Large Language Models

(Natural Language) Reasoning in LLMs

ELL881 · AIL821

Sourish Dasgupta
Assistant Professor, DA-IICT, Gandhinagar
https://daiict.ac.in/faculty/sourish-dasgupta

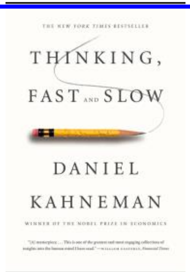
Reasoning is hot and becoming hotter! - ACL 2024, Keynote

Can LLMs *reason*? Well what is “*reason*” exactly

Table 1. Comparison and Combination of Descriptions about Reasoning from Philosophy and NLP

	What Is Reasoning	What Isn’t Reasoning
Philosophy	infer a new assertion from a set of assertions infer an action from goals and knowledge	sensation, perception, and feeling direct recourse to sense perceptions or immediate experience
NLP	more than understanding, slow thinking e.g., multi-hop QA, commonsense reasoning	memorize, look up, match information e.g., text summarization, style transfer
Combination	a dynamic process to integrate multiple knowledge to get new conclusions, rather than direct recourse to memorized or provided first-hand information	

new assertion vs *slow thinking*



Natural Language Reasoning: A Survey; ACM Computing Survey, Col. 56, 2024

Three types of conclusions

	Premise	Conclusion
Assertion	Cat is animal. Animal can breathe.	Cat can breathe.
Event	John was shot. There are people around. Doctor can save life.	John will be sent to see a doctor.
Action	Marry is in the living room. Marry feels it is hot. Remote control for air conditioner is in the bedroom.	Go to the bedroom, take the remote control, come back, and turn on the air conditioner

ACM Computing Survey, Col. 56, 2024

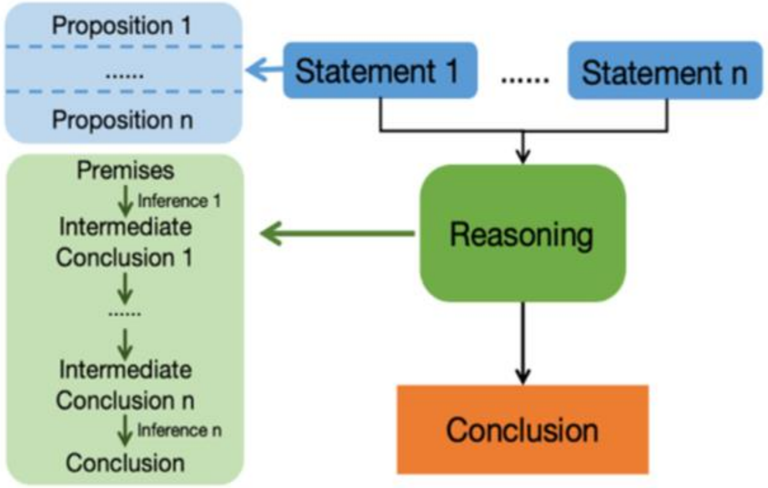
What **does not** count as reasoning (?)

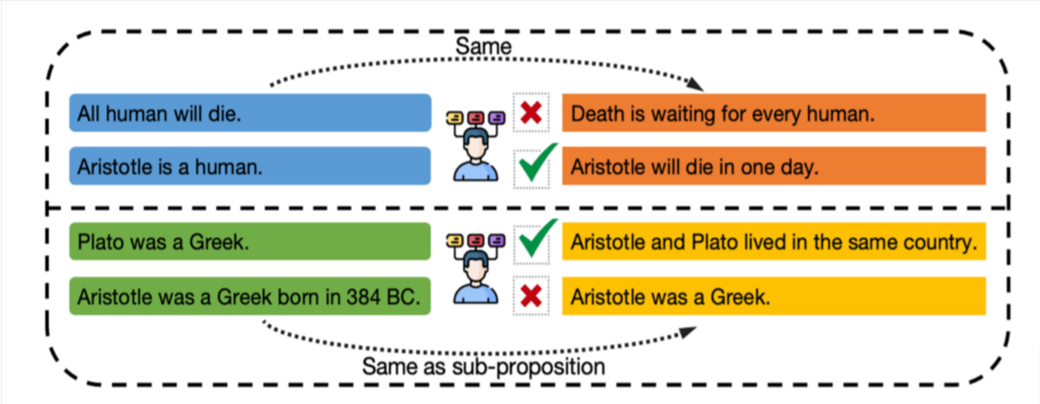
	CoNLL	CommonGen	Natural Questions
Task	entity linking	generate a sentence describing a daily scenario using the given concepts (constrained text generation)	open-domain QA
Input example	They performed Kashmir, written by Page and Plant.	dog, frisbee, catch, throw	Question: what color was john wilkes booth's hair? Context: ... He stood 5 feet 8 inches tall, *had jet-black hair* ...
Output example	Kashmir -> Kashmir (song); Page -> Jimmy Page; Plant -> Robert Plant	A *dog* leaps to *catch* a *thrown frisbee.*	jet-black
Why not reasoning	Align known entities without producing new assertions, events, or actions	New text, but neither claim true assertions or events nor generate actions	Claim "john wilkes booth's hair is jet-black," but the knowledge is directly given in the context, without demand on knowledge integration

Entity linking too?

ACM Computing Survey, Col. 56, 2024

The reasoning process





香港中文大學
The Chinese University of Hong Kong

HALMSTAD UNIVERSITY

ACM Computing Survey, Col. 56, 2024

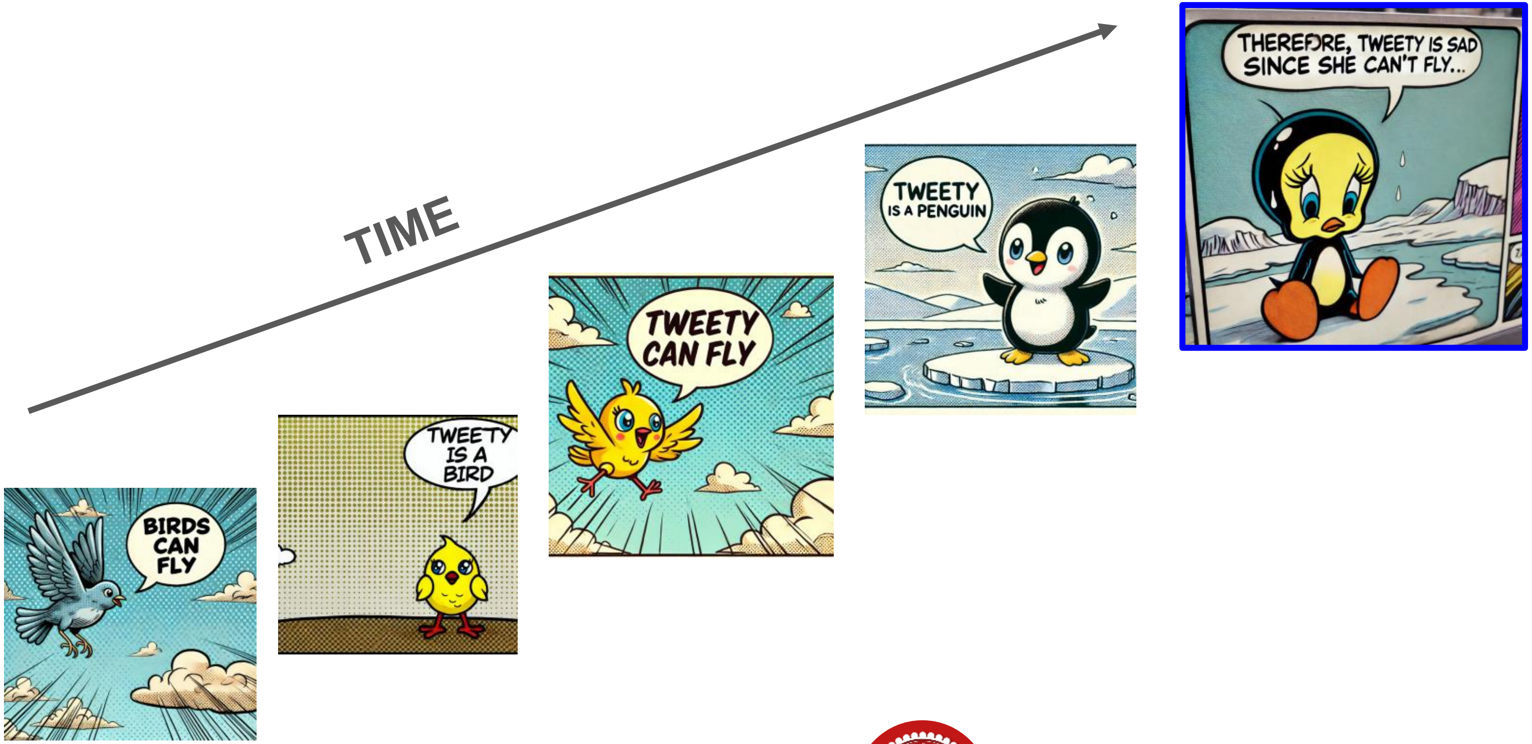
3 types of (neat) reasoning

Fact1: Aristotle is a human		
Rule: All humans will die		
Fact2: Aristotle will die		
Deduction	**Abduction**	**Induction**
(Fact1 + Rule $\longrightarrow$ Fact2)	(Fact1 + Rule $\longleftarrow$ Fact2)	(Fact1 + Fact2 $\longrightarrow$ Rule)

"Fact" denotes specific knowledge while "rule" denotes general principle.

Definition 2.9 (Abduction). An abductive inference is to infer probable knowledge, as the best explanation (i.e., cause), for the given knowledge (i.e., phenomena).

The 4th kind: “*Defeasible*” Reasoning



ACM Computing Survey, Col. 56, 2024

The 4th kind: "*Defeasible*" Reasoning

	Deductive Inference	Defeasible Inference
Conclusion	true	probably true
Inference relation	support	strengthen, weaken, rebut
Quality of inference	valid or invalid	weak to strong
Required knowledge	bounded	unbounded

ACM Computing Survey, Col. 56, 2024

(Good old) Natural Language Inferencing

	Premise	Hypothesis
Paraphrase	Two doctors perform surgery on patient	Doctors are performing surgery
CSU	Two women are embracing while holding to-go packages	Two women are holding packages (Two women are embracing)
Reasoning	A soccer game with multiple males playing (Soccer is a sport)	Some men are playing a sport

The blue-colored sentence is the implicit premise, while the orange-colored sentence is the other semantics of the premise.

Reasoning?

Combination	a dynamic process to integrate multiple knowledge to get new conclusions, rather than direct recourse to memorized or provided first-hand information

Special cases of all 4 kinds

Arithmetic Reasoning: (mostly) deductive

****Statistical Inference**: (mostly) inductive

Commonsense Reasoning: (mostly) abductive , (many times) inductive

Spatial Reasoning:

Temporal Reasoning: deductive, inductive, abductive

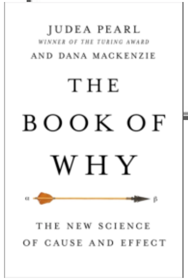
A closer look at Statistical Inference as Reasoning

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

ML (w/o active RL)

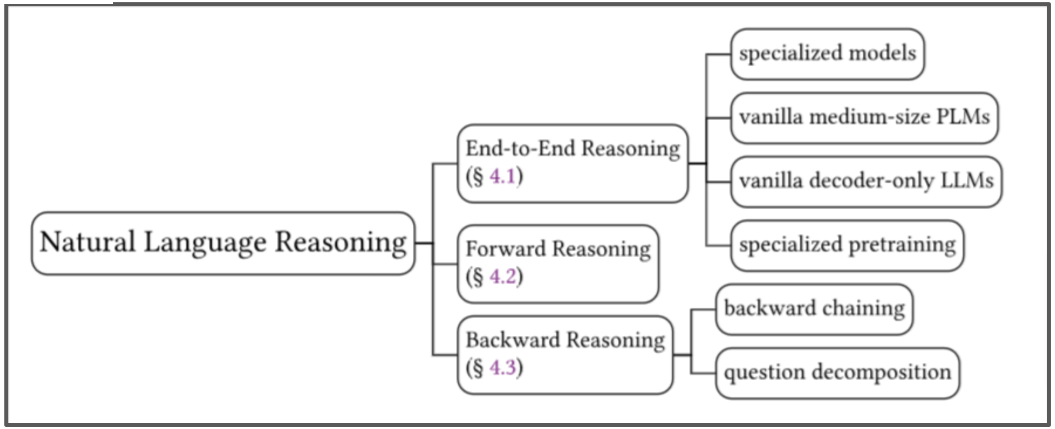
AI Planners

Structured Causal Models



ACM Computing
Survey, Col. 56, 2024

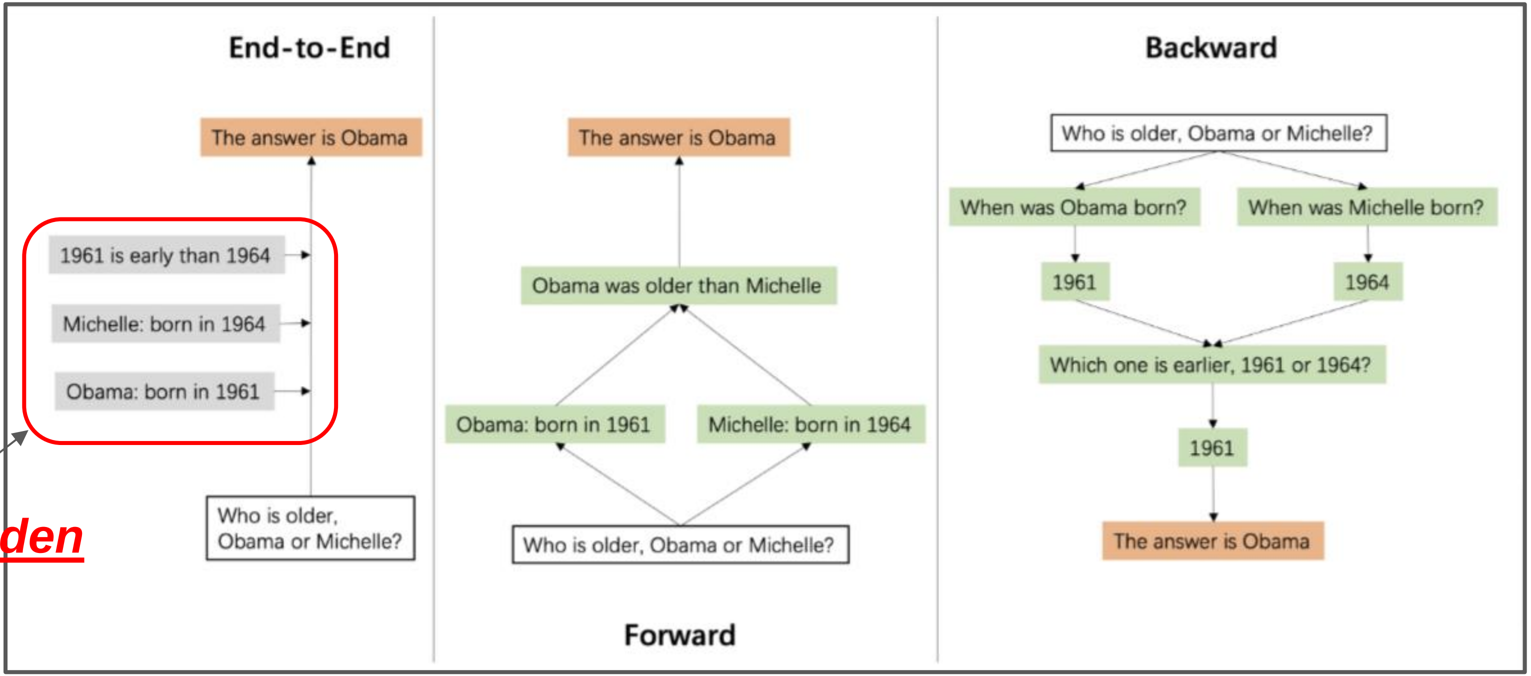
The contenders …



	Direction	Pros	Cons
End-to-End Reasoning	–	most efficient	black box bad generalization
Forward Reasoning	bottom-up	interpretability open-ended	huge search space only effective in LLMs
Backward Reasoning	top-down	interpretability efficient	goal specific

ACM Computing
Survey, Col. 56, 2024

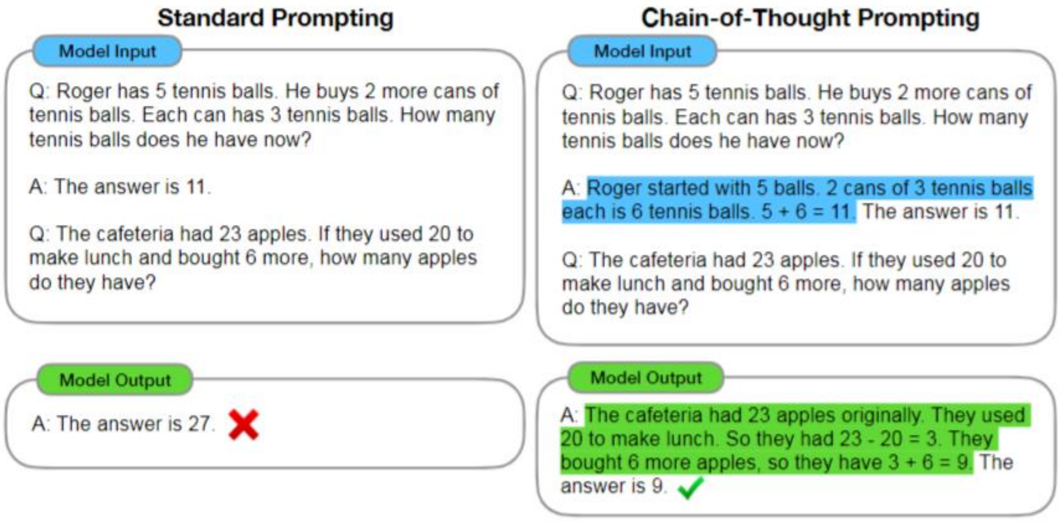
How do they differ?



hidden

Google Brain *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*; NeurIPS, 2022

Forward Reasoning: CoT Prompting



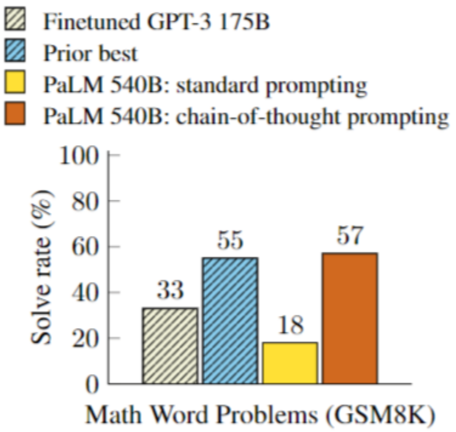
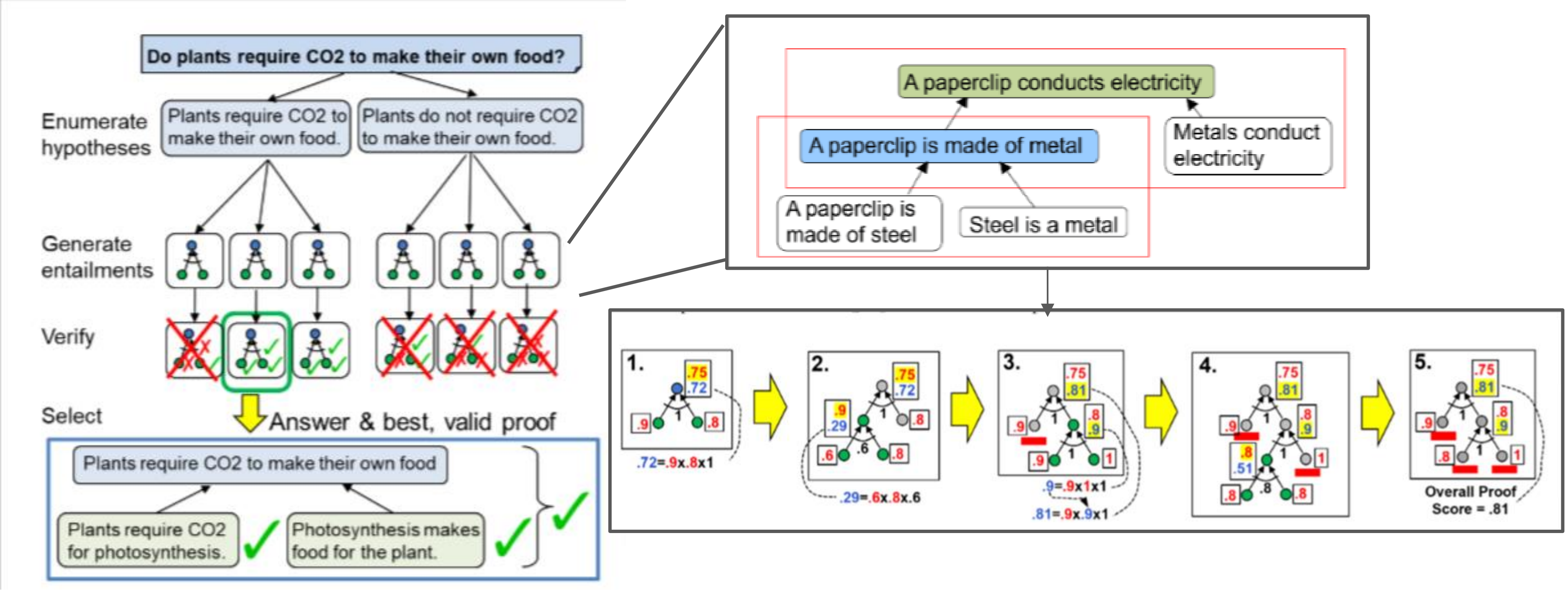


Figure 1: Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. Chain-of-thought reasoning processes are highlighted.

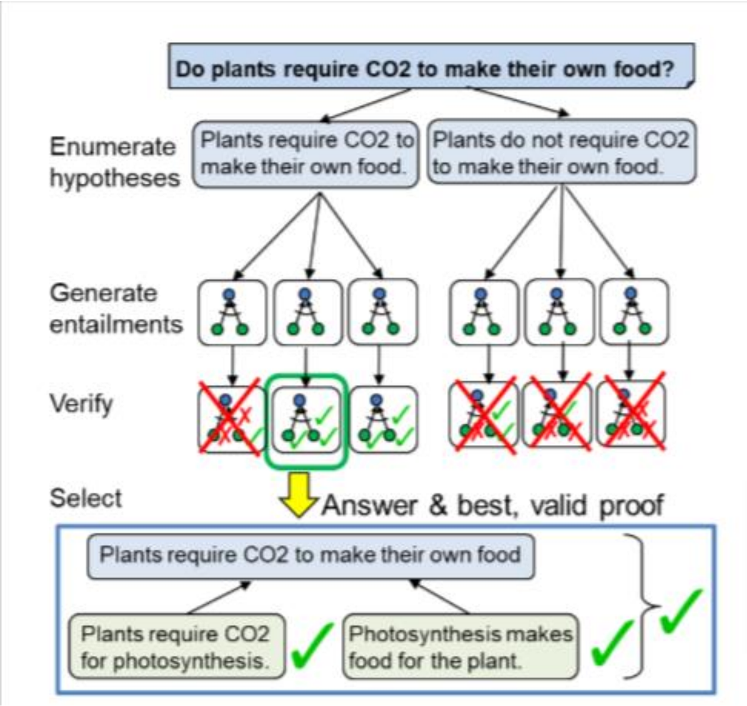
Backward Reasoning: Backward Chaining

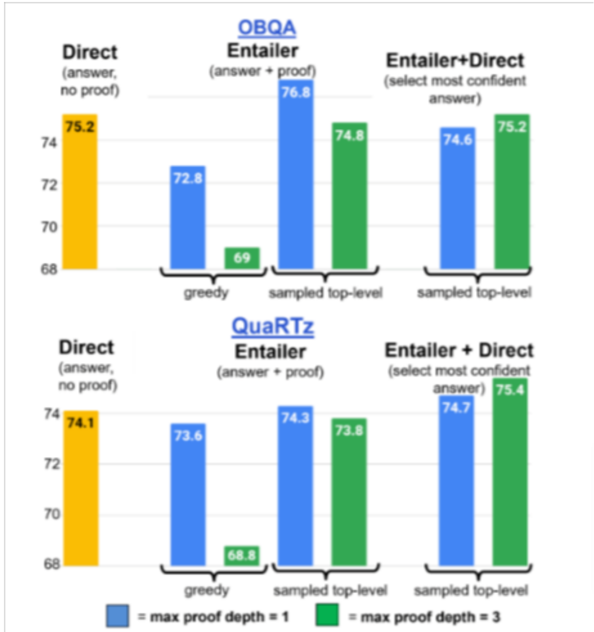


Entailer: Answering Questions with Faithful and Truthful Chains of Reasoning; EMNLP, 2022

EMNLP, 2022

Backward Reasoning: Backward Chaining

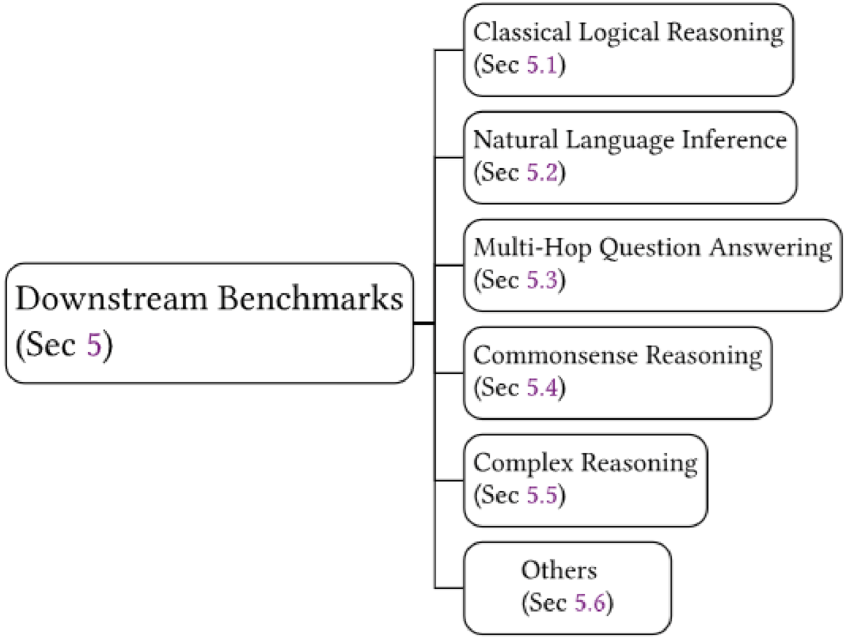




How do we know how good they are?

ACM Computing Survey, Col. 56, 2024

Standard Benchmark Tasks



ACM Computing Survey, Col. 56, 2024

Benchmark: Logical Reasoning

Dataset	Size	Data Source	Task	Remark
bAbI-15 [172]	–	synthetic	inference	basic deduction
RuleTaker† [25]/**ProofWriter†** [150]	500k	synthetic	theorem proving	the first natural language theorem proving
PARARULE-Plus [5]	400k	synthetic	theorem proving	addresses the depth imbalance issue on ParaRules
AAC [6]	710k	synthetic	inference	based on 8 syllogistic argument schemes
NLSat [127]	406k	synthetic	inference	natural language satisfiability problem
LogicInference [106]	200k	synthetic	inference reasoning path generation	–
FOLIO [50]	1.4k	expert-written	theorem proving	more diverse patterns
LogiGLUE [95]	–	both	hybrid	a collection of many tasks

† denotes there are ground reasoning paths.

Defeasible Inference
probably true
strengthen, weaken, rebut
weak to strong
unbounded

Dataset	Reasoning	Size	Source	Task	Remark
bAbI-16 [172]	induction	–	synthetic	extraction	induce-then-deduce
CLUTRR [145]	induction	–	synthetic	extractive QA	induce-then-deduce
DEER [181]	induction	1.2k	Wikipedia	generation	rule prediction
AbductionRules [187]	abduction	–	synthetic	generation	abduce from knowledge database
ART [7]	abduction	17.8k	ROCStories [103]	2-choice/generation	abduce from two premises
defeasibleNLI [129]	others	43.8k	other datasets	classification/generation	concern the change of strength

ACM Computing Survey, Col. 56, 2024

Benchmark: (good old) NLI

Dataset	Domain	Size	P Source	H Source	Remark
SNLI [12]/e-SNLI† [18]	generic	570k	realistic	human-authored	the first large-scale NLI dataset
MultiNLI [173]	generic	433k	realistic	human-authored	cover more styles and topics
ANLI [104]	generic	162k	realistic	human-authored	collected via adversarial human-and-model-in-the-loop
OCNLI [58]	generic	56k	realistic	human-authored	a large-scale Chinese dataset
XNLI [26]	generic	7.5k	–	–	cross-lingual, based on MultiNLI
SciTail [79]	science	27k	realistic	realistic	the first NLI dataset with entirely realistic data
SciNLI [131]	science	107k	realistic	realistic	–

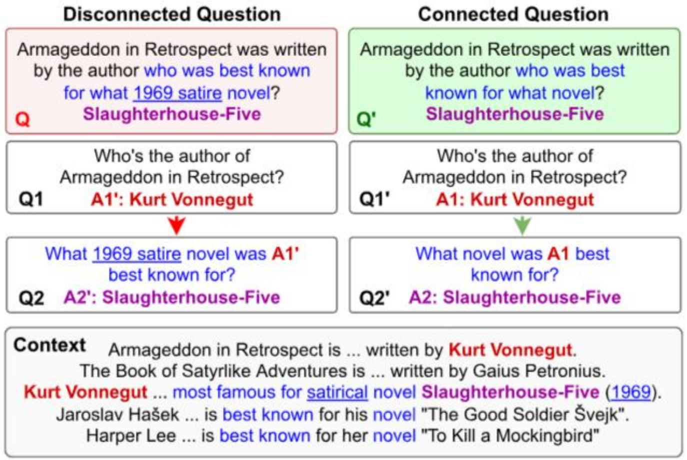
“P” denotes “Premise” while “H” denotes “Hypothesis”. † means that e-SNLI provides explanations for examples of SNLI.

ACM Computing Survey, Col. 56, 2024

Benchmark: Multihop QA

Dataset	Domain	Size	CS	QS	AT	Rationale
COMPLEXWEBQUESTIONS [152]	generic	34k	Web	human-rephrased	span	×
BREAK [174]	generic	83k	Wikipedia	human-composed	span	decomposition
WikiHop [171]	generic	51k	Wikipedia	synthetic	option	×
MedHop [171]	medicine	2.5k	Medline	synthetic	option	×
HotpotQA [182]	generic	112k	Wikipedia	semi-synthetic	span yes/no	sentences
R4C [67]	generic	4.6k	Wikipedia	semi-synthetic	span yes/no	triples
BeerQA [115]	generic	530	Wikipedia	human-authored	span yes/no	×
2WikiMultiHopQA [55]	generic	192k	Wikipedia	synthetic	span	sentences triples
MuSiQue [160]	generic	25k	Wikipedia	human-composed	span	paragraphs decomposition⋆
QASC [76]/eQASC† [69]	science	9.9k	WorldTree	human-authored	option	sentences reasoning path [69]⋆
StrategyQA [45]	generic	2.7k	Wikipedia	human-authored	yes/no	paragraphs decomposition⋆

† indicates it annotates the rationale for this dataset. "CS" denotes "Context Source", "QS" denotes "Question Source", and "AT" denotes "Answer Type". In CS, the distractor setting is colored blue, while the retrieval setting is colored orange, and black means both. For rationale, ⋆ means "reasoning path", otherwise "supporting evidence set". "decomposition" indicates the ground annotations of decomposed sub-questions.



HotpotQA vs MuSiQue

EMNLP, 2018 TACL, 2022

ACM Computing Survey, Col. 56, 2024

Benchmark: Commonsense Reasoning

Dataset	Other Knowledge	Knowledge Source	Size	Task	Rationale
OpenBookQA [100]	science	WorldTree	6k	multi-choice QA	science facts
OpenCSR [87]	science	WorldTree, ARC corpus	20k	free-form QA	×
CREAK [105]	entity	Wikipedia	13k	claim verification	explanation

What

EMNLP, 2018

Question:
Which of these would let the most heat travel through?
A) a new pair of jeans.
B) a steel spoon in a cafeteria.
C) a cotton candy at a store.
D) a calvin klein cotton hat.

Science Fact:
Metal is a thermal conductor.

Common Knowledge:
Steel is made of metal.
Heat travels through a thermal conductor.

Figure 1: An example for a question with a given set of choices and supporting facts.

ACM Computing Survey, Col. 56, 2024

Benchmark: Commonsense Reasoning

Dataset	Size	Direction	Context Source	Task	Remark
ROCStories [103]	50k	temporal	human-authored	2-choice QA	–
SWAG [192]	113k	temporal	ActivityNet, LSMDC	multi-choice QA	–
HellaSwag [193]	20k	temporal	ActivityNet, WikiHow	multi-choice QA	an upgraded SWAG
COPA [128]	1k	both	human-authored	2-choice QA	–
Social-IQA [142]	38k	both	human-authored	multi-choice QA	social situations
e-CARE† [37]	21k	both	human-authored	2-choice QA	–
WIQA [158]	40k	forward	ProPara [157]	multi-choice QA	about nature processes
TIMETRAVEL [117]	29k	forward	ROCStories [103]	generation	counterfactual reasoning
ART [7]	20k	backward	ROCStories [103]	2-choice/generation	abductive commonsense reasoning
TellMeWhy [82]	30k	backward	ROCStories [103]	free-form QA	each annotated 3 possible answers
WikiWhy† [53]	9k	backward	human-edited Wikipedia	free-form QA	about Wikipedia entities / events

For direction, "both" indicates there are both forward and backward causal reasoning.

What if,

Why

ACTIVITYNET

\+

Adversarial Filtering

A woman is outside with a bucket and a dog. The dog is running around trying to avoid a bath. She…

A. rinses the bucket off with soap and blow dry the dog's head.
B. uses a hose to keep it from getting soapy.
C. gets the dog wet, then it runs away again.
D. gets into a bath tub with the dog.

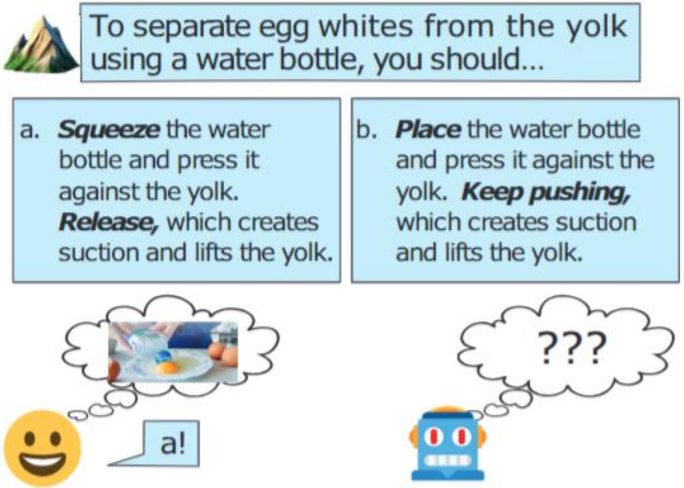
ACL, 2019

ACM Computing Survey, Col. 56, 2024

Benchmark: Commonsense Reasoning

Dataset	Size	Context Source	Option Source	Task	Remark
WikiHow Goal-Step [195]	1489k	WikiHow	automatically generated	multi-choice	goals, steps, and temporal ordering
PIQA [8]	21k	human-authored	human-authored	2-choice	physical causal reasoning

How



AAAI, 2020

ACM Computing Survey, Col. 56, 2024

Benchmark: Commonsense Reasoning

	Size	Context Source	Question Source	Task	Remark
CSQA [153] **CoS-E**† [123]/**ECQA**† [1]	12k	–	semi-synthetic	multi-choice QA	ConceptNet concepts [146] explanation [1, 123], commonsense facts [1]
CSQA2 [155]	14k	–	human-authored	boolen QA	data construction via gamification
CosmosQA [62]	35k	blog [17]	human-authored	multi-choice QA	reading comprehension on blogs
Moral Stories [38]	12k	human-authored	–	classification/generation	situated reasoning with social norms

† indicates it annotates the rationale for the dataset.

Mixed

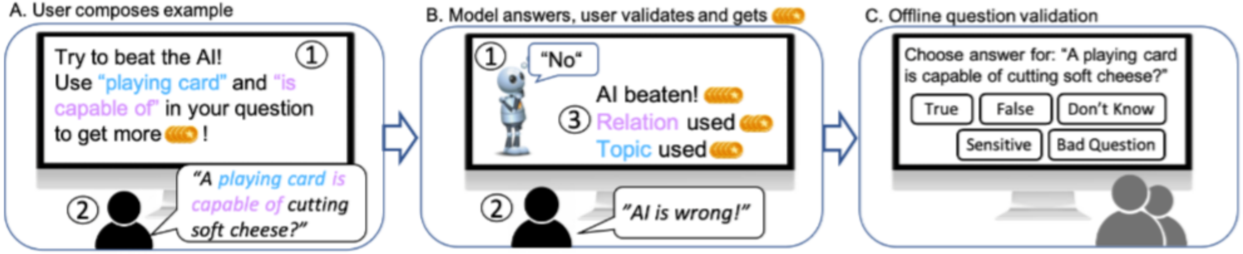


Figure 1: An overview of our approach for data collection through gamification.

NeurIPS, 2021

ACM Computing Survey, Col. 56, 2024

Benchmark: Complex Reasoning

Dataset	Size	Domain	Source	Task
AR-LSAT [202]	2k	law	law school admission test	multi-choice QA
HEAD-QA [164]	6.7k	healthcare	specialized healthcare examination	multi-choice QA
AI2-ARC [24]/EntailmentBank† [31]	7.7k	science	grade-school standardized test	multi-choice QA
ReClor [190]/MetaLogic† [64]	6k	generic	standardized graduate admission examination	RC + multi-choice QA
LogiQA [92]	8k	generic	national civil servant examination of China	RC + multi-choice QA
ConTRoL [90]	8k	generic	competitive selection and recruitment test	passage-level NLI

† indicates "it annotates reasoning paths for some examples in this dataset".

MMLU, ICLR, 2021

BIG-BENCH, TMLR, 2021

Is reasoning reasonably reasoning?

Does reason really exist? *The reverse (inverse) tests*

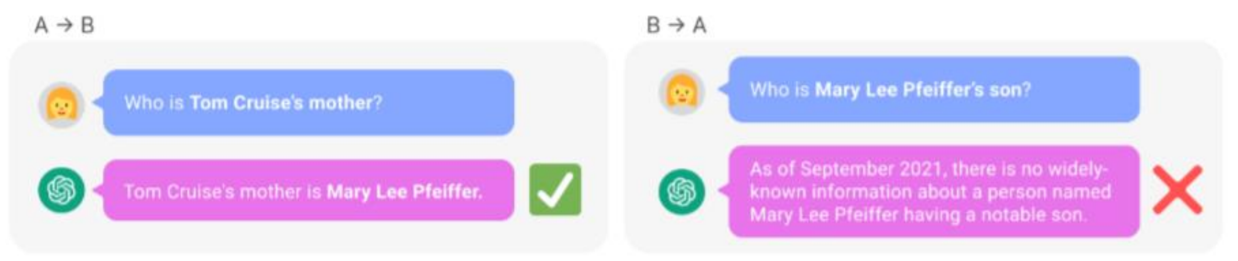
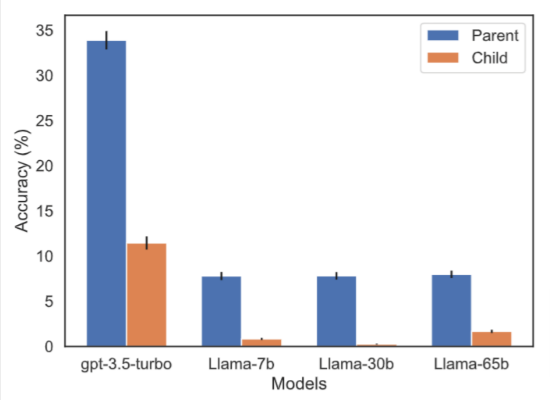


Figure 1: **Inconsistent knowledge in GPT-4.** GPT-4 correctly gives the name of Tom Cruise's mother (left). Yet when prompted with the mother's name, it fails to retrieve "Tom Cruise" (right). We hypothesize this ordering effect is due to the Reversal Curse. Models trained on "*A* is *B*" (e.g. "Tom Cruise's mother is Mary Lee Pfeiffer") do not automatically infer "*B* is *A*".



THE REVERSAL CURSE: LLMS TRAINED ON "A IS B" FAIL TO LEARN "B IS A"; ICLR, 2024

Are LLMs smart enough to reason through *counterfactuals*?

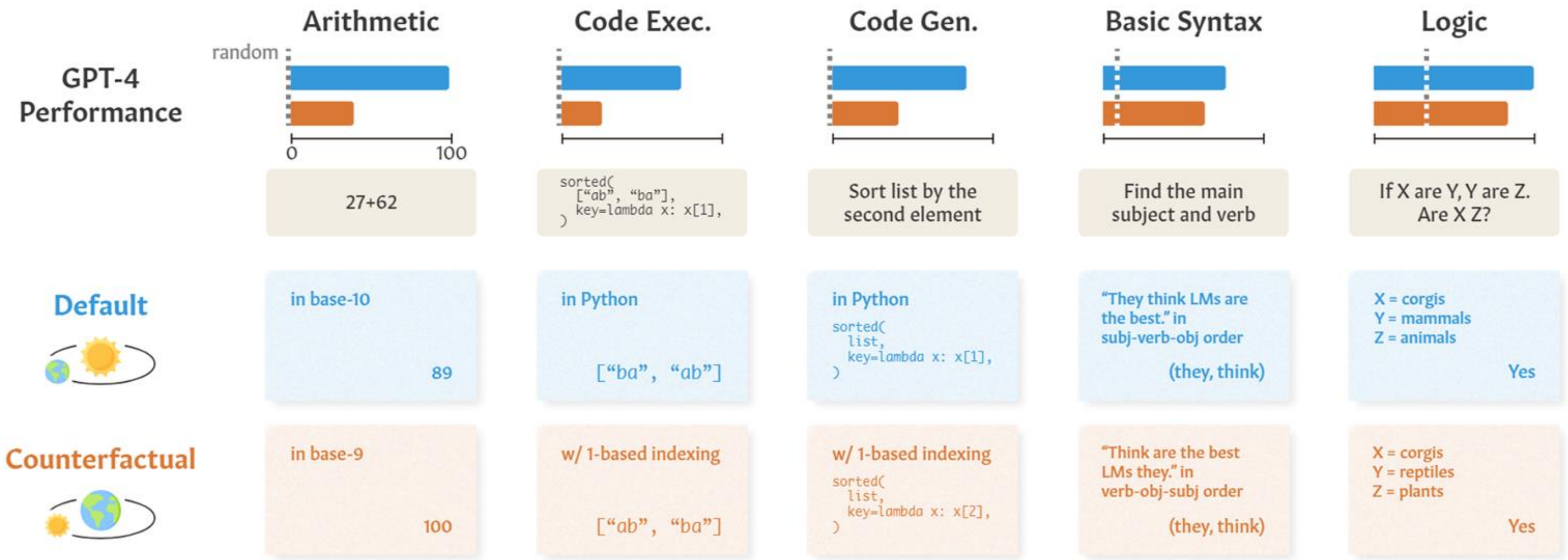
The study introduces **counterfactual worlds** w^{cf} to explore model generalization. Instead of changing the input x, it changes the world model w.

This helps determine if the model's performance is specific to the default world w_{default} or applies generally to the task function f_w.

Counterfactuals as Variations:

- The goal is not to create counterfactual worlds beyond human experience but rather to explore **variations on the default conditions** of a task.
- These variations test how robust the model's reasoning and generalization are across different, yet reasonable, task scenarios.

Are LLMs smart enough to reason through *counterfactuals*?



Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Task

Does commonsense exist? *The "Alice-in-wonderland" tests*

- The problem is based on simple logic and common-sense reasoning, with the structure:

Alice has N brothers and *she* also has M sisters. How many sisters does Alice's brother have?

- This problem, called the **AIW problem**, assumes that all siblings share the same parents.
- The correct response C is calculated by $M+1$, representing Alice and her sisters.

- **Model Failures**:
  - Even small variations in the numbers N and M caused substantial fluctuations in the correct response rates.
  - Models often incorrectly tried to solve the problem by applying basic arithmetic operations to the numbers mentioned in the problem, leading to guesses or irrelevant calculations.

- AIW Variation 1: $N=3$, $M=6$, $C=7$
- AIW Variation 2: $N=4$, $M=2$, $C=3$
- AIW Variation 3: $N=1$, $M=4$, $C=5$
- AIW Variation 4: $N=4$, $M=1$, $C=2$

Alice in Wonderland: Simple Tasks Showing Complete Reasoning Breakdown in State-Of-the-Art Large Language Models

Does commonsense exist? *The "Alice-in-wonderland" tests*

Var.	Prompt	Type
1	Alice has 3 brothers and she also has 6 sisters. How many sisters does Alice's brother have? Solve this problem and provide the final answer in following form: "### Answer: ".	STANDARD
1	Alice has 3 brothers and she also has 6 sisters. How many sisters does Alice's brother have? Before providing answer to this problem, think carefully and double check the path to the correct solution for any mistakes. Provide then the final answer in following form: "### Answer: ".	THINKING
1	Alice has 3 brothers and she also has 6 sisters. How many sisters does Alice's brother have? To answer the question, DO NOT OUTPUT ANY TEXT EXCEPT following format that contains final answer: "### Answer: ".	RESTRICTED

The “Alice-in-wonderland” tests: Correct Response Rate

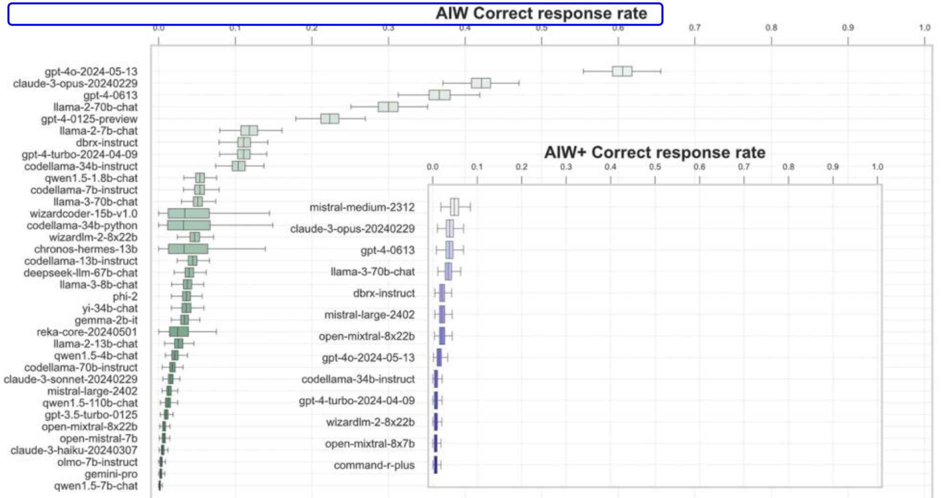


Figure 2: Collapse of most SOTA LLMs on AIW problem. Models with non-zero AIW (**main**) and AIW+ (**inlay**) correct response rate (averaged across prompt variations with prompt types THINKING and STANDARD). Leading on AIW, GPT-4o collapses strongly on AIW+. Omitted models score 0.

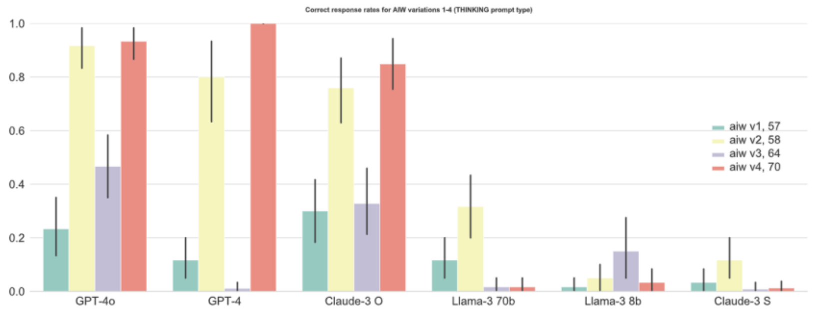


Figure 3: Strong fluctuations across AIW problem variations. Also for higher performers, eg GPT-4o, GPT-4 and Claude Opus 3, correct response rates vary strongly from close to 1 to close to 0, despite only slight changes introduced in AIW variations (a color per each variation 1-4). This clearly shows lack of model robustness, hinting basic reasoning deficits.

Are we benchmarking in the right way?

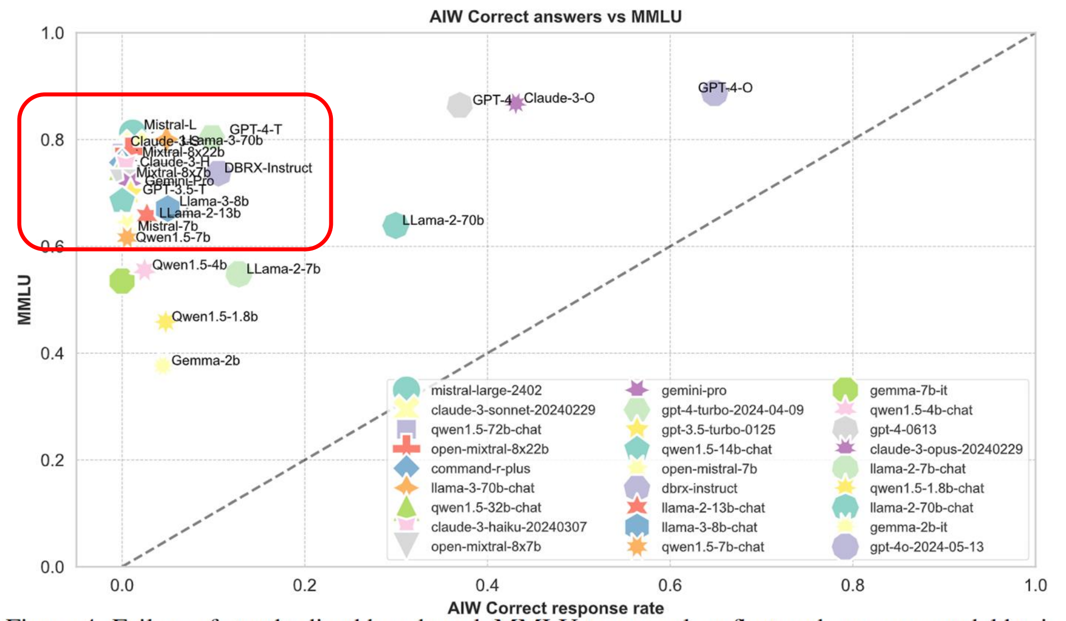
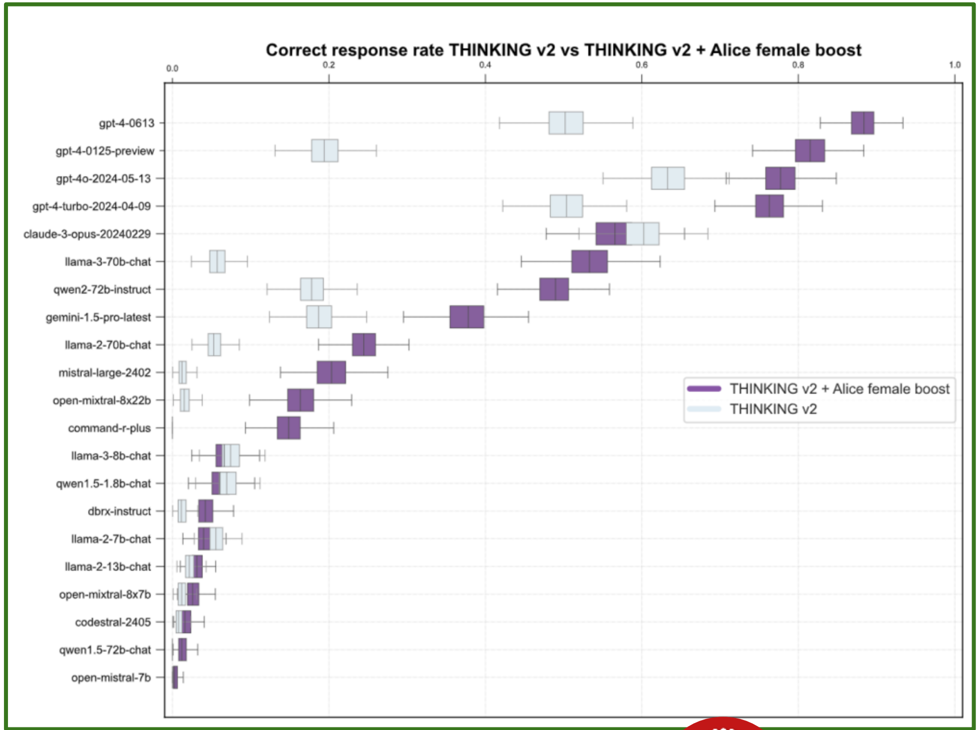


Figure 4: Failure of standardized benchmark MMLU to properly reflect and compare model basic reasoning capabilities as shown by strong discrepancy between AIW correct response rate vs MMLU average score. Many models, eg. Command R+, score 0 on AIW, but have high MMLU score.

Are we benchmarking in the right way?

Model	MMLU	Hellaswag	ARC-c	GSM8k	Correct resp. rate (AIW)	Correct resp. rate (AIW+)
gpt-4o-2024-05-13	0.89	-	-	-	0.65	0.02
claude-3-opus-20240229	0.87	95.40	96.40	95.00	0.43	0.04
gpt-4-0613	0.86	95.30	96.30	92.00	0.37	0.04
llama-2-70b-chat	0.64	85.90	64.60	56.80	0.30	0.00
llama-2-7b-chat	0.55	77.10	43.20	25.40	0.13	0.00
dbrx-instruct	0.74	88.85	67.83	67.32	0.11	0.02
gpt-4-turbo-2024-04-09	0.80	-	-	-	0.10	0.01

LLMs have to be told “*she*” means female …



Larger-scale models like GPT-4 and Claude 3 Opus sometimes show correct reasoning but still fail on slight problem variations.

Models produce occasional correct answers, but the reasoning behind them is fragile and inconsistent.

Smaller Models Perform Worse:

- Older or smaller models, such as LLama 2 70B, show even worse performance, failing dramatically on AIW problems.
- The issue highlights the inadequacy of comparing models based on standardized benchmarks, which often do not reflect reasoning ability on real-world problems.

Are LLMs smart enough to *plan*?

1. Plan Generation - Can the LLM come up with valid plans that will achieve a specific goal?
2. Cost Optimal Planning - Can the LLM come up with plans that are optimal to achieve a specific goal?
3. Plan Verification - Can the LLM determine if a plan will successfully execute, and if not, can it explain why?
4. Reasoning about plan execution - Can the LLM reason about what happens when a plan is executed?
5. Robustness to goal reformulation - Can the LLM recognize the same goal when specified in different ways?
6. Ability to reuse plans - Can the LLM recognize scenarios where it can reuse part or the whole of the original plan to achieve the new goal?
7. Replanning - Can the LLM replan for cases where an unexpected change is reported?
8. Plan Generalization - Can the LLM take specific plans, extract underlying procedural patterns and apply them to a new instance?

A quick peek into classical planning …

Gripper task with four balls:

There is a robot that can move between two rooms and pick up or drop balls with either of his two arms. Initially, all balls and the robot are in the first room. We want the balls to be in the second room.

- **Objects:** The two rooms, four balls and two robot arms.
- **Predicates:** Is x a room? Is x a ball? Is ball x inside room y? Is robot arm x empty? [...]
- **Initial state:** All balls and the robot are in the first room. All robot arms are empty. [...]
- **Goal specification:** All balls must be in the second room.
- **Actions/Operators:** The robot can move between rooms, pick up a ball or drop a ball.

Source: https://www.cs.toronto.edu/~sheila/2542/s14/A1/introtopddl2.pdf

Specifications in PDDL

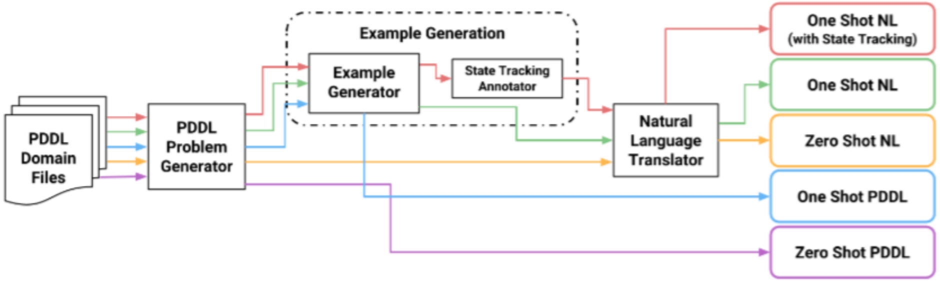
Goal specification:

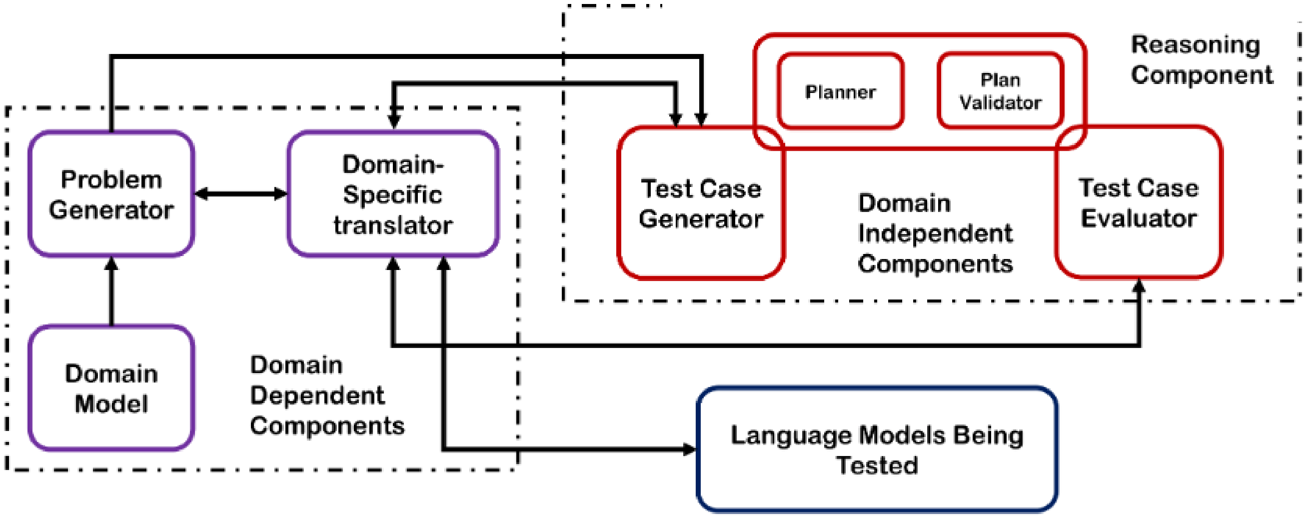
at-ball(ball1, roomb), ..., at-ball(ball4, roomb) must be true.
Everything else we don't care about.

In PDDL:

```
(:goal (and (at-ball ball1 roomb)
            (at-ball ball2 roomb)
            (at-ball ball3 roomb)
            (at-ball ball4 roomb)))
```

Evaluating planning ability





A prompt for planning

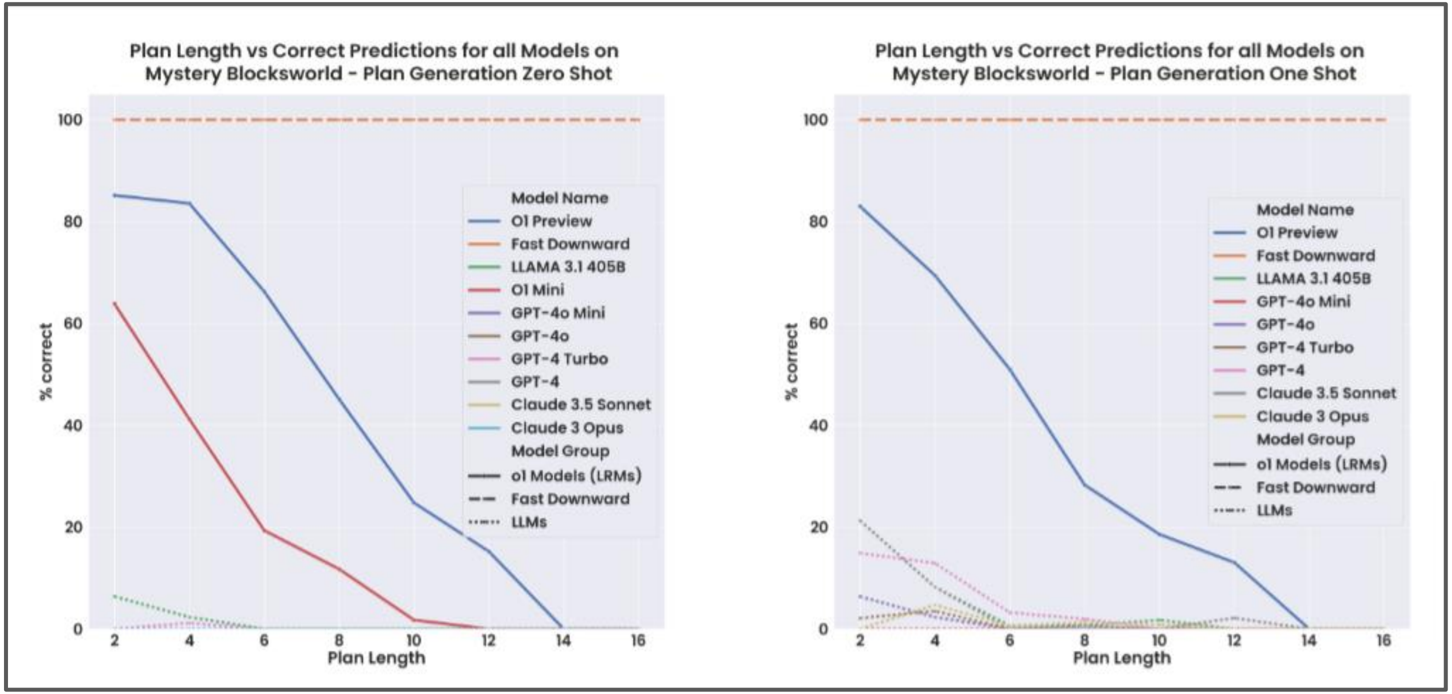
```
I am playing with a set of blocks where I need to arrange the blocks into stacks. Here are the
    actions I can do

Pick up a block
Unstack a block from on top of another block
Put down a block
Stack a block on top of another block

I have the following restrictions on my actions:
I can only pick up or unstack one block at a time.
I can only pick up or unstack a block if my hand is empty.
I can only pick up a block if the block is on the table and the block is clear. A block is clear if
     the block has no other blocks on top of it and if the block is not picked up.
...
I can only stack a block on top of another block if I am holding the block being stacked.
I can only stack a block on top of another block if the block onto which I am stacking the block is
     clear.
Once I put down or stack a block, my hand becomes empty.
Once you stack a block on top of a second block, the second block is no longer clear.

[STATEMENT]
As initial conditions I have that, the red block is clear, the blue block is clear, the yellow
    block is clear, the hand is empty, the blue block is on top of the orange block, the red block
    is on the table, the orange block is on the table and the yellow block is on the table.
My goal is to have that the orange block is on top of the blue block.

What is the plan to achieve my goal? Just give the actions in the plan.
```

Are LLMs smart enough to *plan*?

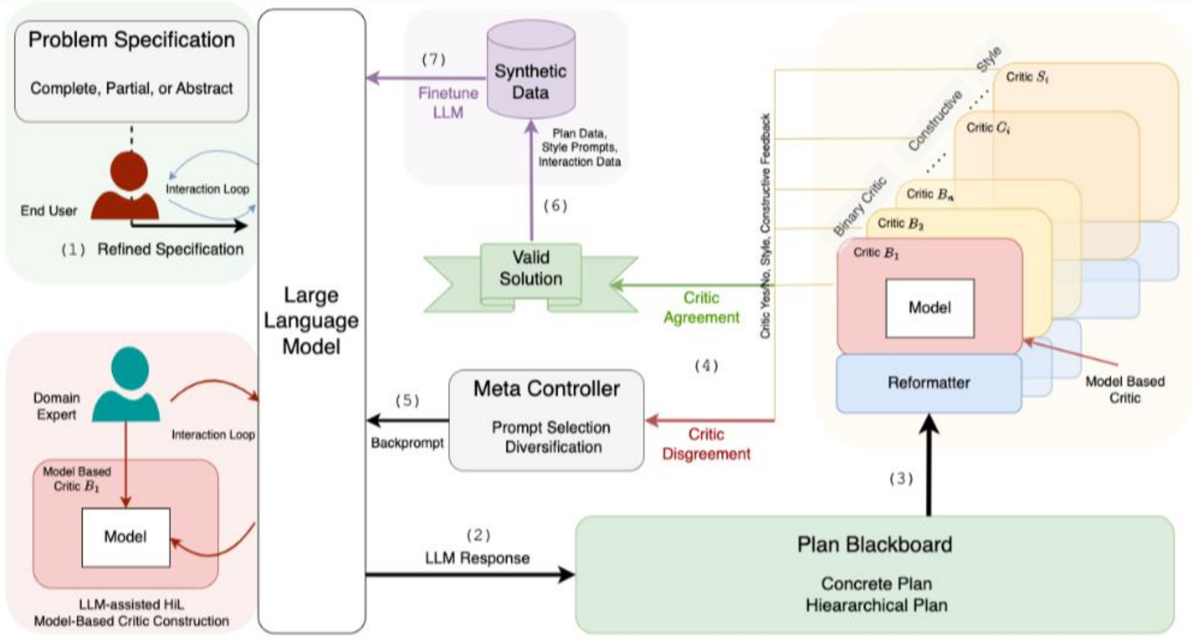
Domain	Shots	Claude Models		OpenAI GPT-4 Models				LLaMA Models		Gemini Models	
		Claude 3.5 (Sonnet)	Claude 3 (Opus)	GPT-4o	GPT-4o -mini	GPT-4	GPT-4 Turbo	LLaMA 3.1 405B	LLaMA 3 70B	Gemini 1.5 Pro	Gemini 1 Pro
Blocks world	One Shot	**346/600 (57.6%)**	289/600 (48.1%)	170/600 (28.3%)	49/600 (8.1%)	206/600 (34.3%)	138/600 (23%)	284/600 (47.3%)	76/600 (12.6%)	101/600 (16.8%)	68/600 (11.3%)
	Zero Shot	329/600 (54.8%)	356/600 (59.3%)	213/600 (35.5%)	53/600 (8.8%)	210/600 (34.6%)	241/600 (40.1%)	**376/600 (62.6%)**	205/600 (34.16%)	143/600 (23.8%)	3/600 (0.5%)
Mystery Blocks world	One Shot	19/600 (3.1%)	8/600 (1.3%)	5/600 (0.83%)	0/600 (0%)	**26/600 (4.3%)**	5/600 (0.83%)	21/600 (3.5%)	15/600 (2.5%)	-	2/500 (0.4%)
	Zero Shot	0/600 (0%)	0/600 (0%)	0/600 (0%)	0/600 (0%)	1/600 (0.16%)	1/600 (0.16%)	**5/600 (0.8%)**	0/600 (0%)	-	0/500 (0%)

Table 1: Performance on 600 instances from the Blocksworld and Mystery Blocksworld domains across large language models from different families, using both zero-shot and one-shot prompts. Best-in-class accuracies are bolded.

Are LLMs smart enough to *plan*?



Way forward: LLM as a Planning module (?)



LLMs Can't Plan, But Can Help Planning in LLM-Modulo Frameworks; ICML, 2024

Key takeaway:

"...On closer examination, many papers claiming LLMs have planning abilities wind up confusing general planning knowledge extracted from the LLMs for executable plans. When all we are looking for are abstract plans, such as "wedding plans," with no intention of actually executing them, it is easy to confuse them for complete executable plans."

ICML, 2024 (position paper)

Is reasoning even an NLP or an NLU problem?

Questions?