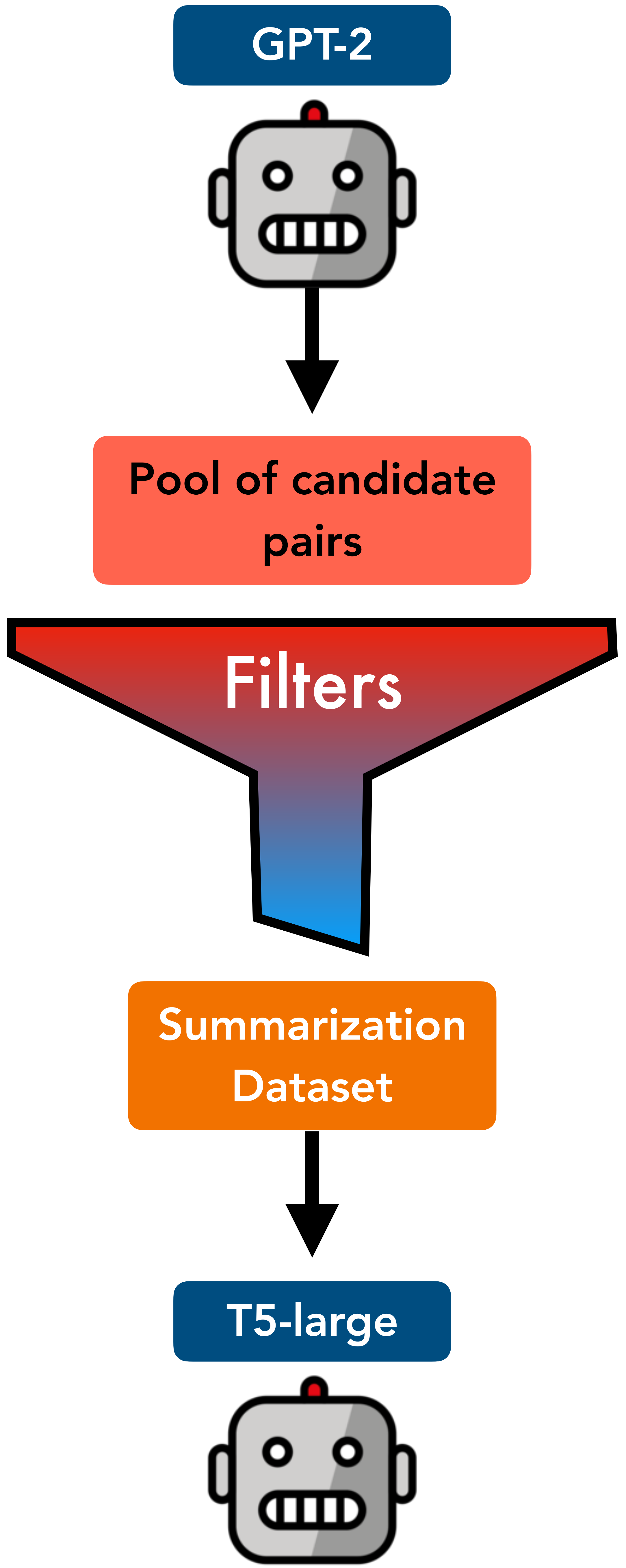
NEURO-LOGIC Decoding* with keywords: *NLP, field, study

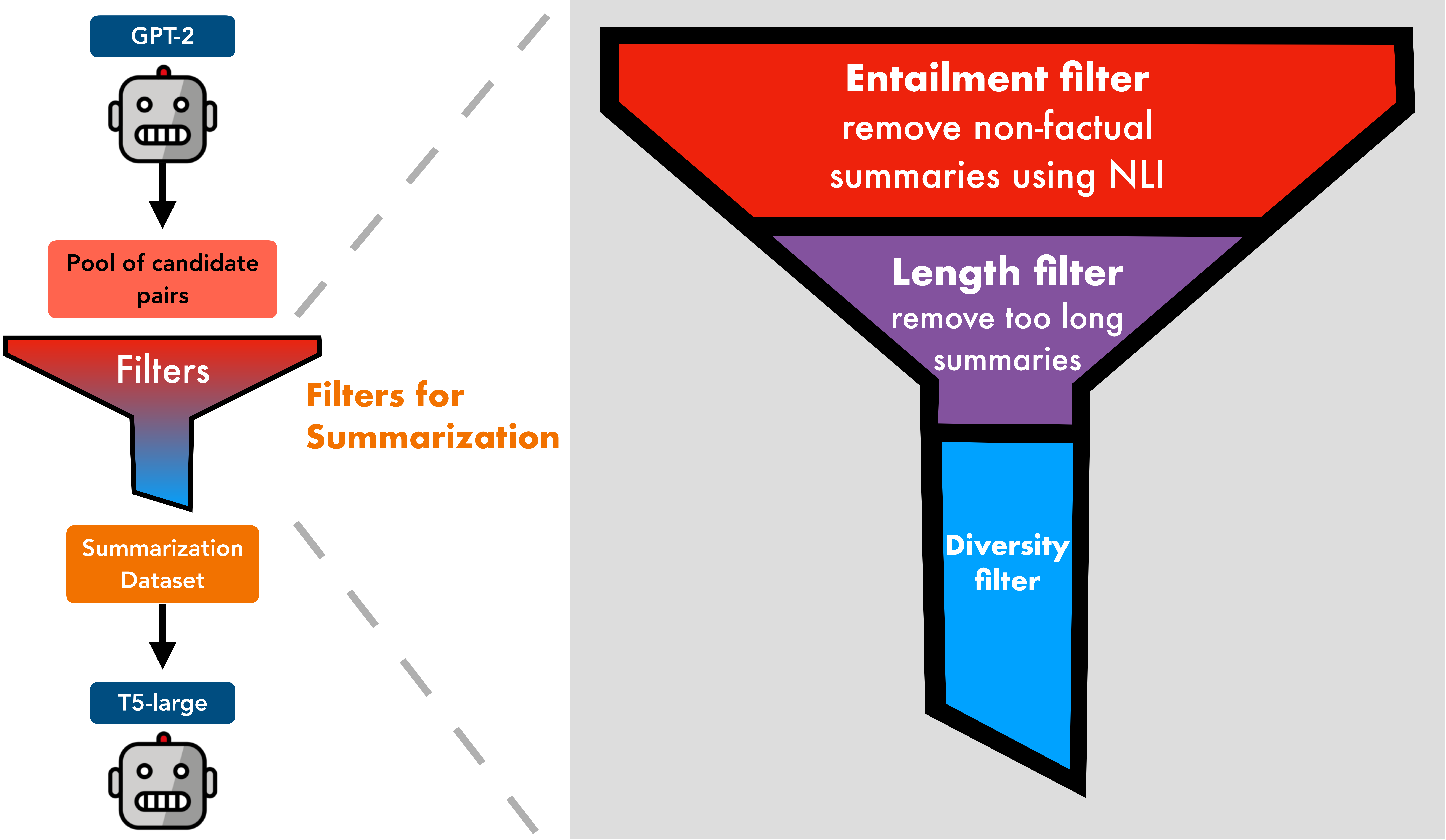
Allowing GPT-2 to generate *text to summarize*,

Intuition: Next sentences generated from same left context are likely to be semantically coherent!

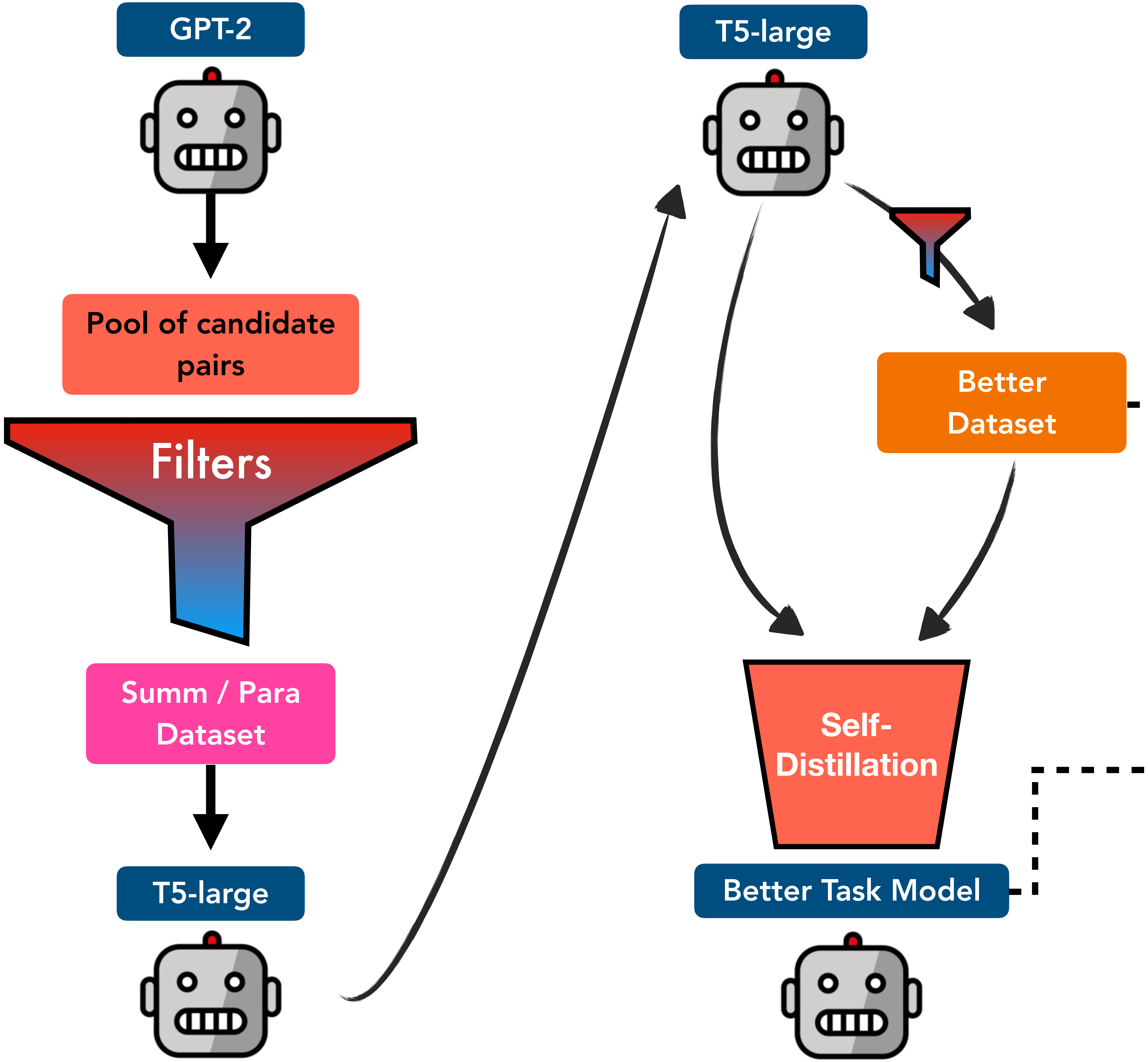
Allowing GPT-2 to generate *text to summarize*, it now generates *>10% correct pairs*!

Overall Framework





GPT-2
Pool of candidate pairs
Filters
Filters for Summarization
Summarization Dataset
T5-large
Entailment filter
remove non-factual summaries using NLI
Length filter
remove too long summaries
Diversity filter



Self-Distillation yields better dataset, stronger task model

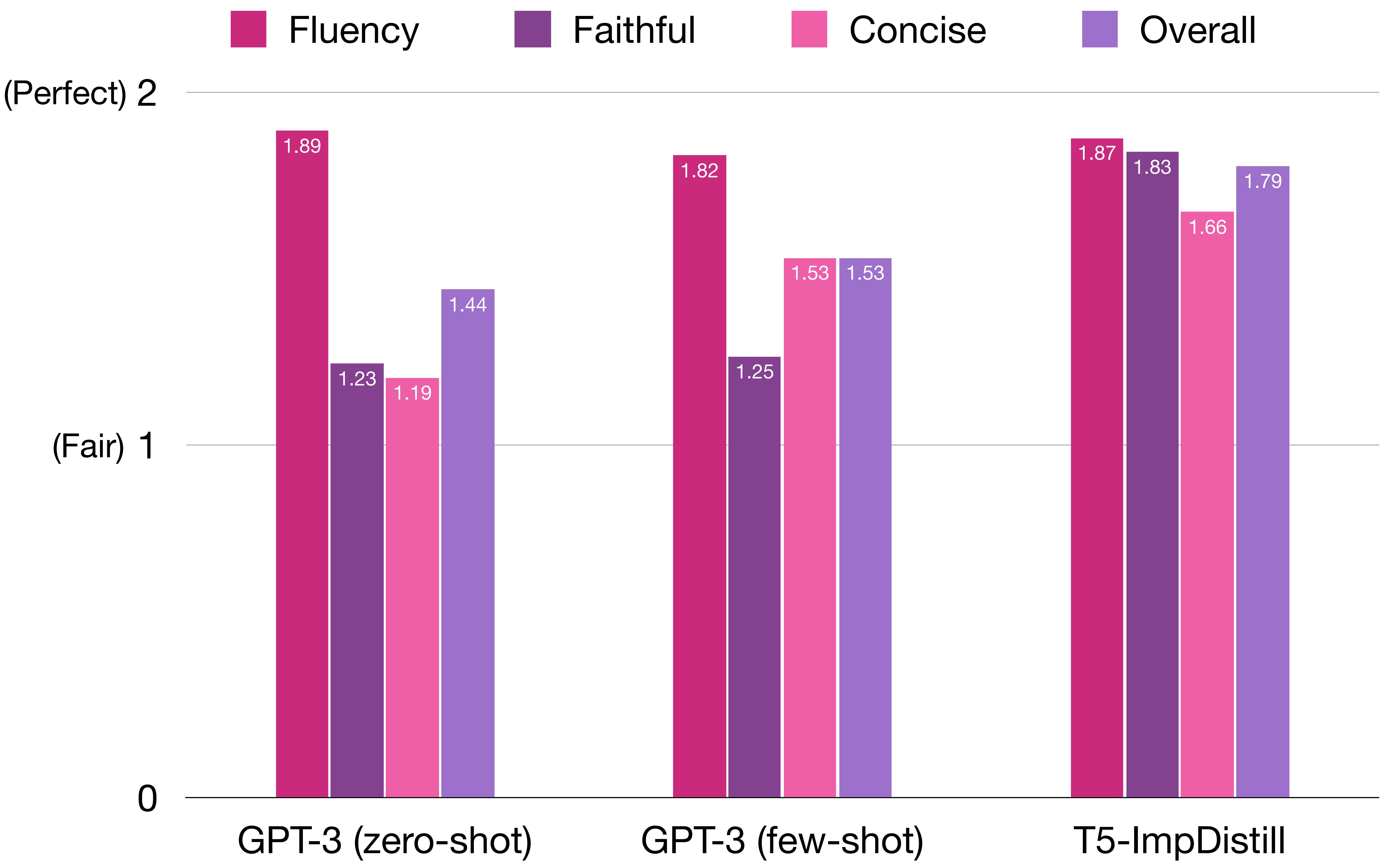
🥟 DimSum+

3.4M samples for
sentence summarization + paraphrasing,
spanning news / reddit / bio domains

🔥 T5-ImpDistill

770M LM capable of both
controllable summarization + paraphrasing,
distilled purely from < 2B LMs

Stronger than **200x larger GPT-3** in human evaluation!



Thank you!