

# **Implications for Cognitive Science**

**LLMs: Implications for Linguistics, Cognitive Science & Society**

Polina Tsvilodub & Michael Franke, Session 5

# Understanding understanding

## 1. Do LLMs **understand** language?

Depends on what it means to understand language.

## 2. Do LLMs **understand** the world?

Depends on what it means to understand the world.

## 3. How can we **understand** how LLMs work?

Requires familiarity w/ LLMs and w/ facility of human interpretation.

## 4. Do LLMs help us **understand** language or mind?

Depends on what we consider a useful **explanation** in science.

Wenn ein Löwe sprechen könnte, wir könnten ihn nicht **verstehen**.

meet the lion [here](#)





# On understanding

# What is $X$ in “ $S$ understands $X$ ”?

subject  $S$

- ▶  $X$  is a phenomenon (single observation or recurrent pattern)
  - “**objective understanding**” (Kelp 2015, 2017, Dellsén 2020)
  - may comprise:
    - “I understand you” (the way you act or feel)
    - “I understand Wagner” (the appeal or success of his music)
- ▶  $X$  is a theory
  - “theoretical understanding”
  - understanding theory  $\neq$  understanding phenomenon
    - I can understand phlogiston theory w/o understanding heat
- ▶  $X$  is a topic or subject matter
  - “topic understanding” (Brun & Baumberger 2017, Carter & Gordon 2016, Khalifa 2017)
  - umbrella concept? (“understands many  $X$ ’ falling under topic  $X$ ”)
- ▶  $X$  is a sign, a linguistic expression, a communicative action
  - “**linguistic understanding**” / “pragmatic understanding” (Longworth 2009)

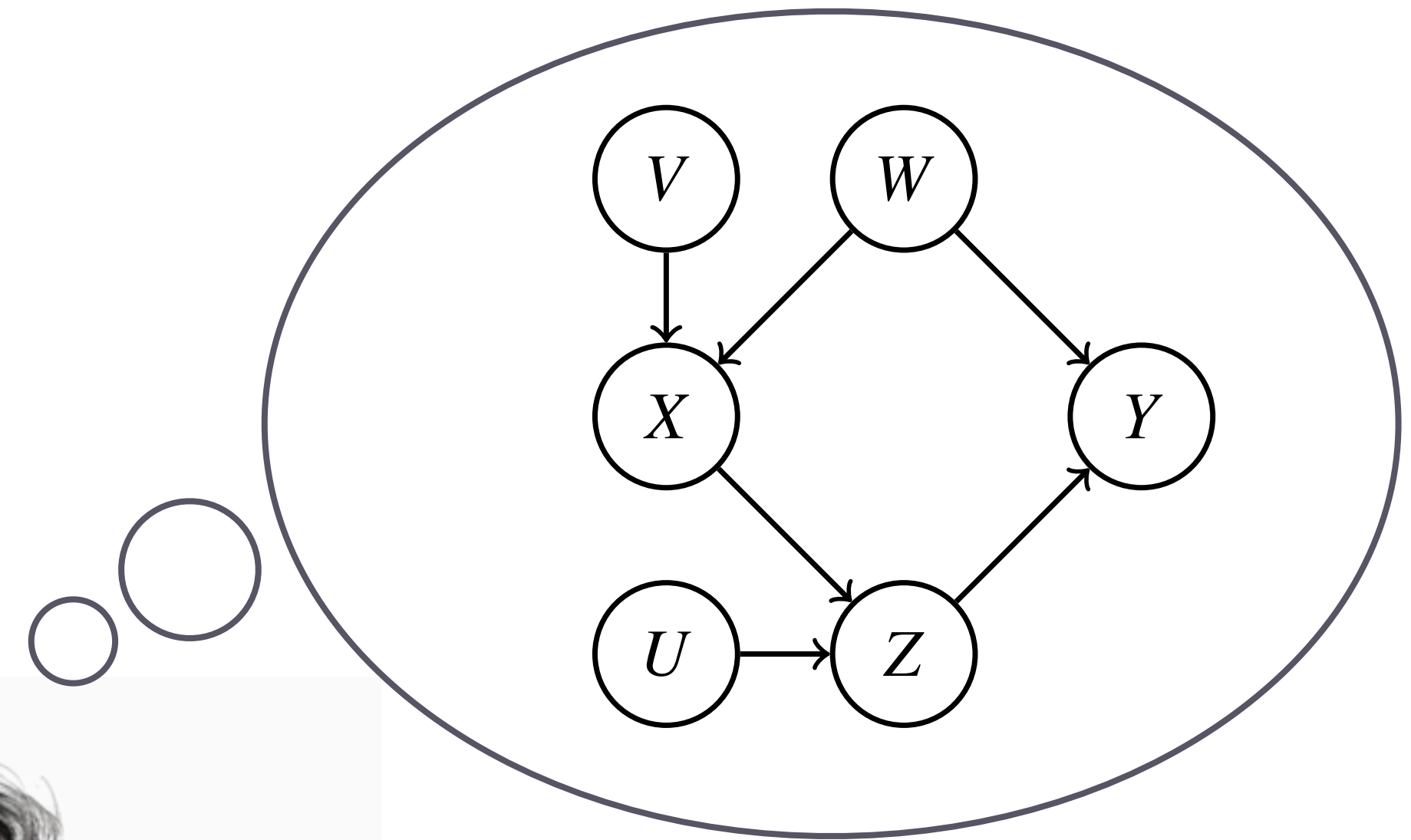
# What does “*S* understands *X*” mean?


subject *S* & phenomenon *X*

- ▶ subjective feeling of understanding
  - less important (for normative notion)
  - think: conspiracy theory (delusional view but strong emotional endorsement)
- ▶ ability to give verbal explanation
  - current debate about importance of explain-ability
    - pro: Strevens (2013), de Regt (2017), Khalifa (2017)
    - con: Wilkenfeld (2013, 2017), Kelp (2015, 2017), Dellsén (2020)
- ▶ ability to predict *X* (involving generalization, analogy, extrapolation, ...)
  - mere prediction-ability is not enough; some sort of “getting the gist” must be present
  - maybe entailed by a “true representation of the generative process of *X*”
- ▶ having an adequate, veridical **dependency model of *X*** (Dellsén 2020)
  - captures “data-generating process” around *X* in terms of dependency relations
    - causal relations, in-virtue-of relations

# The “theory theory” of mind

- ▶ developmental hypothesis about conceptual and causal learning
- ▶ statistical observation-based learning of probabilistic, causal Bayes nets
- ▶ explicit (probabilistic) models of the “**observation-generating process**”
- ▶ abstract, generative, causal concepts





**On understanding  
language**

# Turing test

- ▶ communicating solely through a text-based computer terminal, can we tell whether we are conversing with a human or a robot?





# Chinese room argument

Imagine a native English speaker who knows no Chinese locked in a room full of boxes of Chinese symbols (a data base) together with a book of instructions for manipulating the symbols (the program). Imagine that people outside the room send in other Chinese symbols which, unknown to the person in the room, are questions in Chinese (the input). And imagine that by following the instructions in the program the man in the room is able to pass out Chinese symbols which are correct answers to the questions (the output). The program enables the person in the room to pass the Turing Test for understanding Chinese but he does not understand a word of Chinese.

The point of the argument is this: if the man in the room does not understand Chinese on the basis of implementing the appropriate program for understanding Chinese then neither does any other digital computer solely on that basis because no computer, qua computer, has anything the man does not have.

Searle (1999)



# The "Clever Hans" effect



1.1. or 1.2. or 1.3. or 1.4. or 1.5. or 1.6. or 1.7. f

2.1. w 2.2. v 2.3. u 2.4. t or 2.5. m 2.6. s 2.7. g

3.1. f 3.2. r 3.3. i 3.4. j 3.5. k 3.6. l 3.7. m

4.1. w 4.2. o 4.3. o 4.4. p 4.5. q 4.6. r 4.7. t

5.4. h 5.5. b 5.6. s 5.7. t

6.4. n 6.5. m 6.6. o 6.7. z

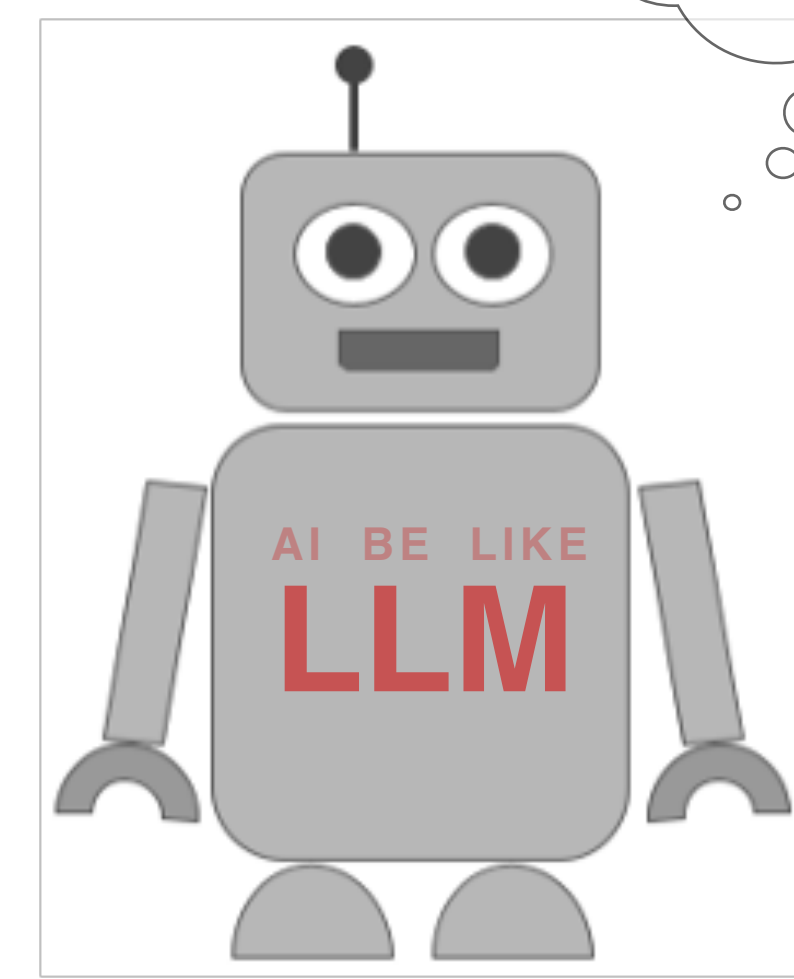
$$\frac{2}{3} + \frac{3}{4} =$$

$$26743 : 8 =$$



All penguins are black & white.  
Some old TV shows are black & white.  
Therefore, all penguins are old TV shows.

valid ● invalid



# Prompt understanding

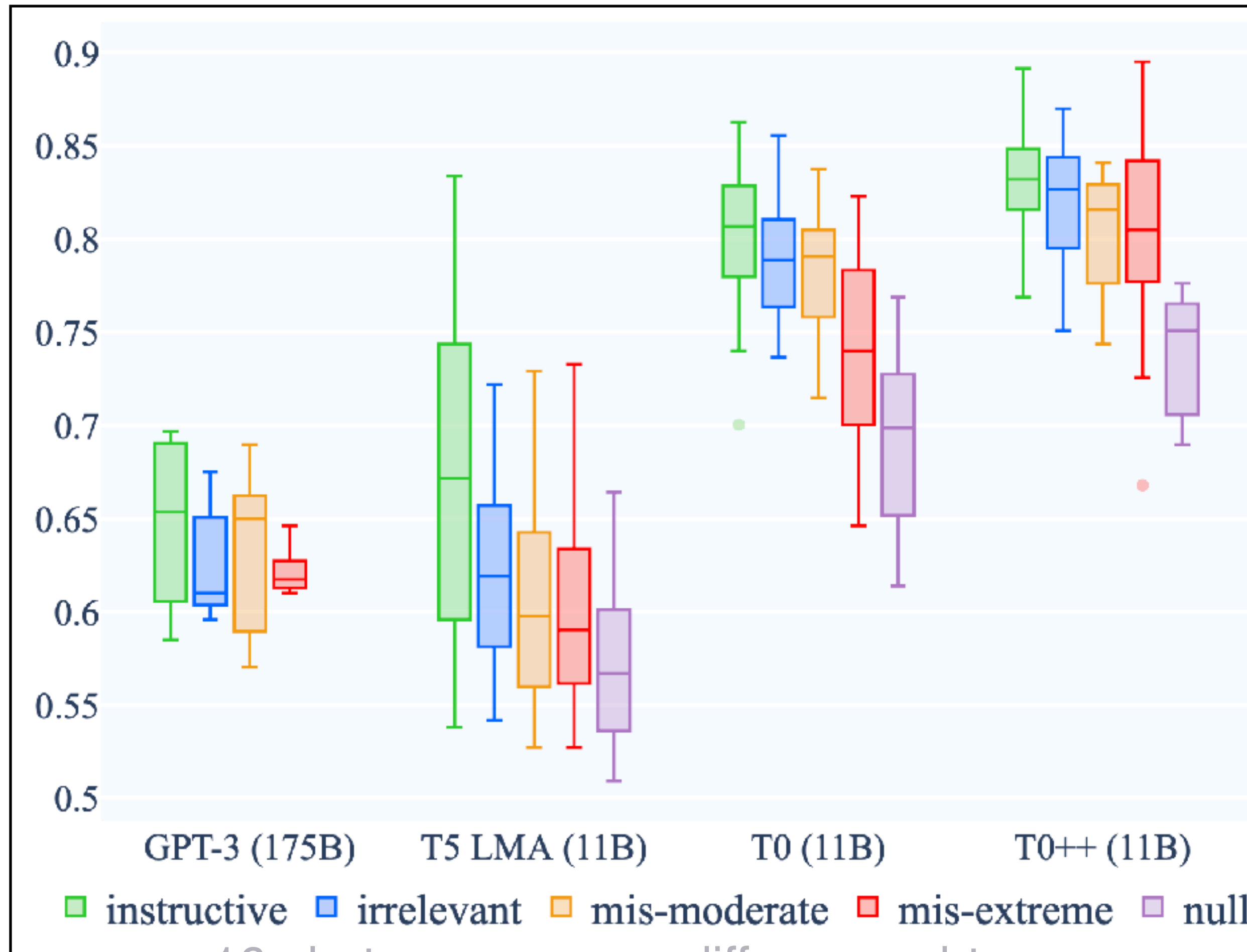
## setup

- ▶ zero-shot, and  $k$ -shot in-context learning
  - $k \in \{0,4,8,16,32,64,128,256\}$
- ▶ different example templates
  - see →
- ▶ different target words
  - yes/no
  - “yes/no”-like (true/false, positive/negative)
  - arbitrary (cat/dog, Jane/Jake)
  - reversed (no/yes)

Category	Examples
instructive	{premise} Are we justified in saying that “{hypo}”? Suppose {premise} Can we infer that “{hypo}”?
misleading-moderate	{premise} Can that be paraphrased as: “{hypo}”? {premise} Are there lots of similar words in “{hypo}”?
misleading-extreme	{premise} is the sentiment positive? {hypo} {premise} is this a sports news? {hypo}
irrelevant	{premise} If bonito flakes boil more than a few seconds the stock becomes too strong. "{hypo}"?
null	{premise} {hypothesis} {hypothesis} {premise}

# Prompt understanding

results

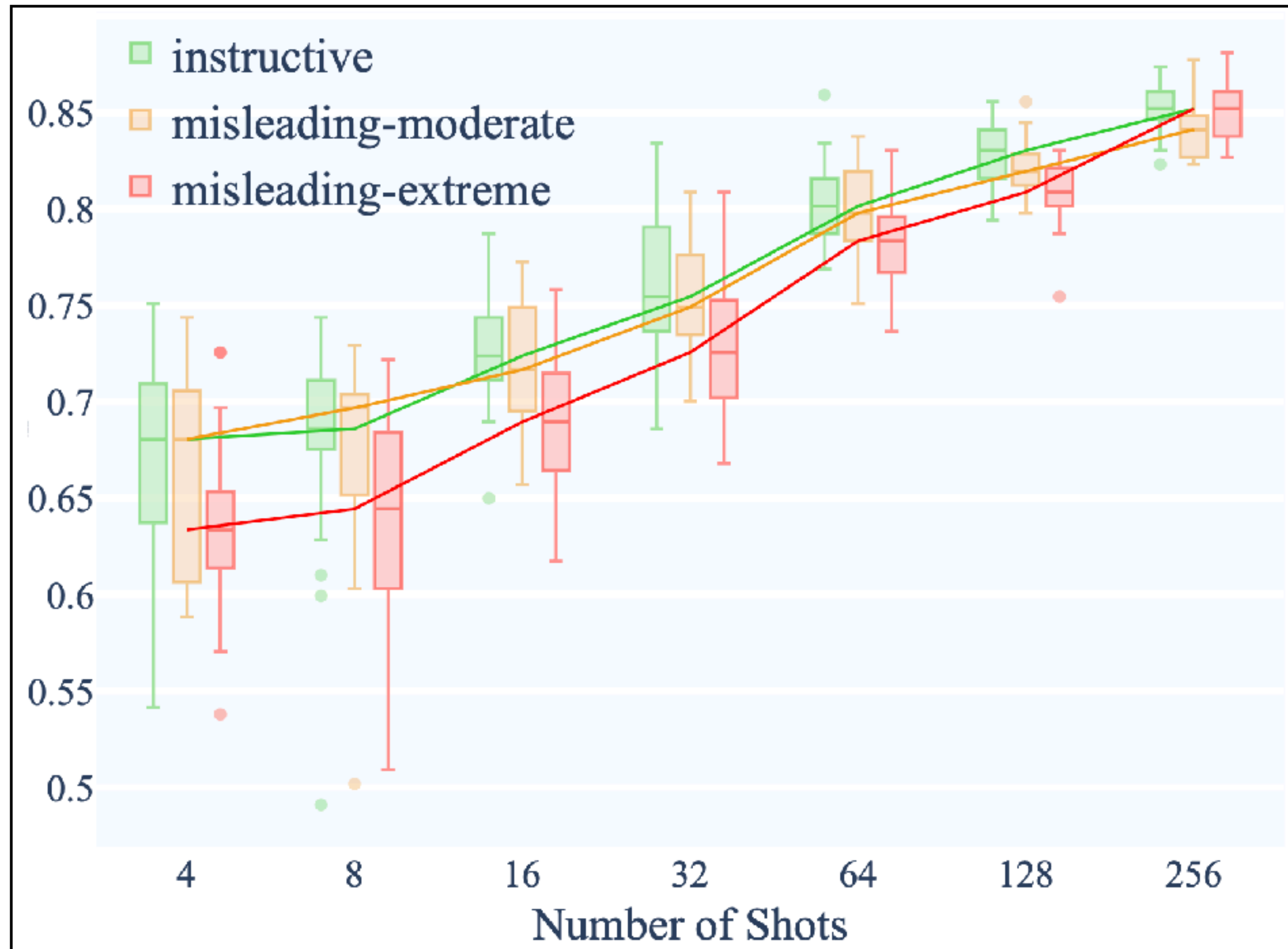


16-shot accuracy: no differences btw. categories except for comparisons against "null"

Category	Examples
instructive	{premise} Are we justified in saying that "{hypothesis}"? Suppose {premise} Can we infer that "{hypothesis}"?
misleading-moderate	{premise} Can that be paraphrased as: "{hypothesis}"? {premise} Are there lots of similar words in "{hypothesis}"?
misleading-extreme	{premise} is the sentiment positive? {hypothesis} {premise} is this a sports news? {hypothesis}
irrelevant	{premise} If bonito flakes boil more than a few seconds the stock becomes too strong. "{hypothesis}"?
null	{premise} {hypothesis} {hypothesis} {premise}

# Prompt understanding

results

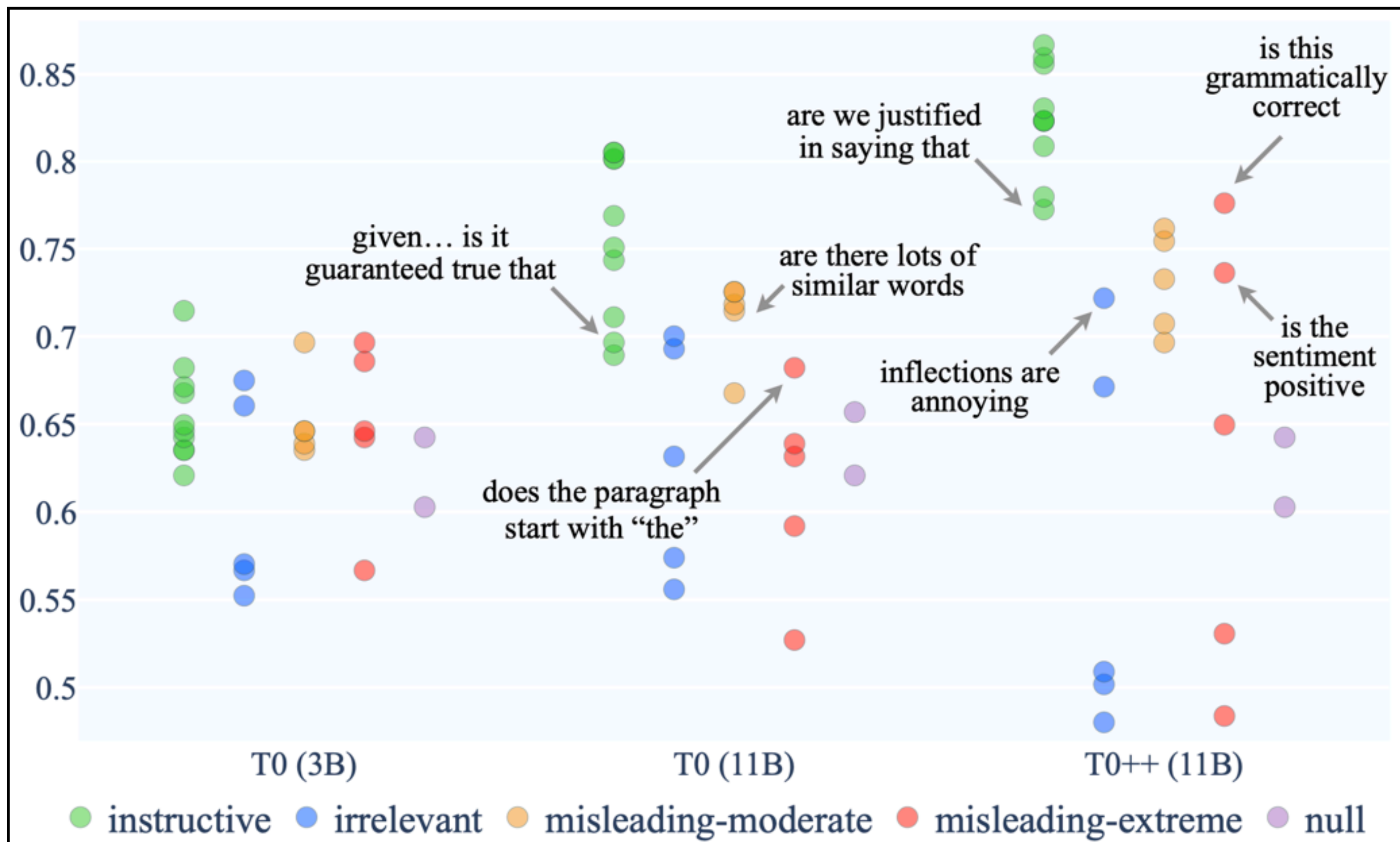


performance of T0 (3B): only difference btw.  
“misleading extreme” for 8 to 128 shots

Category	Examples
instructive	{prem} Are we justified in saying that “{hypo}”? Suppose {prem} Can we infer that “{hypo}”?
misleading-moderate	{prem} Can that be paraphrased as: “{hypo}”? {prem} Are there lots of similar words in “{hypo}”?
misleading-extreme	{prem} is the sentiment positive? {hypo} {prem} is this a sports news? {hypo}
irrelevant	{prem} If bonito flakes boil more than a few seconds the stock becomes too strong. “{hypo}”?
null	{premise} {hypothesis} {hypothesis} {premise}

# Prompt understanding

results



bigger T0 models are only models where instruction category matters for zero-shot

NB: T0 is only instruction-fine tuned model examined

Category	Examples
instructive	{premise} Are we justified in saying that “{hypo}”? {premise} Suppose {premise} Can we infer that “{hypo}”?
misleading-moderate	{premise} Can that be paraphrased as: “{hypo}”? {premise} Are there lots of similar words in “{hypo}”?
misleading-extreme	{premise} is the sentiment positive? {hypo} {premise} is this a sports news? {hypo}
irrelevant	{premise} If bonito flakes boil more than a few seconds the stock becomes too strong. "{hypo}"?
null	{premise} {hypothesis} {hypothesis} {premise}

# Prompt understanding

## results

- ▶ different target words
  - yes/no
  - “yes/no”-like (true/false, positive/negative)
  - arbitrary (cat/dog, Jane/Jake)
  - reversed (no/yes)
- ▶ target word matters for average success
- ▶ but unintuitive interactions with:
  - the former can outperform the latter
    - {premise} Does the paragraph start with “the”?  
{hypothesis} [yes/no]
    - {premise} Based on the previous passage, is it true that  
“{hypothesis}”? [cat/dog]

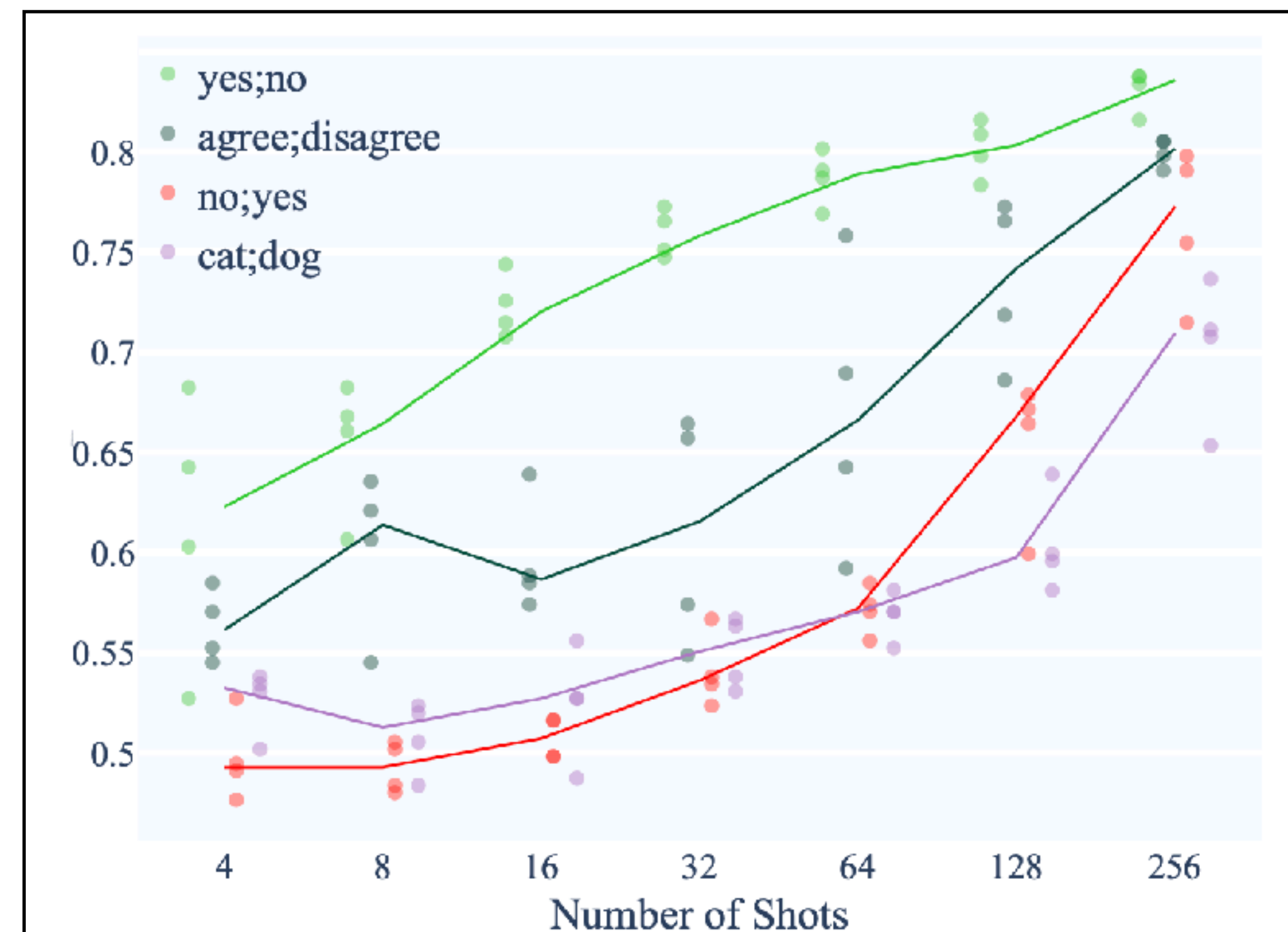


Figure 5: The best-performing instructive template for ALBERT on RTE, {prem} Are we justified in saying that "{hypo}"? with select LM targets from each category.





**On understanding  
the world**

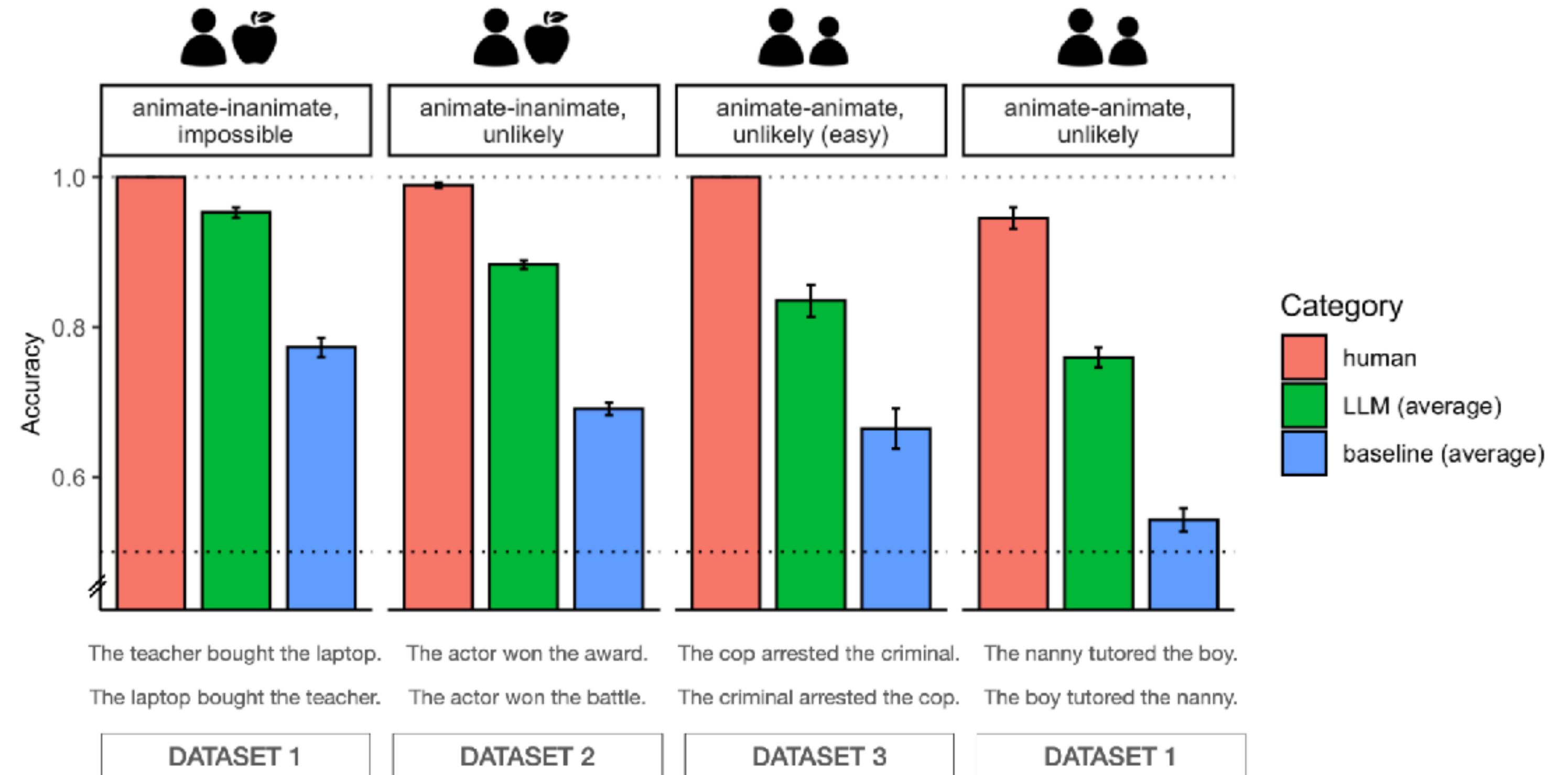
# Probabilistic world knowledge in LLMs

- ▶ can LLMs distinguish impossible, improbable and probable events?

Sentence Set	Plausible?	Voice	Synonym #	Sentence
<b>Dataset 1</b> <i>(Fedorenko et al. 2020)</i>	yes	active	1	The teacher bought the laptop.
			2	The instructor purchased the computer.
		passive	1	The laptop was bought by the teacher.
		2	The computer was purchased by the instructor.	
	no	active	1	The laptop bought the teacher.
			2	The computer purchased the instructor.
		passive	1	The teacher was bought by the laptop.
		2	The instructor was purchased by the computer.	
<b>Dataset 2</b> <i>(Vassallo et al. 2018)</i>	yes	active	-	The actor won the award.
	no	active	-	The actor won the battle.
<b>Dataset 3</b> <i>(Ivanova et al. 2021)</i>	yes	active	-	The cop is arresting the criminal.
	no	active	-	The criminal is arresting the cop.

# Probabilistic world knowledge in LLMs

- ▶ LLM interpretation:
  - compare probability of sentences under next-token prediction
- ▶ LLMs tested:
  - BERT, RoBERTa, GPT-2, GPT-J
- ▶ baseline models:
  - small LSTMs, theory-driven models, distributional models



**Figure 2.** Human accuracy as well as average accuracy of the four LLM models (LLM (average)) and average accuracy of the four baseline models (baseline (average)) on Dataset 1 (the first and fourth set of bars; same data as in Figure 1), as well as Datasets 2 and 3 (the second and third set of bars); results ordered by LLM performance. Dotted lines indicate chance-level performance.

# Causal reasoning

You have previously observed the following chemical substances in different wine casks:

- Cask 1: substance A was present, substance B was present, substance C was present.

- Cask 2: substance A was present, substance B was present, substance C was present.

[...]

- Cask 20: substance A was absent, substance B was absent, substance C was absent.

You have the following additional information from previous research:

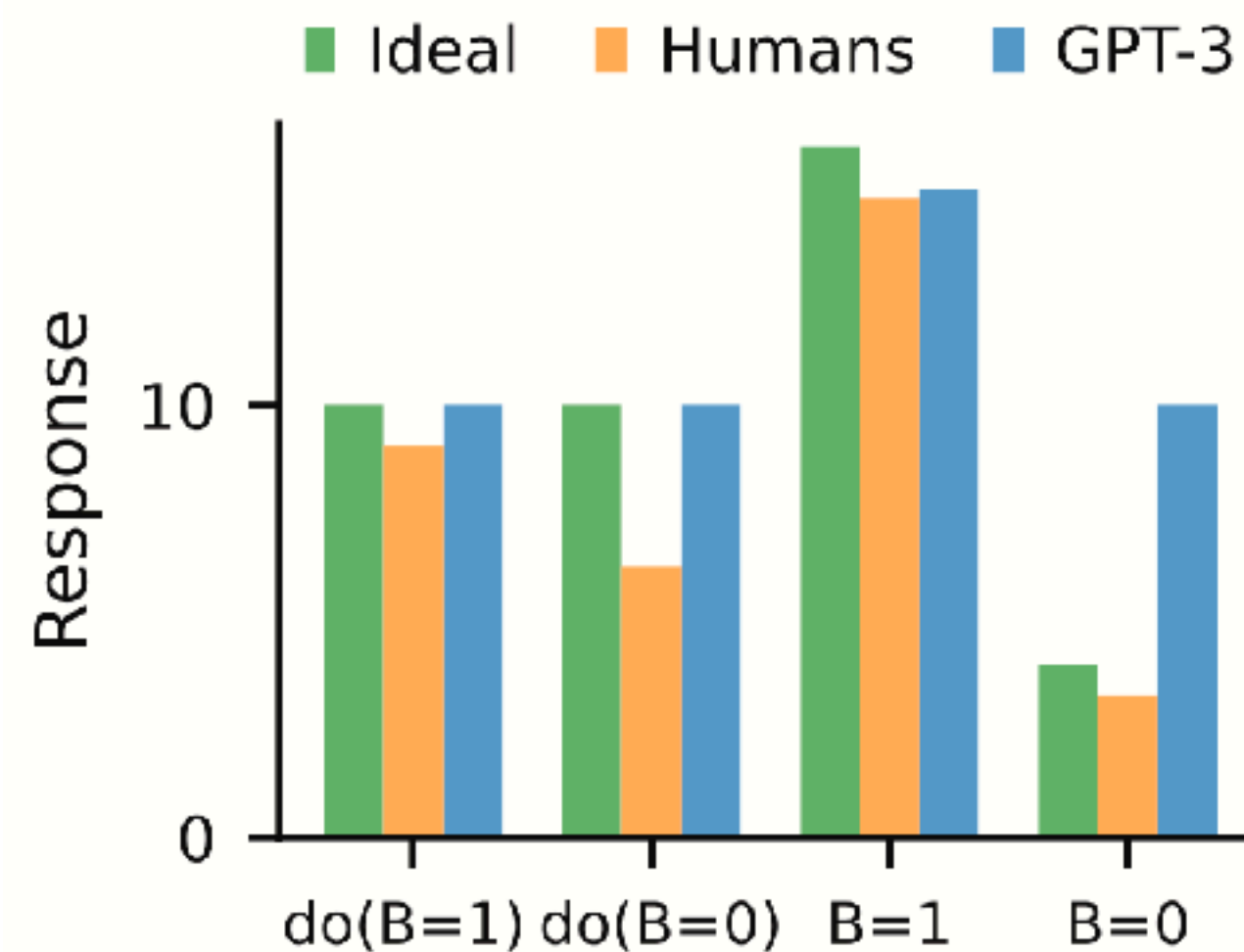
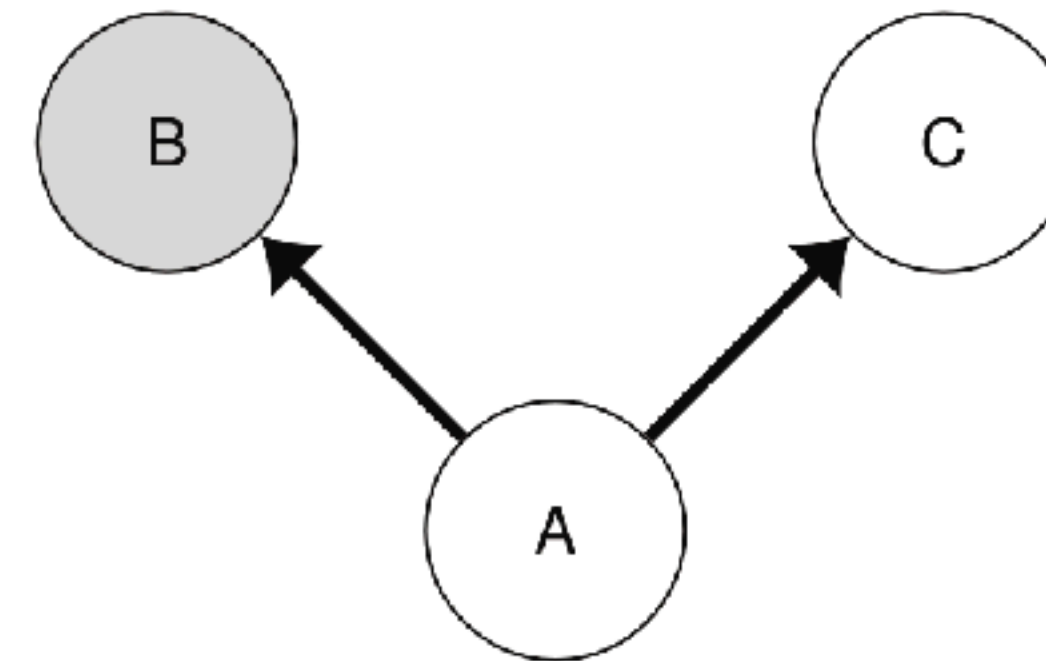
- Substance A likely causes the production of substance B.
- Substance A likely causes the production of substance C.

Imagine that you test 20 new casks in which you have manually added substance B.

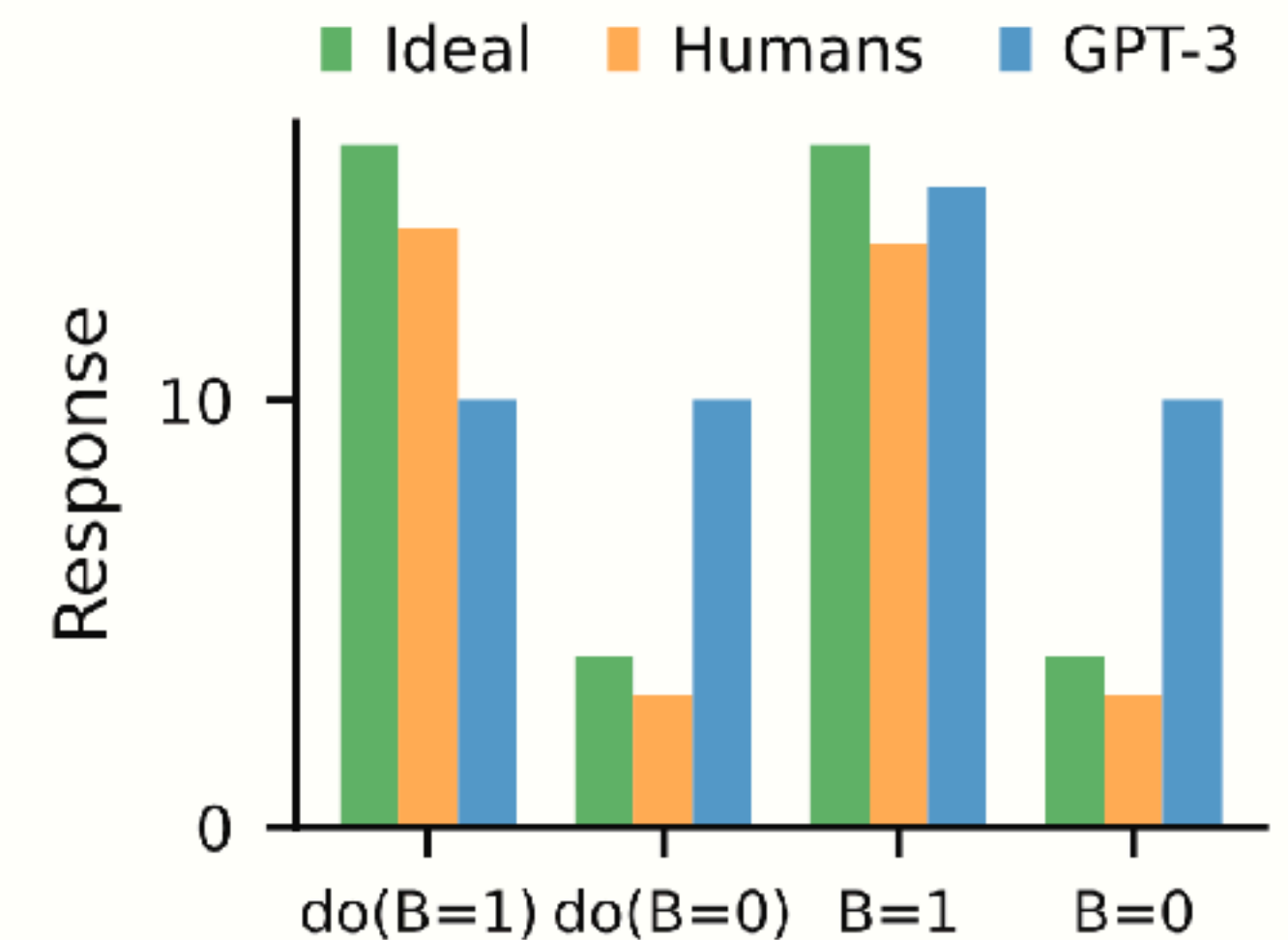
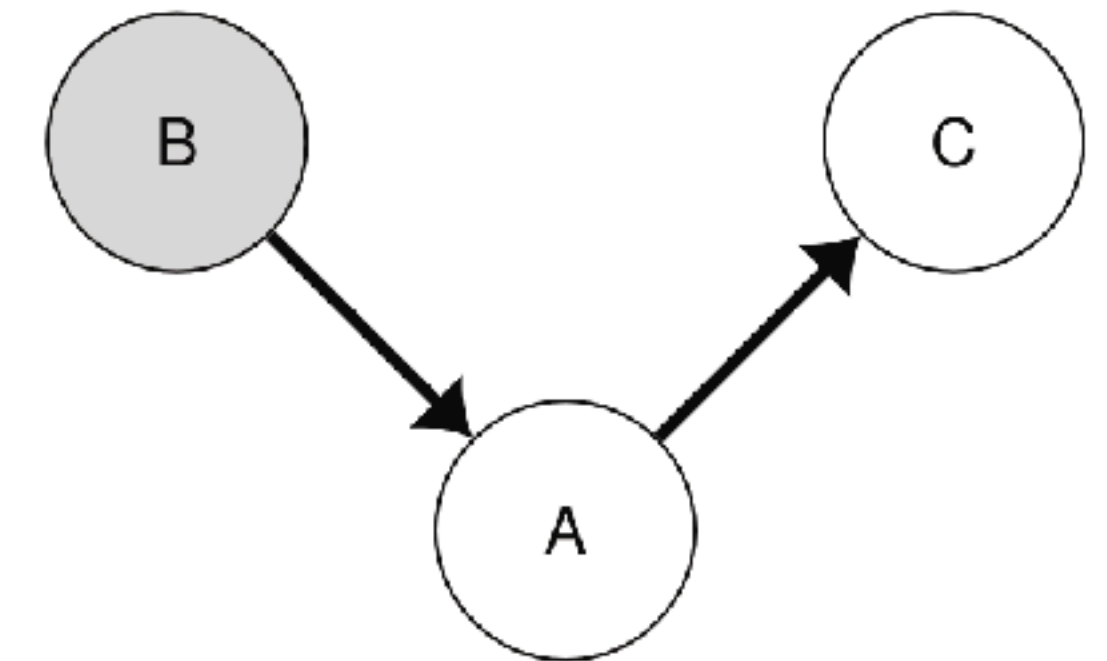
Q: How many of these new casks will contain substance C on average?

A: [insert] casks.

Common-Cause



Causal-Chain




# Think break

Anything here you would agree or disagree with?

Humans learn by connecting with other people, asking them questions, and actively engaging with their environments, whereas large language models learn by being passively fed a lot of text and predicting what word comes next. GPT-3 also failed to learn about and use causal knowledge in a simple reasoning task. We believe it makes sense that GPT-3 struggles to reason causally because acquiring knowledge about interventions from passive streams of data is hard to impossible (32).

Binz & Schulz (2023)





# On understanding language models

# Principle of charity

Norm for critical thinking & proper argumentation

- ▶ interpret a speaker's statements as the most rational, strongest and most coherent claim
- ▶ ask yourself: "What could have motivated or caused this position?" or: "In which light is this a coherent, convincing position to hold?"
- ▶ aspects of charity include ascriptions of ...
  - regular meaning of words and phrases
  - beliefs and perceptions corresponding to what is said
  - an overall consistent belief set / world view
  - common human motivations and goals
  - ...



# Grice's Maxims of Conversation

Assumptions about speaker behavior to infer what was meant

## Maxim of Quality

Try to make your contribution one that is true.

- (i) Do not say what you believe to be false.
- (ii) Do not say that for which you lack adequate evidence.

## Maxim of Quantity

- (i) Make your contribution as informative as is required for the current purposes of the exchange.
- (ii) Do not make your contribution more informative than is required.

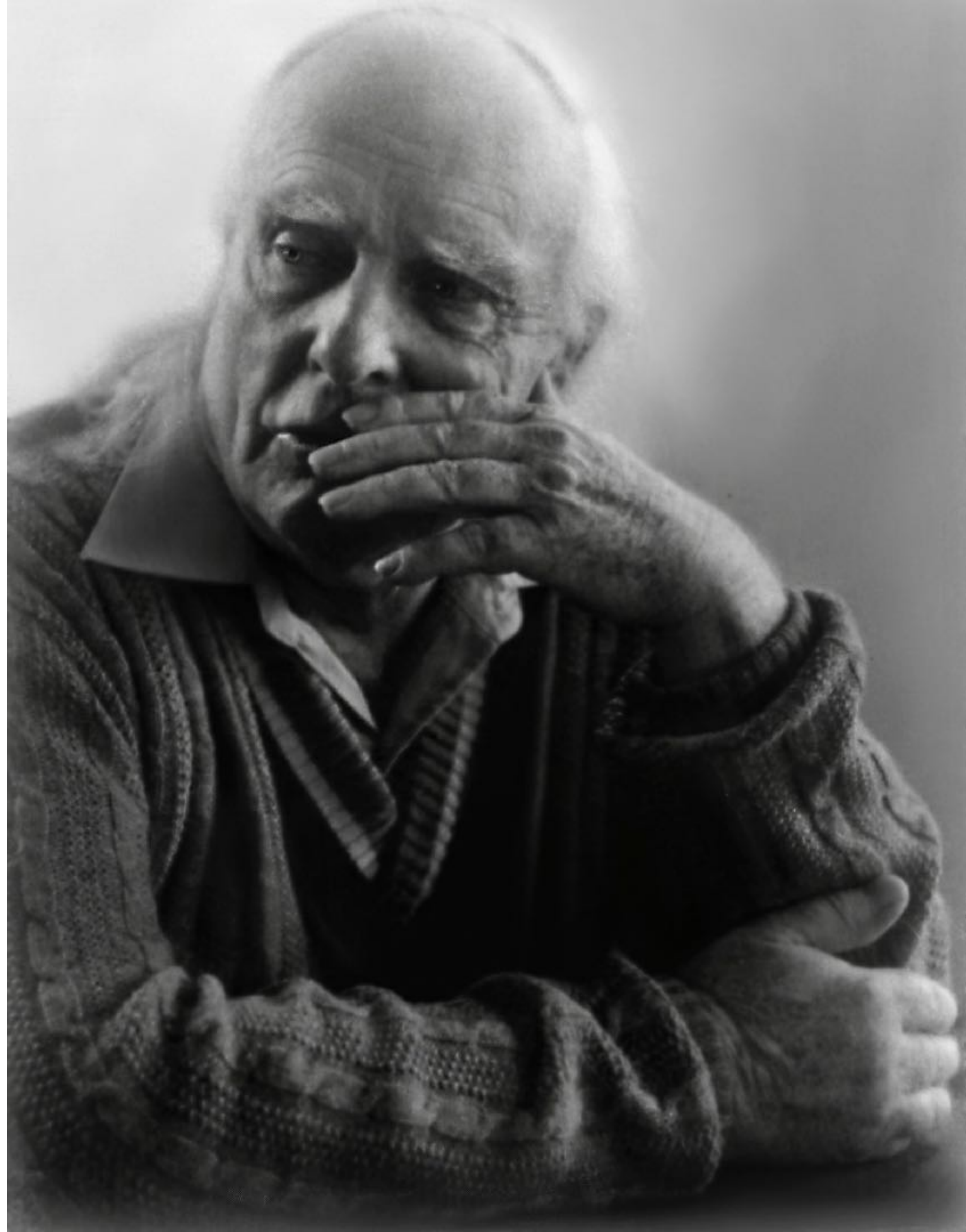
## Maxim of Relation

- (i) Be relevant.

## Maxim of Manner

Be perspicuous.

- (i) Avoid obscurity of expression.
- (ii) Avoid ambiguity.
- (iii) Be brief (avoid unnecessary prolixity).
- (iv) Be orderly.





# Relevance theory

## Cognitive Principle of Relevance

“Human cognition tends to be geared to the maximization of relevance.”

Sperber and Wilson (1995), p. 260

## Relevance of an input to an individual

“[T]he greater the positive **cognitive effects** achieved by processing an input, the greater the relevance, [...] the greater the **processing effort** expended, the lower the relevance of the input to the individual at that time.”

Wilson and Sperber (2004), p. 610

## RT-interpretation mechanism

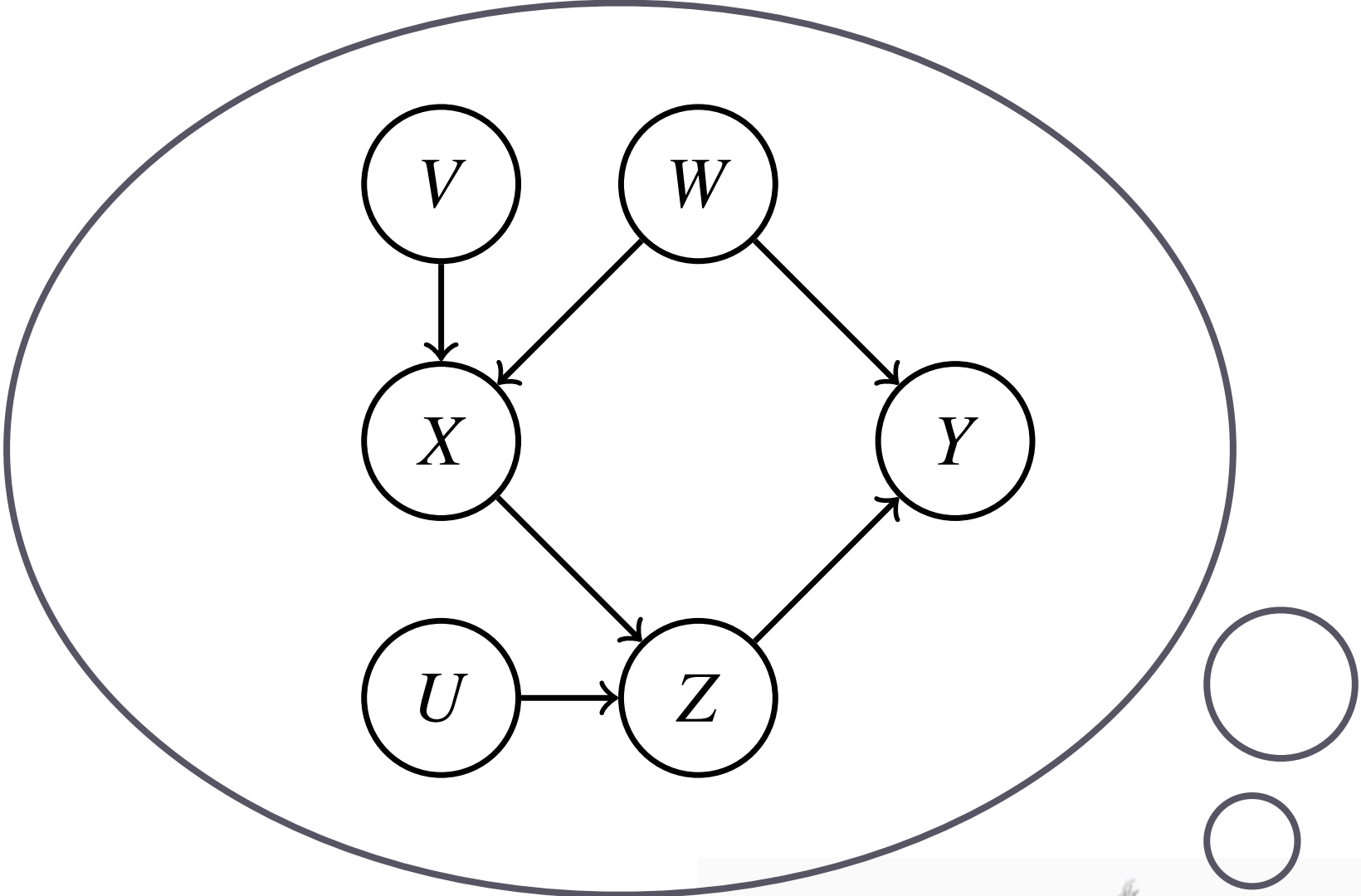
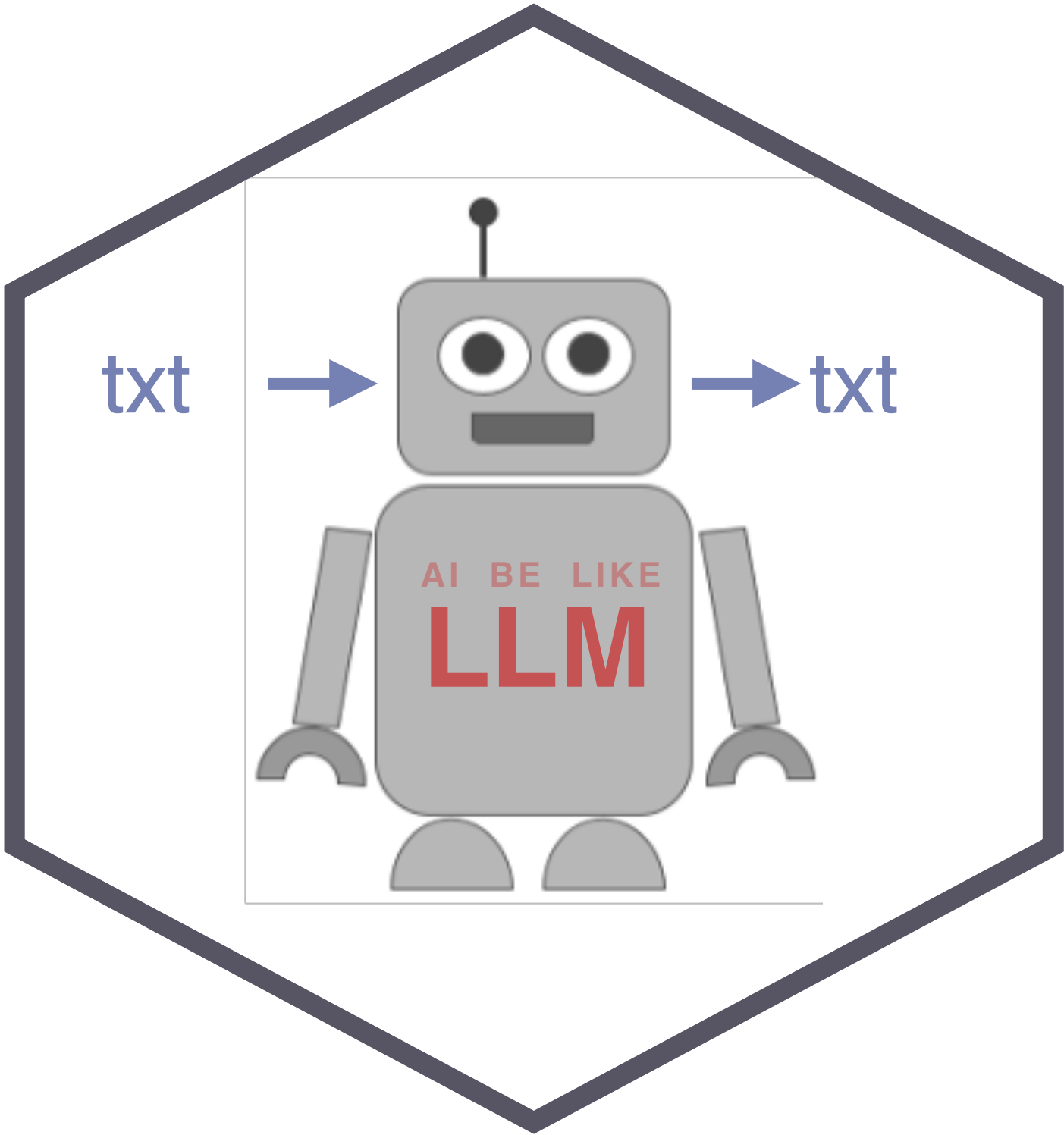
“Follow a path of least effort in computing cognitive effects: Test interpretive hypotheses [...] in order of accessibility [...] and] [s]top when your expectations of relevance are satisfied.”

Wilson and Sperber (2002), p. 260



# Generative-process thinking meets GPT

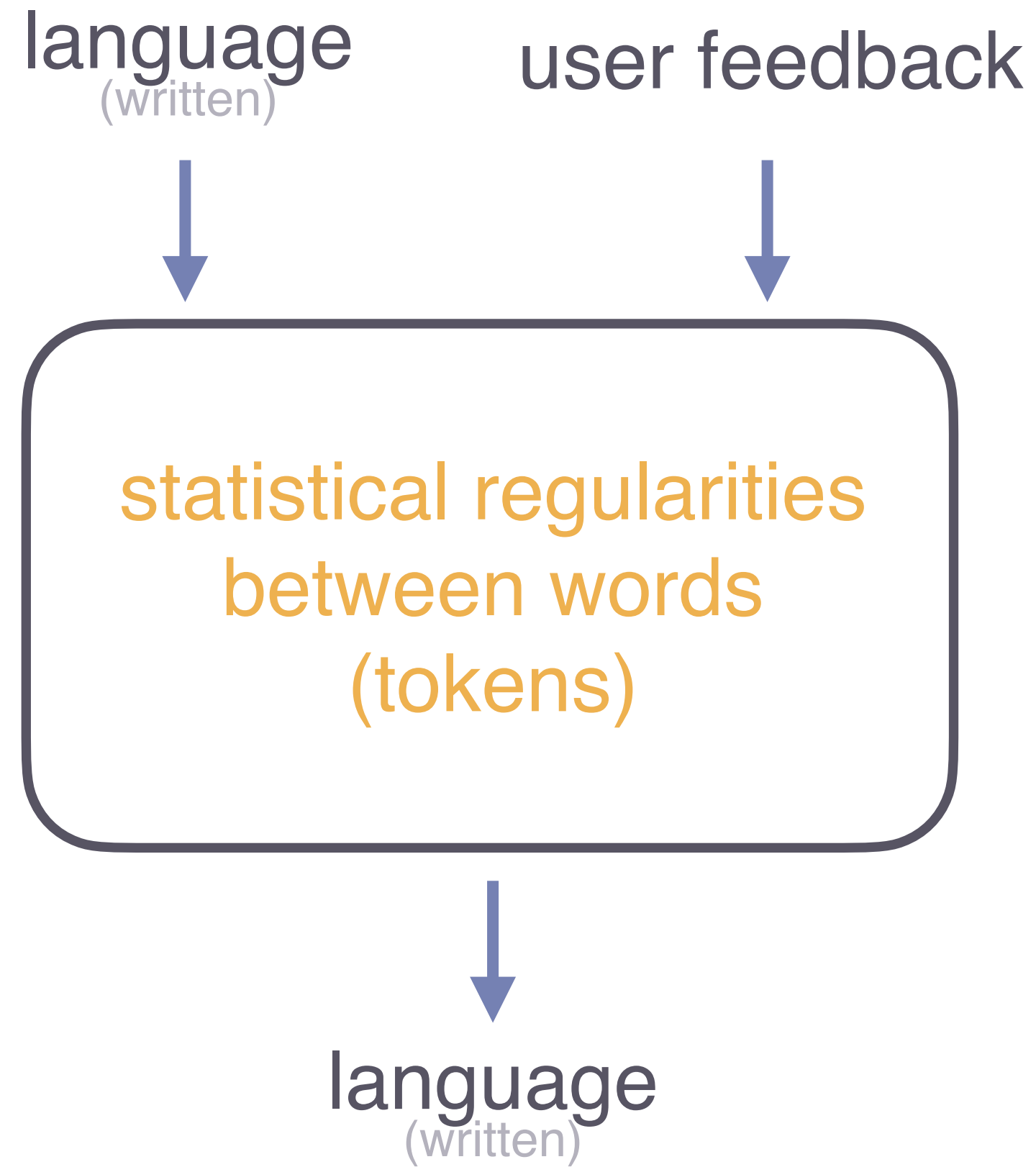
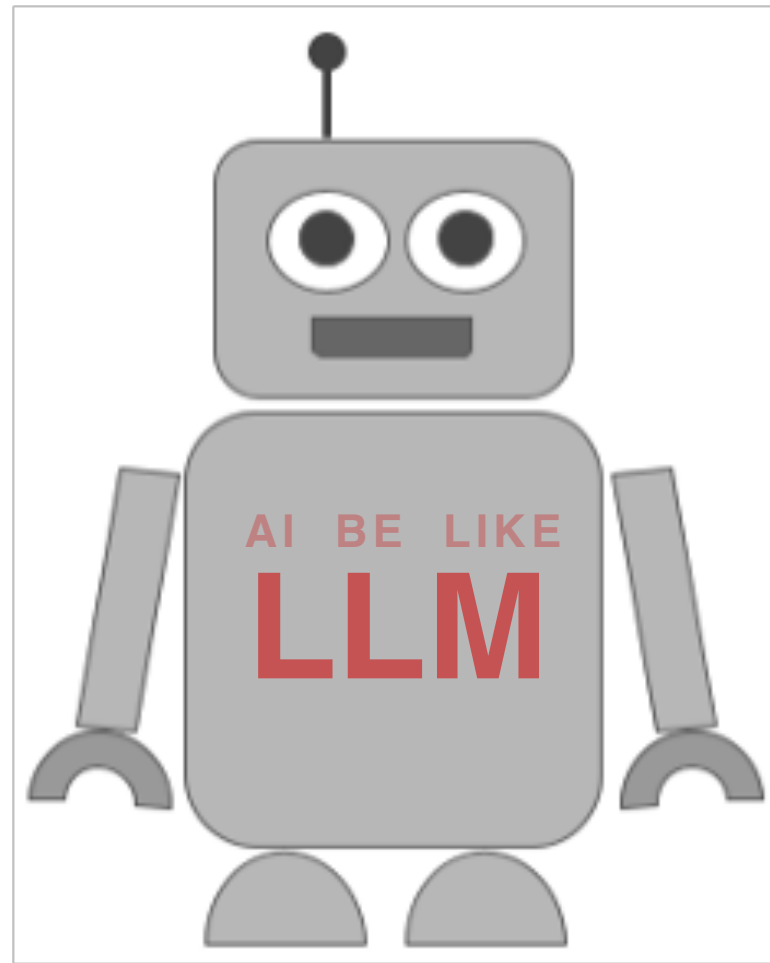
or, how to outwit thousands of years of evolution



# Two forms of intelligence

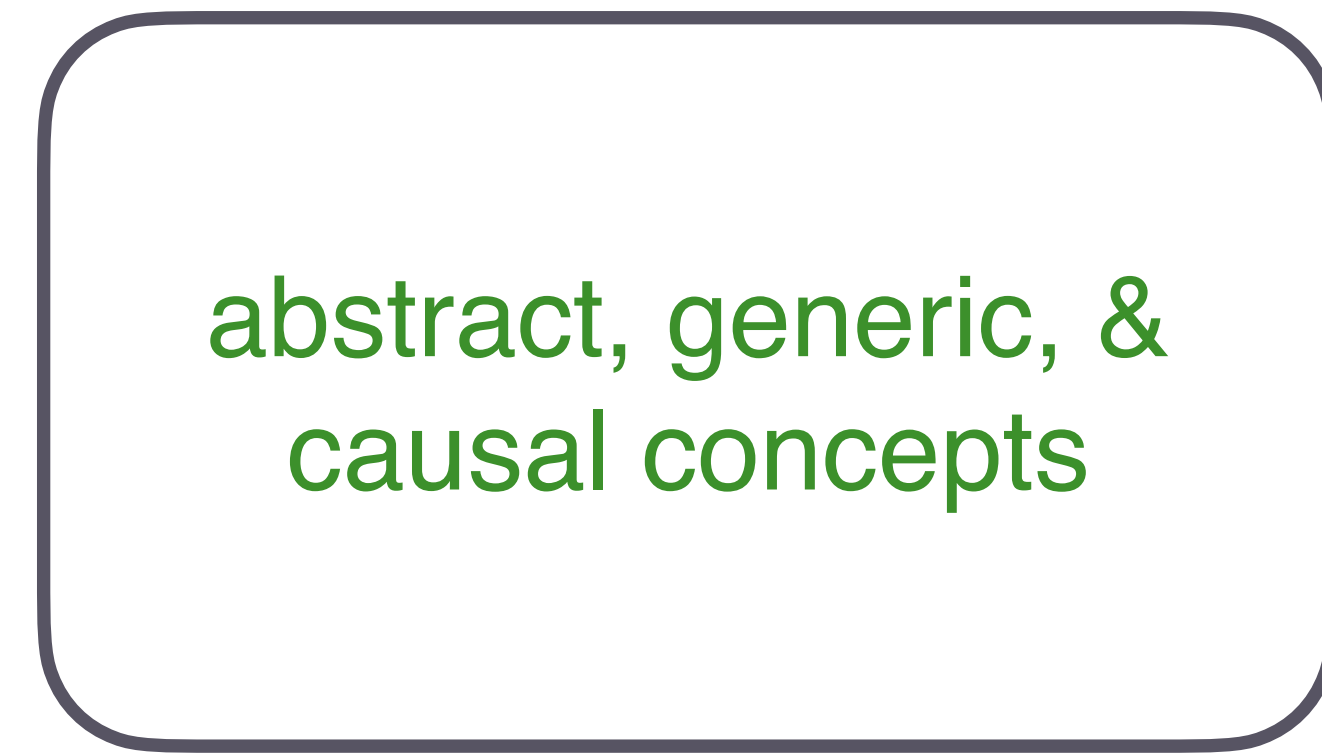
or: the LLM cheat sheet

NEITHER OF WHICH  
ANYONE REALLY FULLY  
UNDERSTANDS



language  
(written, spoken, signed)

world  
(the whole gory mess)





# **Explanations** & (language) models

# Flavors of NLP models



# Dimensions of explanatory value

## 1. performance

- a. training scores
- b. human judgements
- c. benchmark scores
- d. replicability
- e. generalization

## 2. indirect support

- a. prima facie conceptual plausibility
- b. support from extant theory
- c. support from prior data

## 3. parsimony

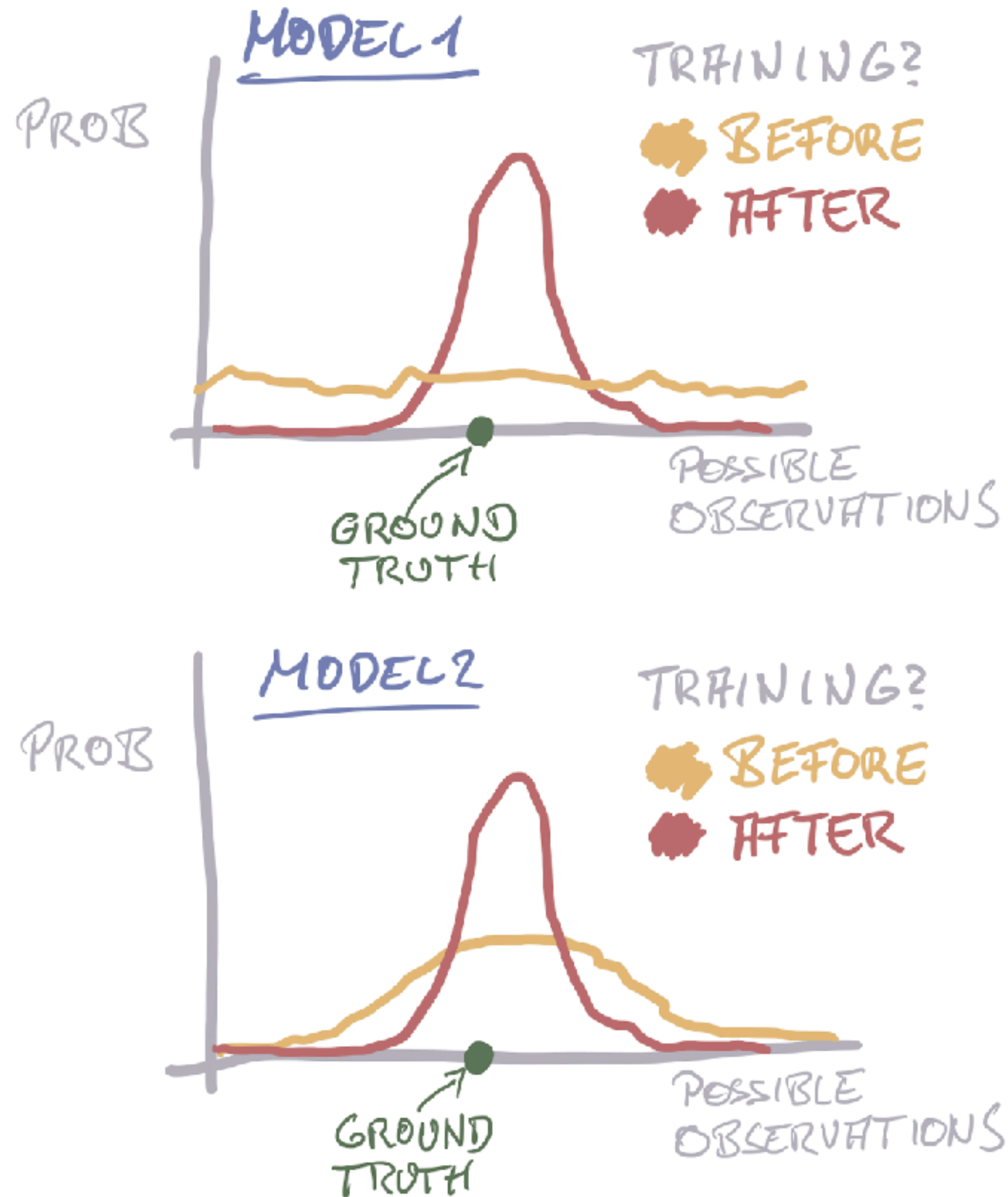
- a. simpler, elegant
- b. more compressible
- c. aligned with prior modeling choices

*oh, nice! a table to compare incomparables! how low can does you go?*

	LLMs	ling. theory
performance	✓	—
indirect support	—	✓
parsimony	—	✓

# Specificity

which model is better?

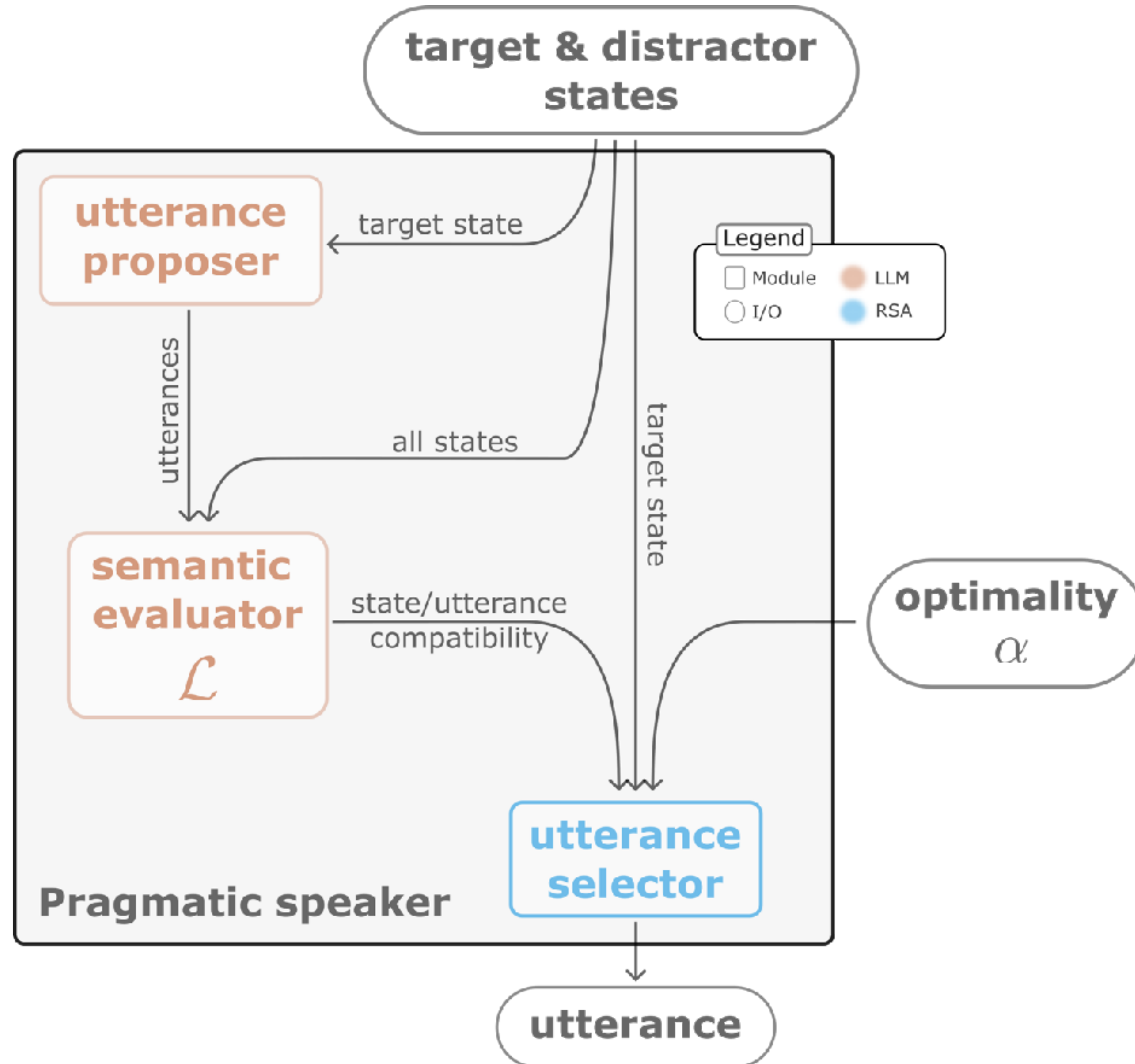


really, guys? another table? you've got to be kidding ...

	LLMs	ling. theory
performance	✓	—
indirect support	—	✓
parsimony	—	✓
specificity	☠	✓

# Hybrid cognitive models

integrating LLMs in explanatory models





# Summary

## Understanding & explanation

- ▶ meaning of “**S understands X**” depends on what *X* is
  - language understanding: flexible & robust generalizations across context
  - world understanding: probabilistic and causal reasoning about the world
  - understanding of LLMs: capturing I/O vs. the generative process behind the behavior
- ▶ **explanatory models** can be evaluated based on
  - performance
  - indirect support
  - parsimony
- ▶ we can build **hybrid** cognitive models with LLM components





# **Chains & Agents**

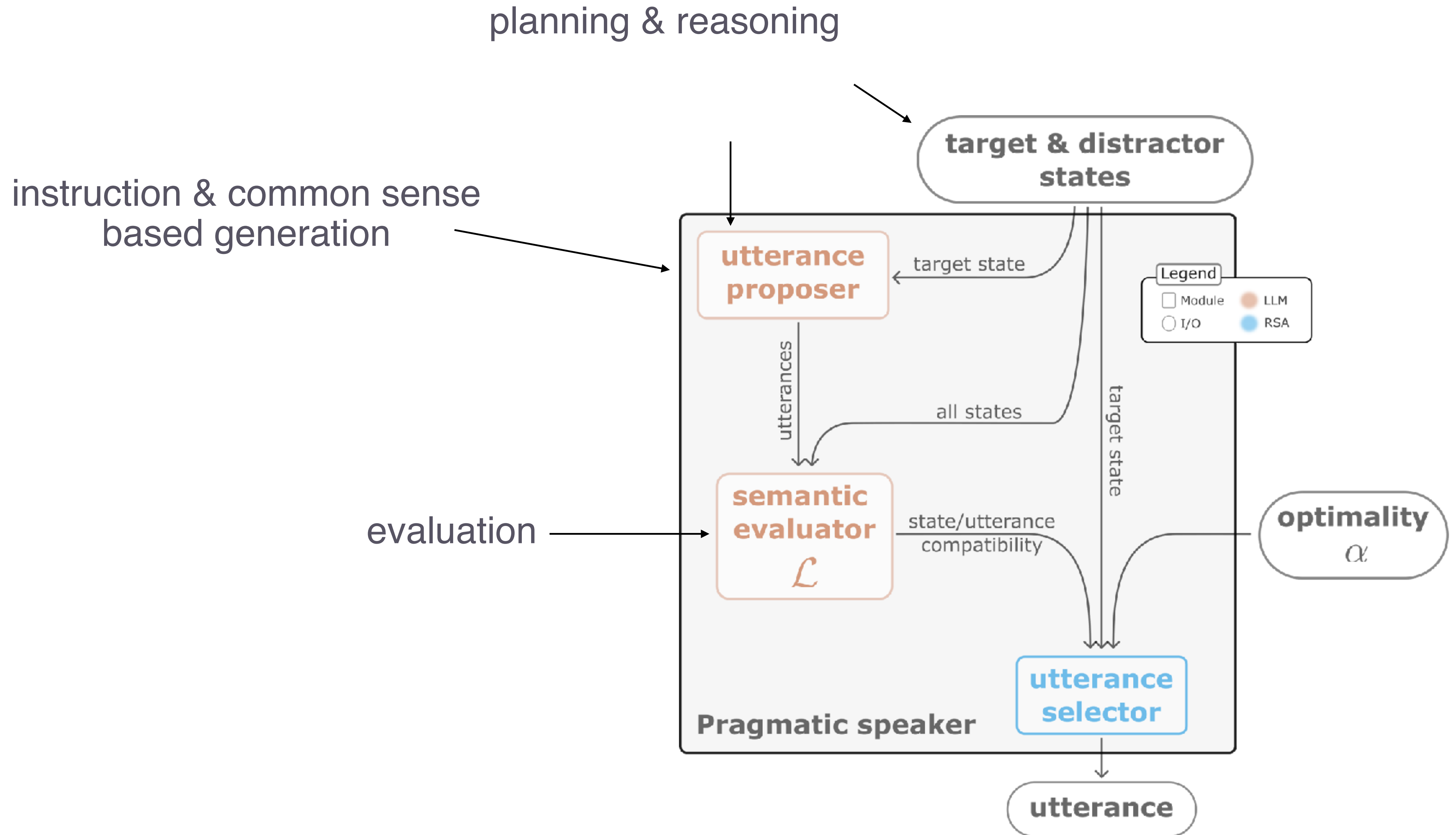
## Building hybrid models

# LLMs as knowledge bases

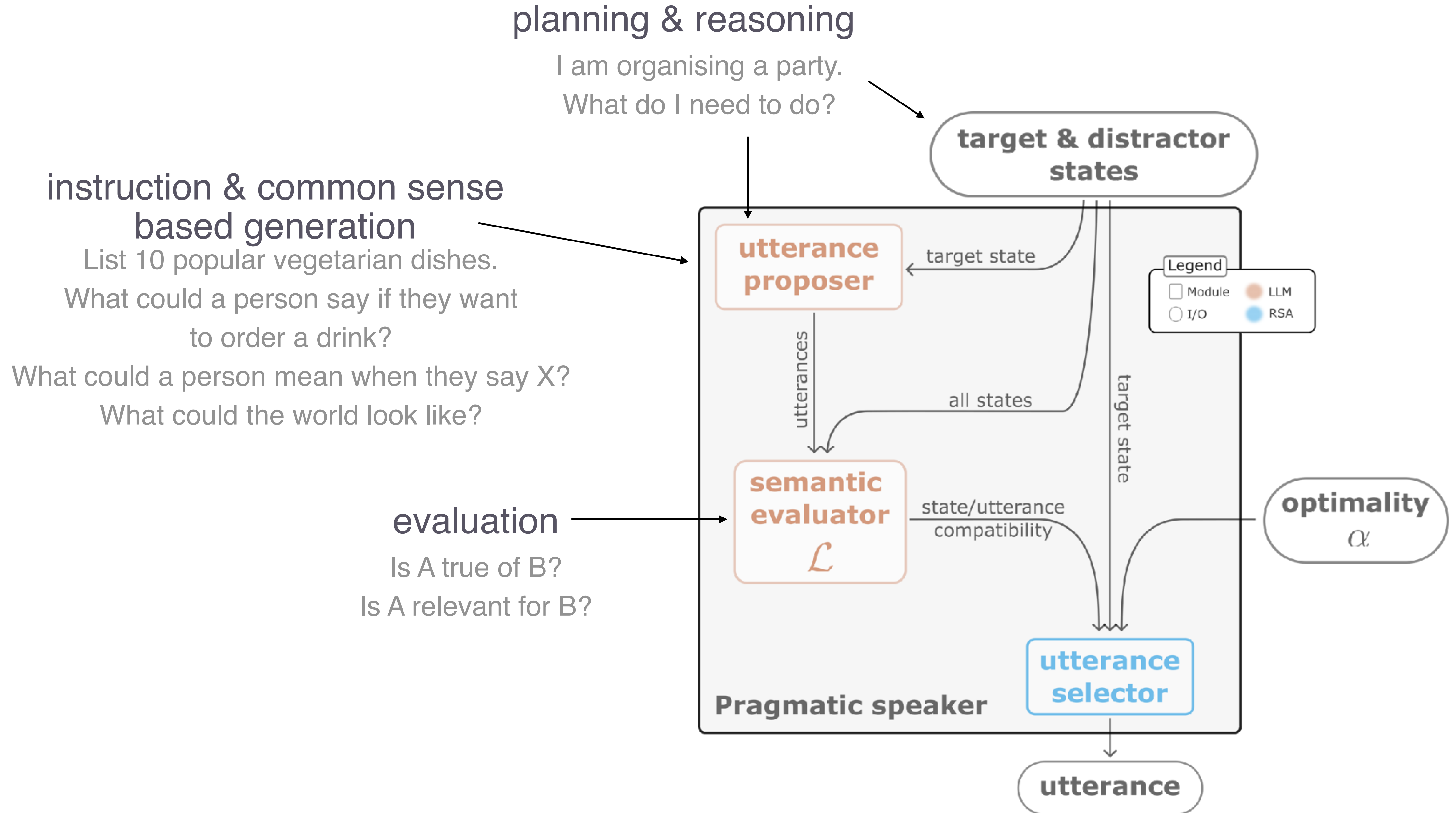
“The key observation is that large language models **encode a wide range of human behavior** represented in their training data. [...] With their ability to **generate and decompose action sequences**, large language models have also been used in planning [...].”

“[...] we compare GPT-4 to ChatGPT throughout to showcase a giant leap in level of **common sense** learned by GPT-4 compared to its predecessor.”

# Building hybrid models



# Building hybrid models



# Few-shot prompting

aka: in-context learning

## RECAP

- ▶ give task instruction
- ▶ give one or more examples
- ▶ works if pattern is recognizable in examples
- ▶ curation, statistics and form of examples matters

### INPUT

A "whatpu" is a small, furry animal native to Tanzania.  
An example of a sentence that uses the word whatpu is:  
We were traveling in Africa and we saw these very cute whatpus.

To do a "farduddle" means to jump up and down really fast.  
An example of a sentence that uses the word farduddle is:

### OUTPUT

When we won the game, we all started to farduddle in celebration.

- ▶ give task instruction
- ▶ give one or more **examples with explicit chain-of-thought reasoning leading to the correct answer**
- ▶ works for example to complex for few-shot prompting
- ▶ requires “right” task analysis in CoT steps

### INPUT

The odd numbers in this group add up to an even number: 4, 8, 9, 15, 12, 2, 1.

A: Adding all the odd numbers (9, 15, 1) gives 25. The answer is False.

The odd numbers in this group add up to an even number: 15, 32, 5, 13, 82, 7, 1.

A:

### OUTPUT

Adding all the odd numbers (15, 5, 13, 7, 1) gives 41. The answer is False.

## ▶ just add “Let’s think step by step”

- even better (Zhou et al. [2022](#)): "Let's work this out in a step by step way to be sure we have the right answer."

(a) Few-shot	(b) Few-shot-CoT
<p>Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now? A: The answer is 11.</p> <p>Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there? A:</p> <p><i>(Output) The answer is 8. ✗</i></p>	<p>Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now? A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. <math>5 + 6 = 11</math>. The answer is 11.</p> <p>Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there? A:</p> <p><i>(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are <math>16 / 2 = 8</math> golf balls. Half of the golf balls are blue. So there are <math>8 / 2 = 4</math> blue golf balls. The answer is 4. ✓</i></p>
(c) Zero-shot	(d) Zero-shot-CoT (Ours)
<p>Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there? A: The answer (arabic numerals) is</p> <p><i>(Output) 8 ✗</i></p>	<p>Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there? A: <b>Let's think step by step.</b></p> <p><i>(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓</i></p>

## INPUT

The odd numbers in this group add up to an even number: 15, 32, 5, 13, 82, 7, 1.

A: Let's think step by step.

## OUTPUT

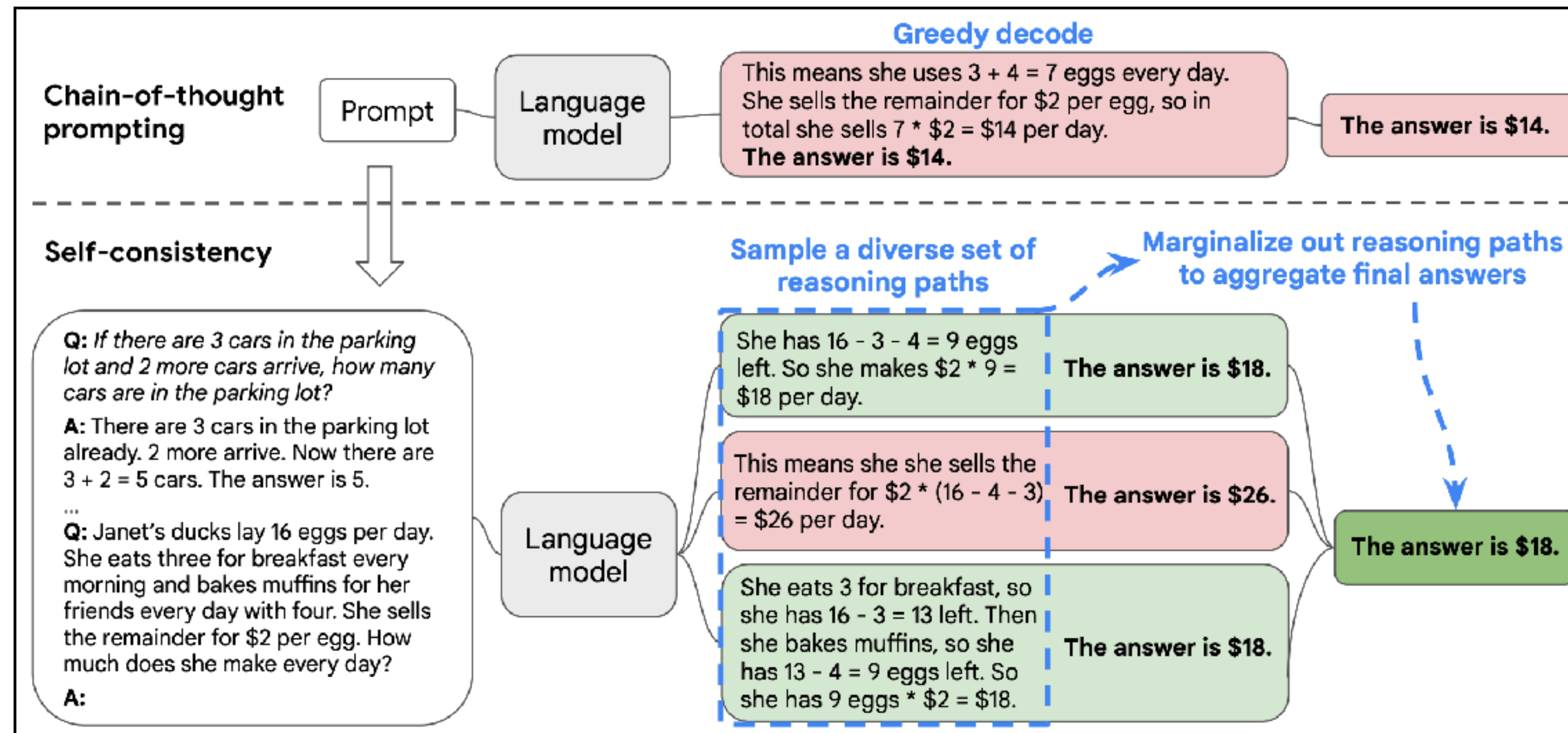
Adding all the odd numbers (15, 5, 13, 7, 1) gives 41. The answer is False.



# Self-consistency prompting

## RECAP

- ▶ few-shot CoT prompting with self-generate CoT sequences (greedily)
- ▶ aggregation over stochastic answer generation

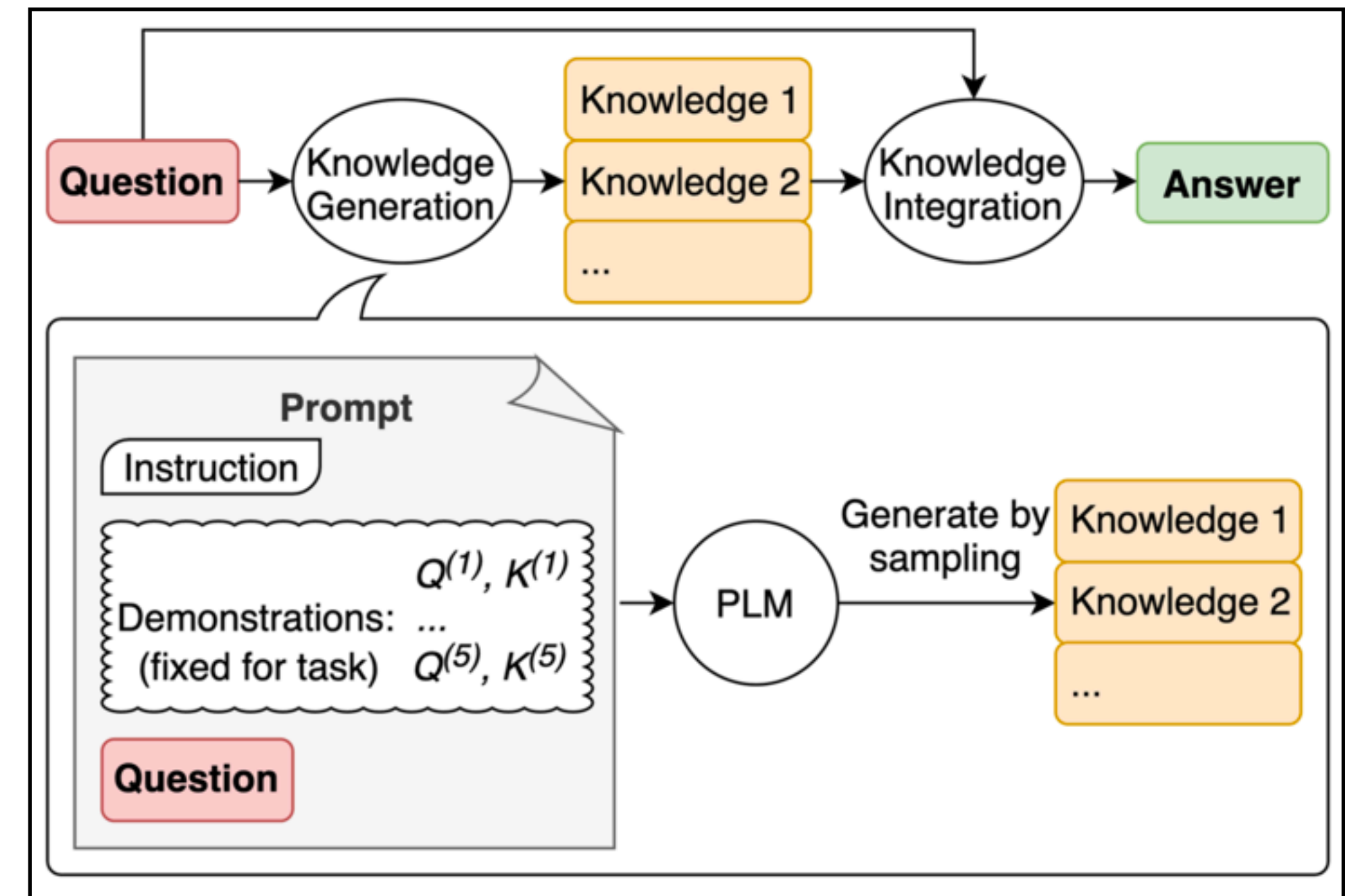


# Generated knowledge prompting

for common sense QA

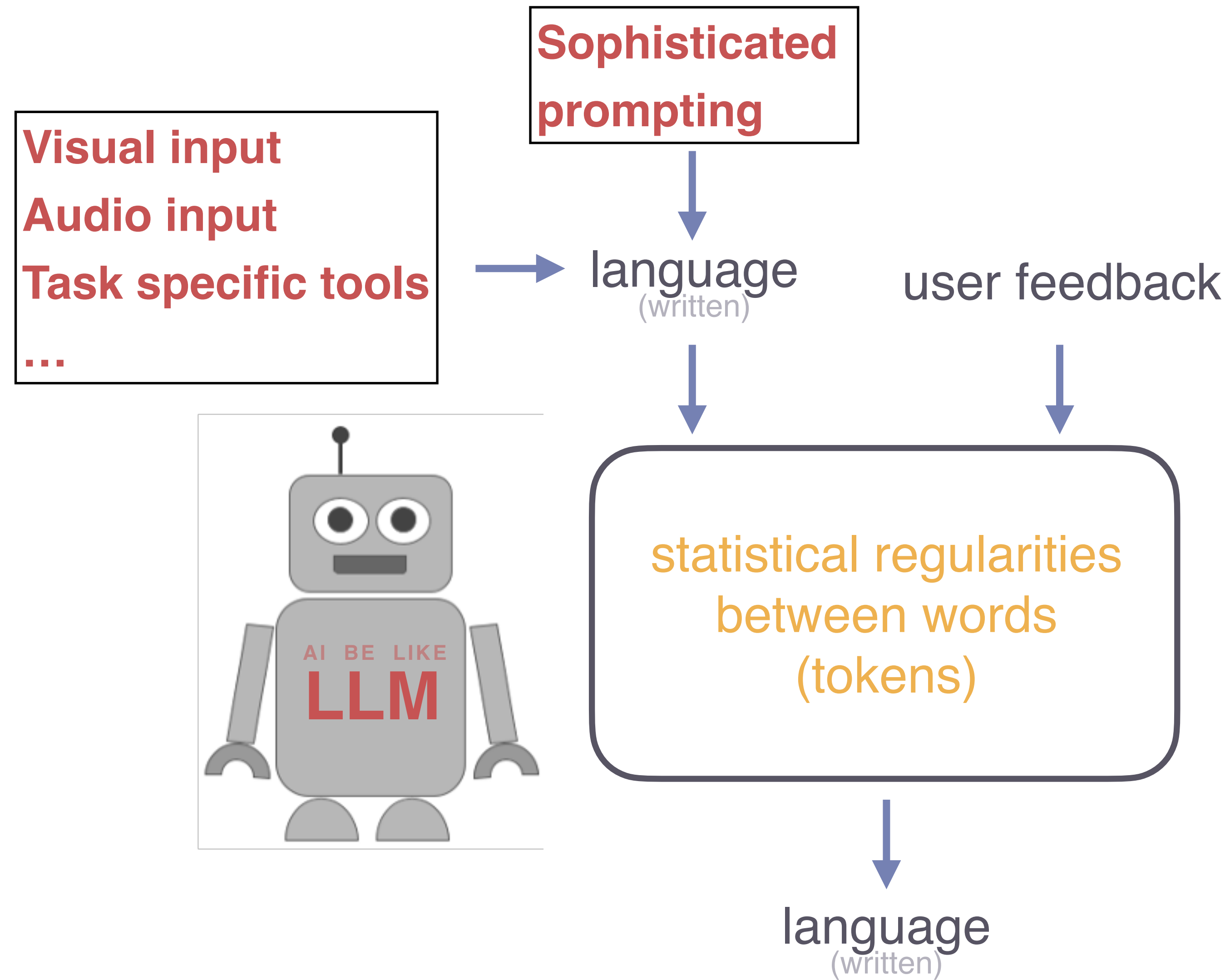
- ▶ generate common knowledge statements  $K$  for  $Q$
- ▶ generate many  $A$ 's for each  $K$
- ▶ final answer to  $Q$  is max of weighted  $A$ 's

## RECAP



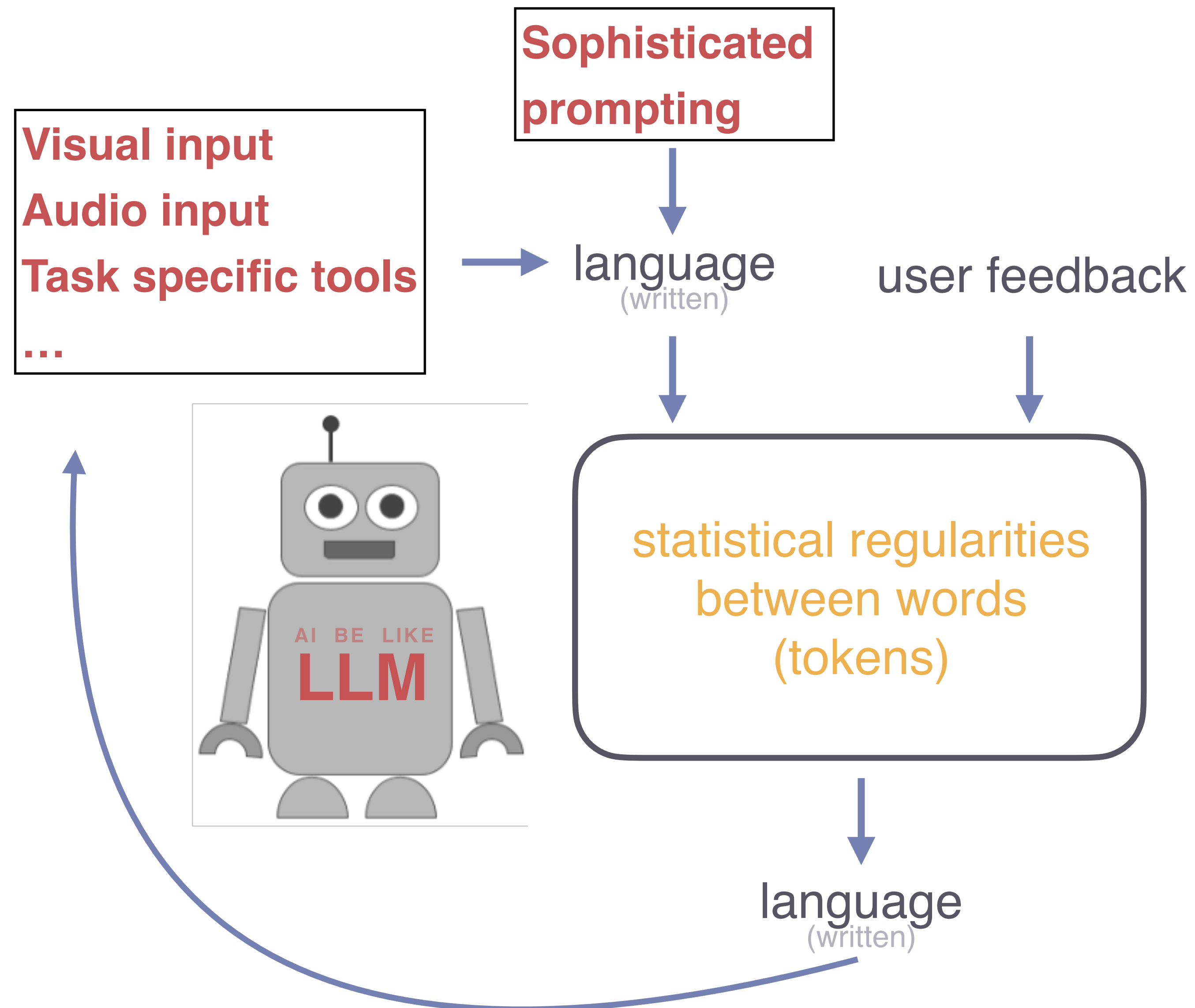
# Prepping prepped LLMs

LLM tools



# Prepping prepped LLMs

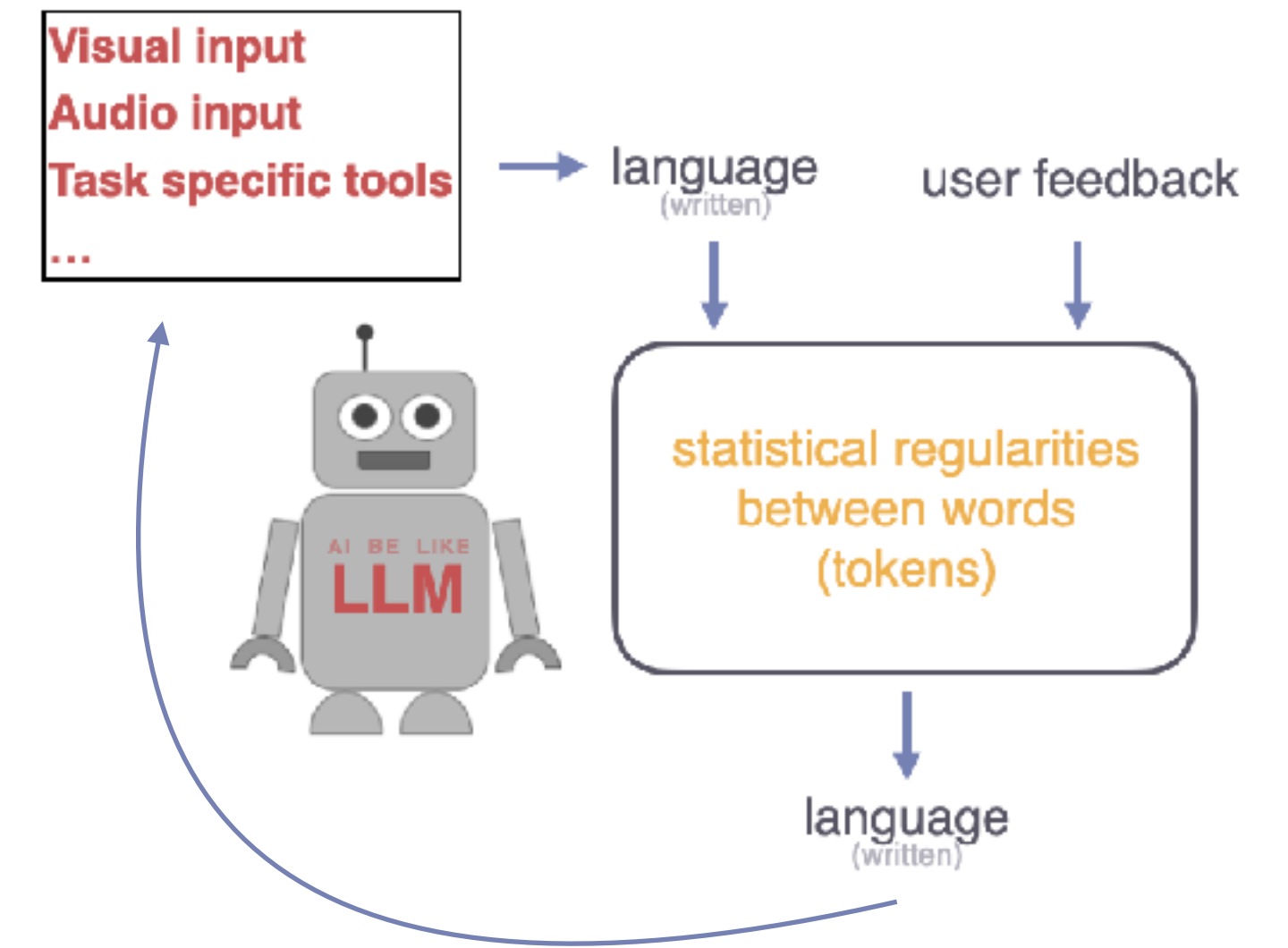
LLM tools



# Prompt, retrieve, repeat!

## Automating LLM requests

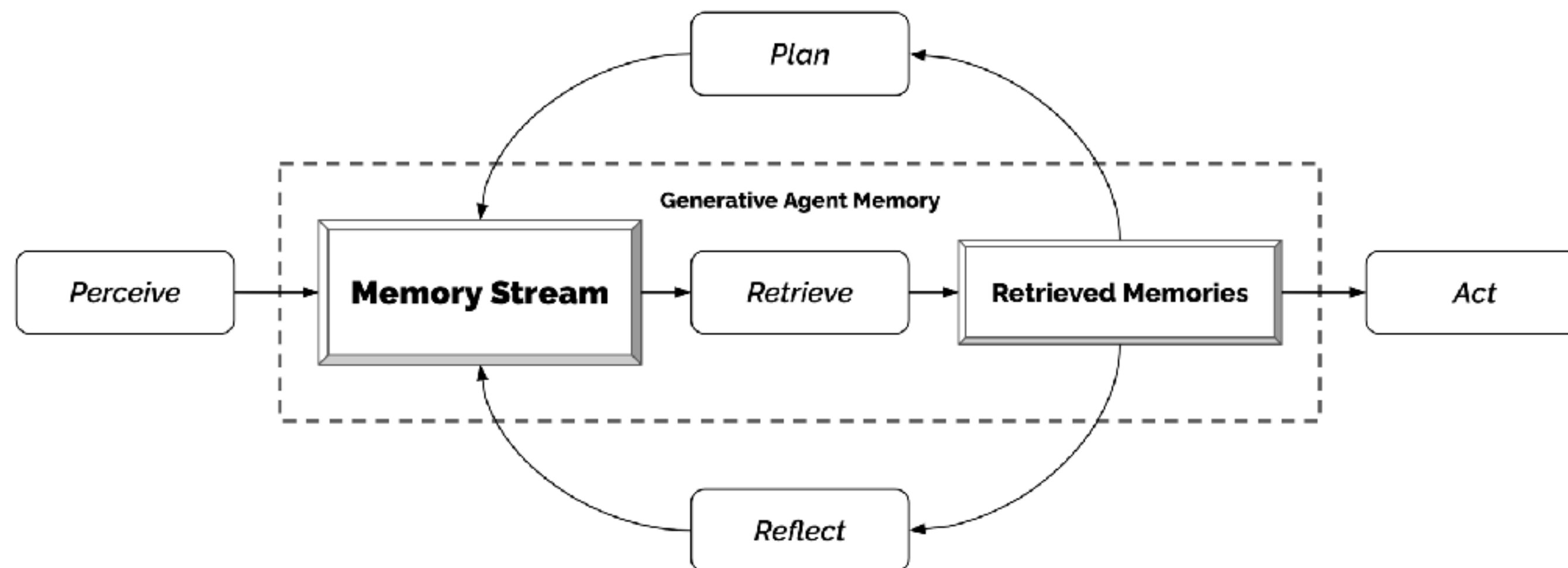
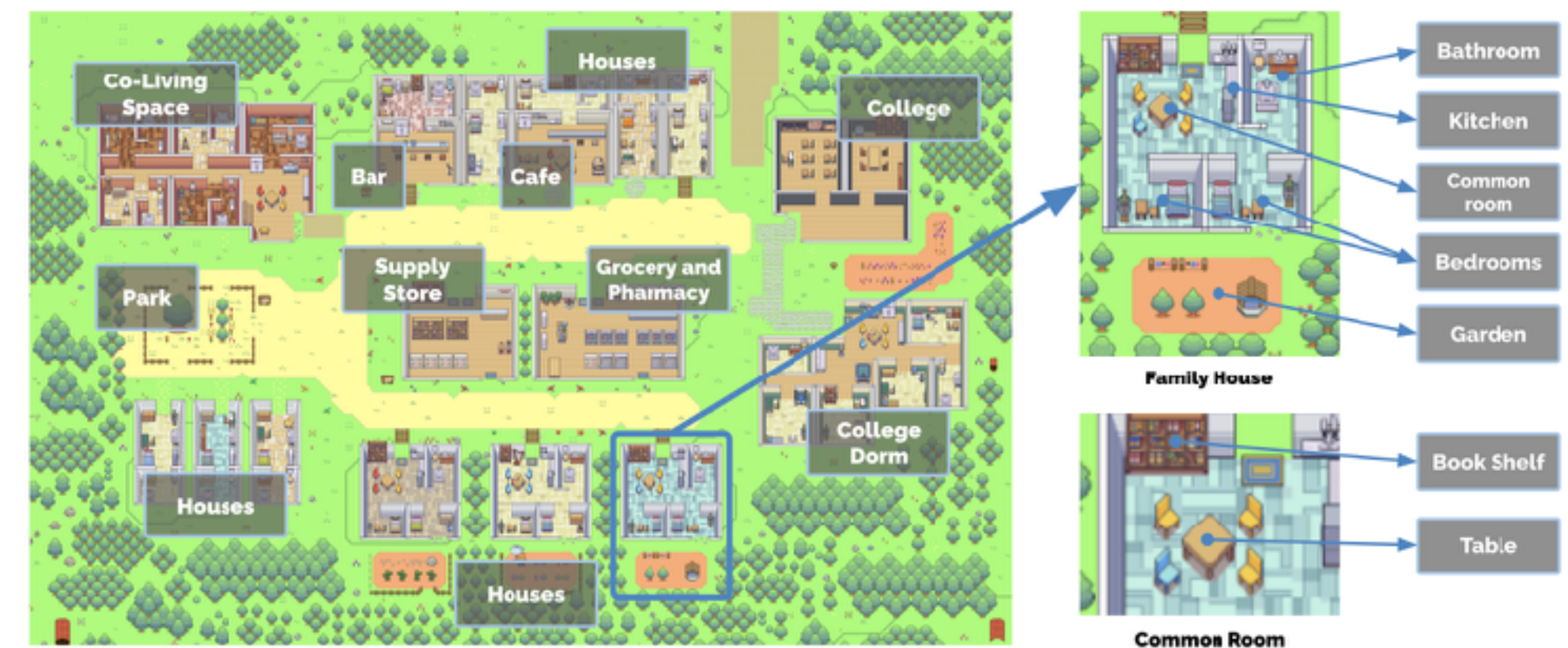
- ▶ generation & I/O with external data sources
  - connection to data bases, APIs & document loaders
  - context & few-shot prompt construction
  - QA, evaluation & selection of responses
- ▶ self-improvement & self-consistency
  - answer option weighting
  - double-checks & revision of text
- ▶ selection of correct processing steps & tools
  - agents
    - planning & evaluation are critical component of problem solving



# Generative agents

Towards autonomous agents???

- ▶ The Sims-style environment Smallville in which LLM based agents dynamically simulate human behavior
- ▶ based on 25 agents (initialized with text bio)
  - interaction with environment via descriptions of actions
  - (emergent) social behavior between agents
  - user intervention via conversation or direct instruction
  - game sandbox movements computed based on LLM output



# AutoGPT, BabyAGI & co

Towards autonomous agents???

## ▶ AutoGPT:

- based on GPT, autonomously generates “thoughts” to achieve a user-specified goal
  - including continuous execution mode
- internet access for searches and information gathering
- memory management
- GPT-4 instances for text generation
- file storage and summarization with GPT-3.5
- extensibility with Plugins
  - TTS, code execution, emails, trading...

## ▶ BabyAGI:

- based on GPT, plans and executes a user-specified task to achieve a goal
- stores subtasks and results in a vector DB
- reprioritises tasks based on results and context

## ▶ JARVIS / HuggingGPT

- a GPT-based controller with different models for solving tasks



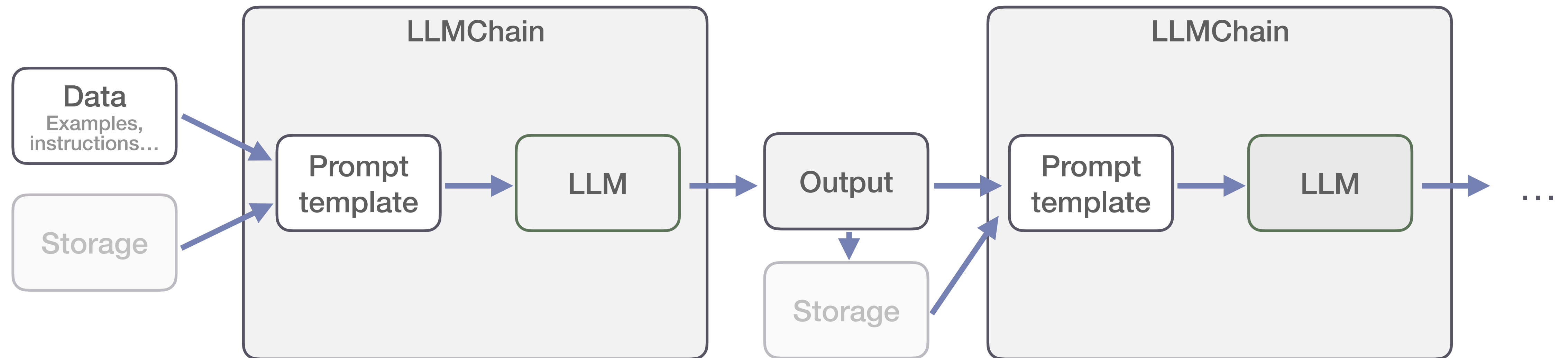
**DO NOT RUN ON YOUR  
MAIN MACHINE!**

# LangChain Chains

\$10 million dollar baby



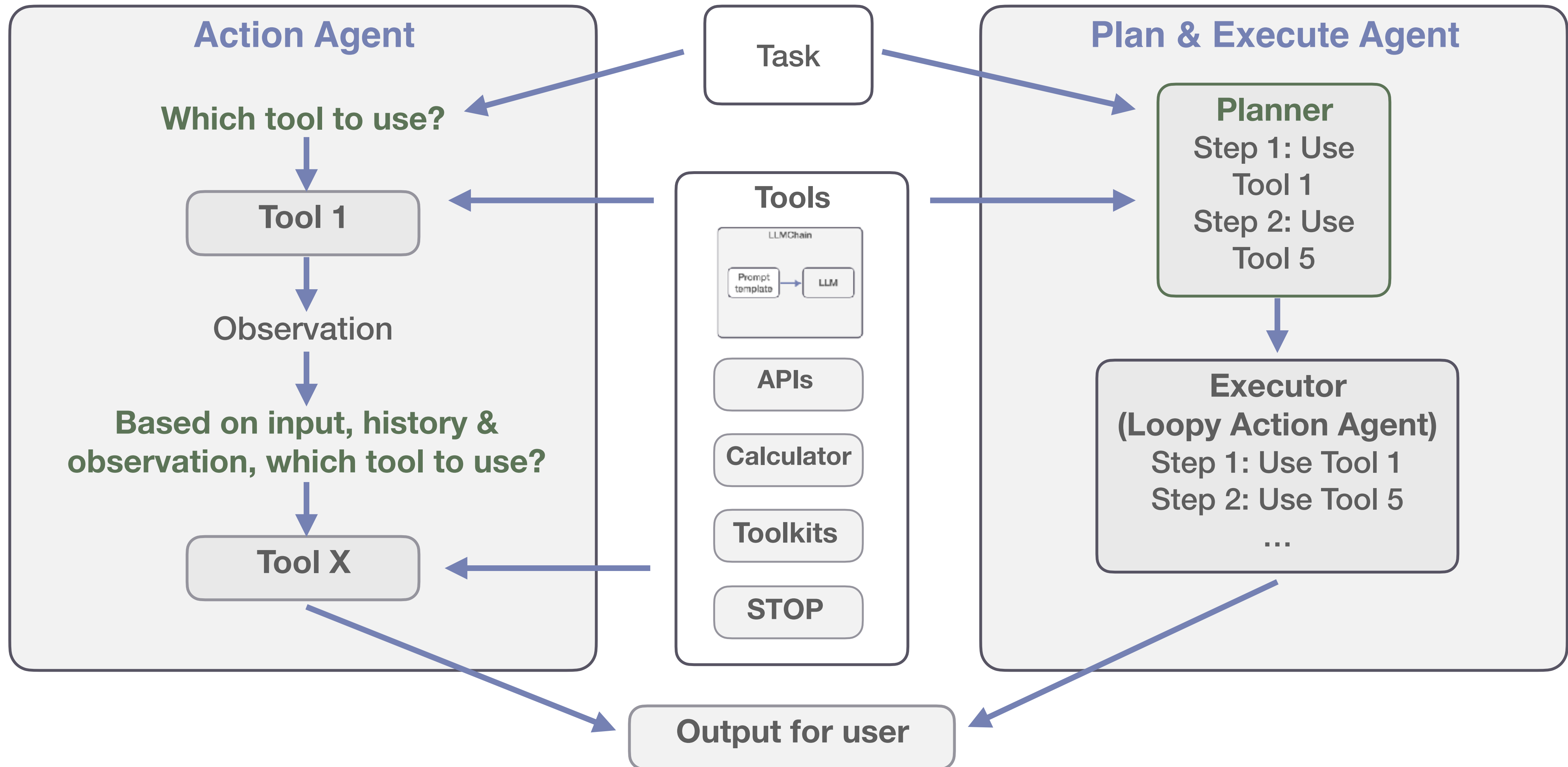
“a framework for developing applications powered by **language models**” can also be *data-aware* and *agentic*





# LangChain Agents

Implementing an unknown chain defined based on input



# LangChain Agents

\$10 million dollar autonomous baby



## Sophisticated prompting

- ▶ single call to LLM
- ▶ call predefined
- ▶ performance optimized via smart prompt engineering
  - examples
  - outputting CoT

## LangChain chain

- ▶ multiple calls to LLM
- ▶ calls predefined
- ▶ performance optimized via repeated use of LLM to complete different tasks
  - I/O
  - calls based on own results & CoT

## LangChain agent

- ▶ multiple calls to LLM
- ▶ calls defined online based on input & results
- ▶ performance optimized via flexible online tool selection
  - agent controller online reasoning about results from different tools
  - calls based on own results, CoT and thoughts (observations)

demo



Using LangChain to build pipelines & agents

# Summary

## Chains & agents

- ▶ we can specify modules of cognitive models which rely on **intuitive knowledge and reasoning** via LLMs
- ▶ the package LangChain enables building LLM powered pipelines via
  - chaining LLM requests
  - providing functionality for an LLM controller selecting tools
  - allowing prompt management



# Homework & Announcements

Optional

Watch an [interview with Ilya Sutskever](#) and think which points you agree / disagree with / are curious about.

No session on May 30th (holiday)!

Session on June 6th **online!**