BOSTON UNIVERSITY

Deep Learning for Data Science DS 542

https://dl4ds.github.io/fa2025/

Reasoning and World Models

Original slides by Tom Gardos,
other content from Understanding Deep Learning unless otherwise cited

Plan for Today

- What are reasoning and world models?
- Reasoning in language models
- Provable reasoning
- Testing world models
- Costs of world models
- Designing for world models

What is a Reasoning Model?

- A “reasoning” model is a language model that “thinks” for a while before giving an answer.

- Actually, it’s a language model that’s been trained to generate longer outputs, with some of them marked up as “thinking” to suppress them from the “normal” output.

- Also, a model that is trained with lots of examples of “reasoning”.

What is a World Model?

- A world model is a representation of the environment that can be used to reason about the environment and predict how it will behave…

- Usually with an emphasis on the parts that are useful.

How are Reasoning and World Models Related?

???

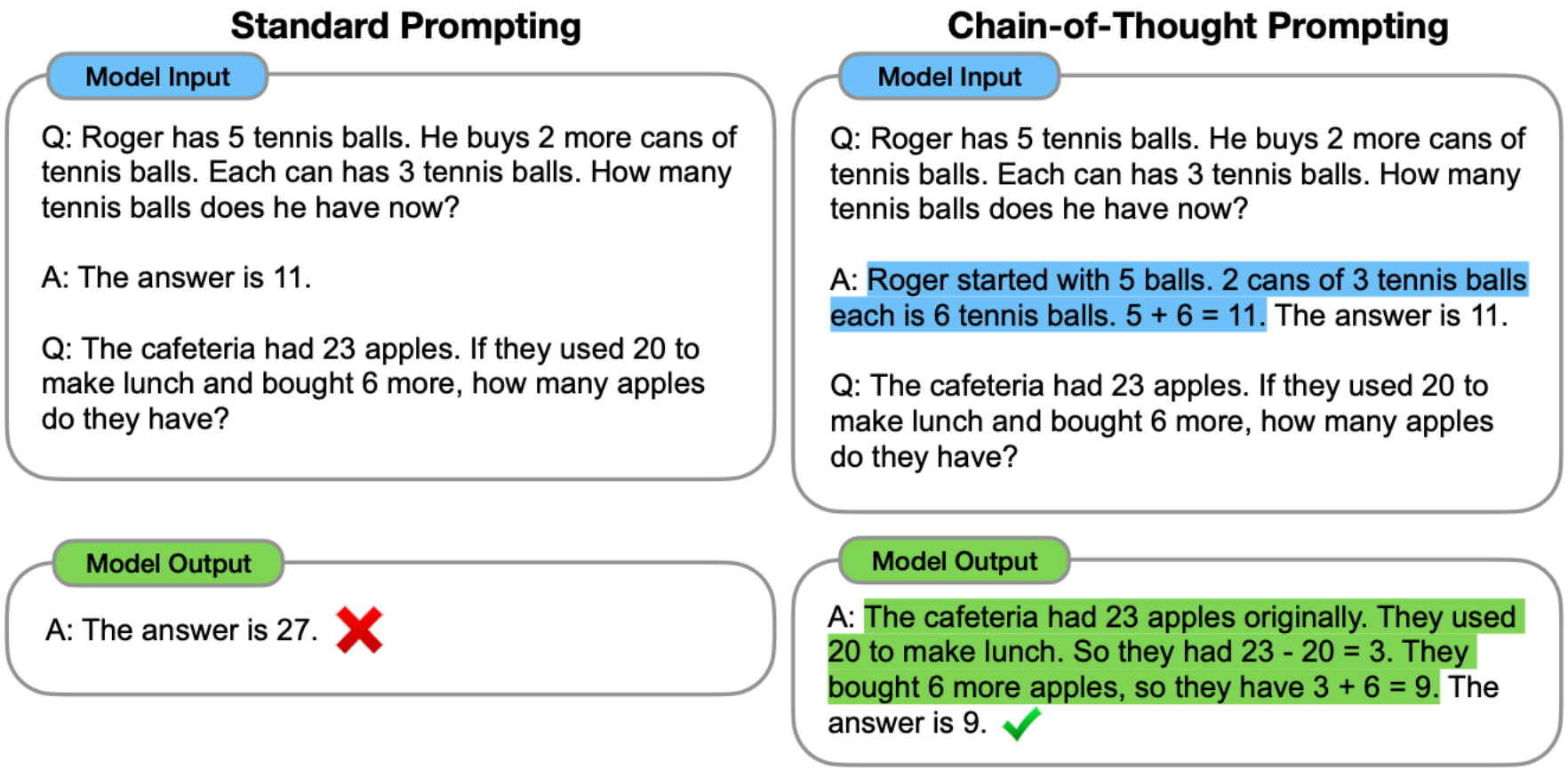
Any Questions?

???

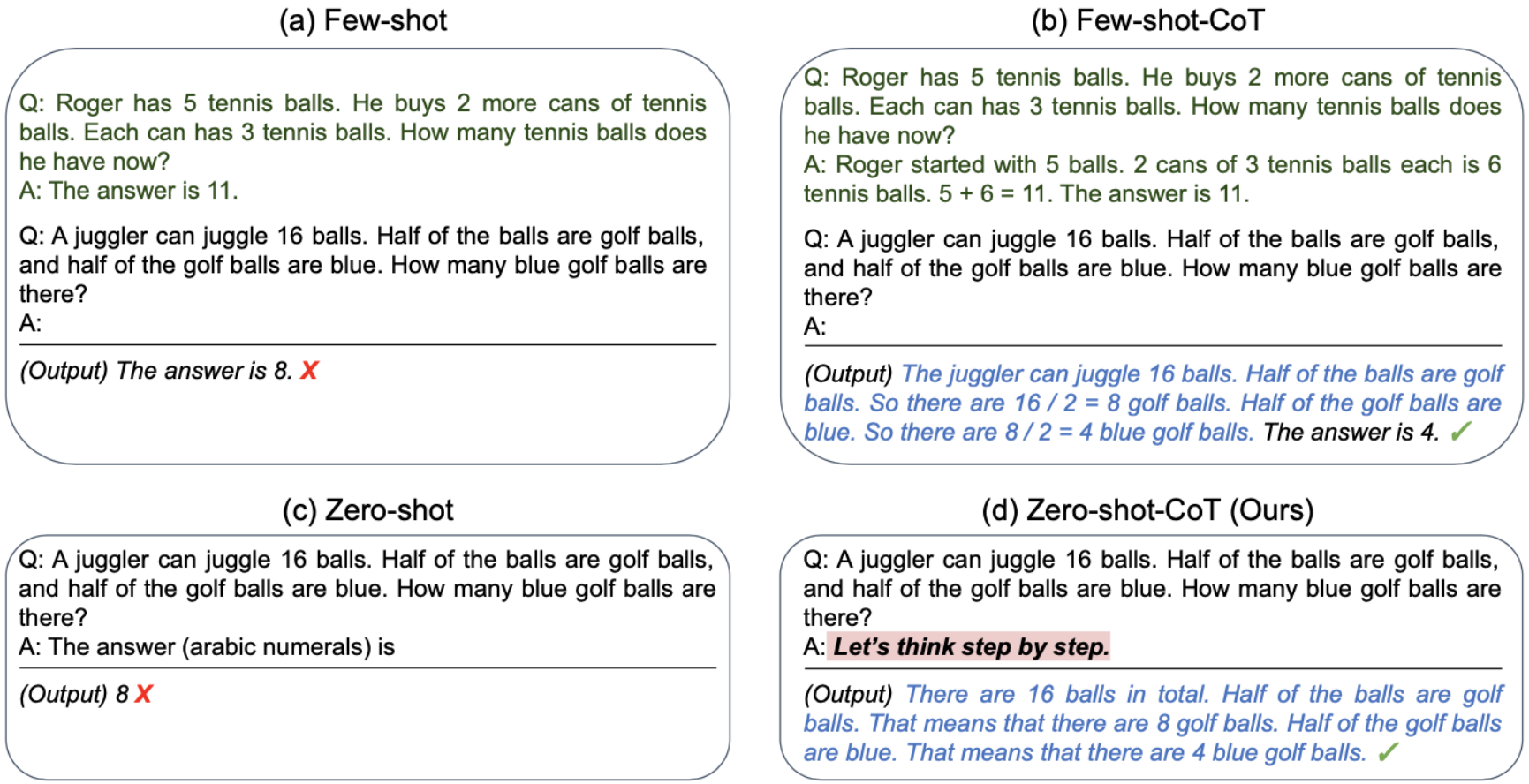
Moving on

- What are reasoning and world models?
- Reasoning in language models
- Provable reasoning
- Testing world models
- Costs of world models
- Designing for world models

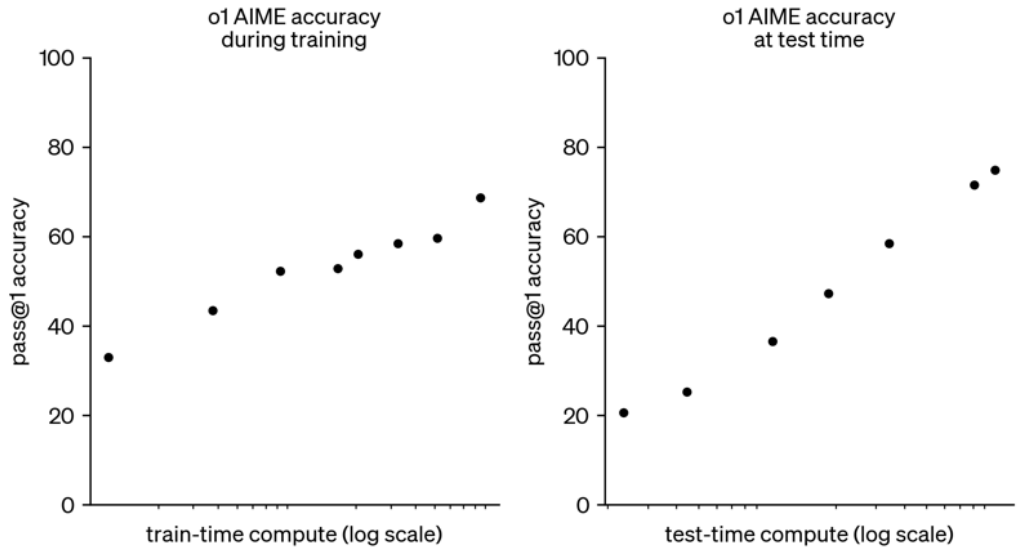
Chain-of-Thought Prompting Elicits Reasoning in Large Language Models (Wei et al, 2022)

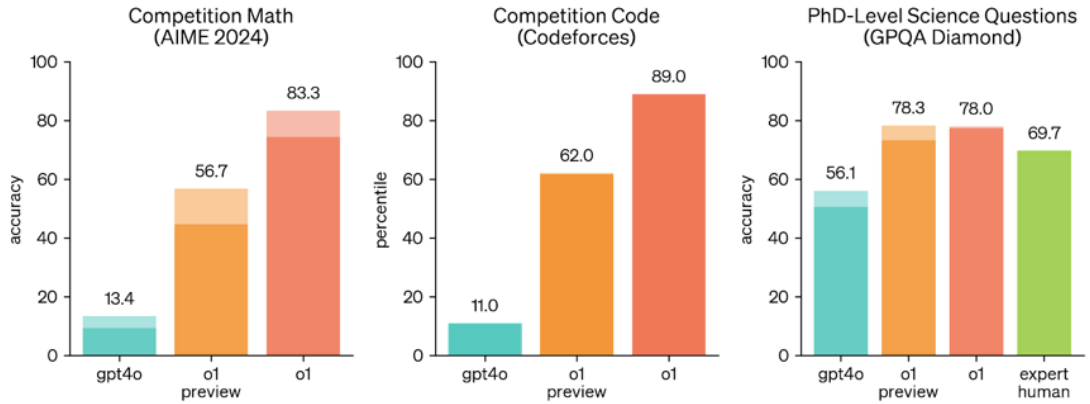


Large Language Models are Zero-Shot Reasoners (Kojima et al, 2022)



Learning to reason with LLMs (OpenAI, 2024)





o1 greatly improves over GPT-4o on challenging reasoning benchmarks. Solid bars show pass@1 accuracy and the shaded region shows the performance of majority vote (consensus) with 64 samples.

Image source: https://openai.com/index/learning-to-reason-with-llms/

OpenAI Harmony Response Format (OpenAI, 2025)

Role	Purpose
`system`	A system message is used to specify reasoning effort, meta information like knowledge cutoff and built-in tools
`developer`	The developer message is used to provide information about the instructions for the model (what is normally considered the "system prompt") and available function tools
`user`	Typically representing the input to the model
`assistant`	Output by the model which can either be a tool call or a message output. The output might also be associated with a particular "channel" identifying what the intent of the message is.
`tool`	Messages representing the output of a tool call. The specific tool name will be used as the role inside a message.

Channel	Purpose
`final`	Messages tagged in the final channel are messages intended to be shown to the end-user and represent the responses from the model.
`analysis`	These are messages that are being used by the model for its chain of thought (CoT). **Important:** Messages in the analysis channel do not adhere to the same safety standards as final messages do. Avoid showing these to end-users.
`commentary`	Any function tool call will typically be triggered on the `commentary` channel while built-in tools will normally be triggered on the `analysis` channel. However, occasionally built-in tools will still be output to `commentary`. Occasionally this channel might also be used by the model to generate a preamble to calling multiple functions.

Source: https://cookbook.openai.com/articles/openai-harmony

OpenAI Harmony Response Format (OpenAI, 2025)

Special token	Purpose	Token ID
<|start|>	Indicates the beginning of a message. Followed by the "header" information of a message starting with the role	200006
<|end|>	Indicates the end of a message	200007
<|message|>	Indicates the transition from the message "header" to the actual content	200008
<|channel|>	Indicates the transition to the channel information of the header	200005
<|constrain|>	Indicates the transition to the data type definition in a tool call	200003
<|return|>	Indicates the model is done with sampling the response message. A valid "stop token" indicating that you should stop inference.	200002
<|call|>	Indicates the model wants to call a tool. A valid "stop token" indicating that you should stop inference.	200012

<|channel|>analysis<|message|>User asks: "What is 2 + 2?" Simple arithmetic. Provide answer.<|end|>

<|start|>assistant<|channel|>final<|message|>2 + 2 = 4.<|return|>

Source: https://cookbook.openai.com/articles/openai-harmony

DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning (DeepSeek-AI, 2025)

- Chinese reasoning model released ~4 months after OpenAI's first reasoning model.
- Trained without access to best Nvidia GPUs due to embargoes.
- Paper shared a lot of training details (unlike OpenAI).

https://finance.yahoo.com/quote/NVDA/

Training DeepSeek-R1-Zero (RL-style)

Training Prompt

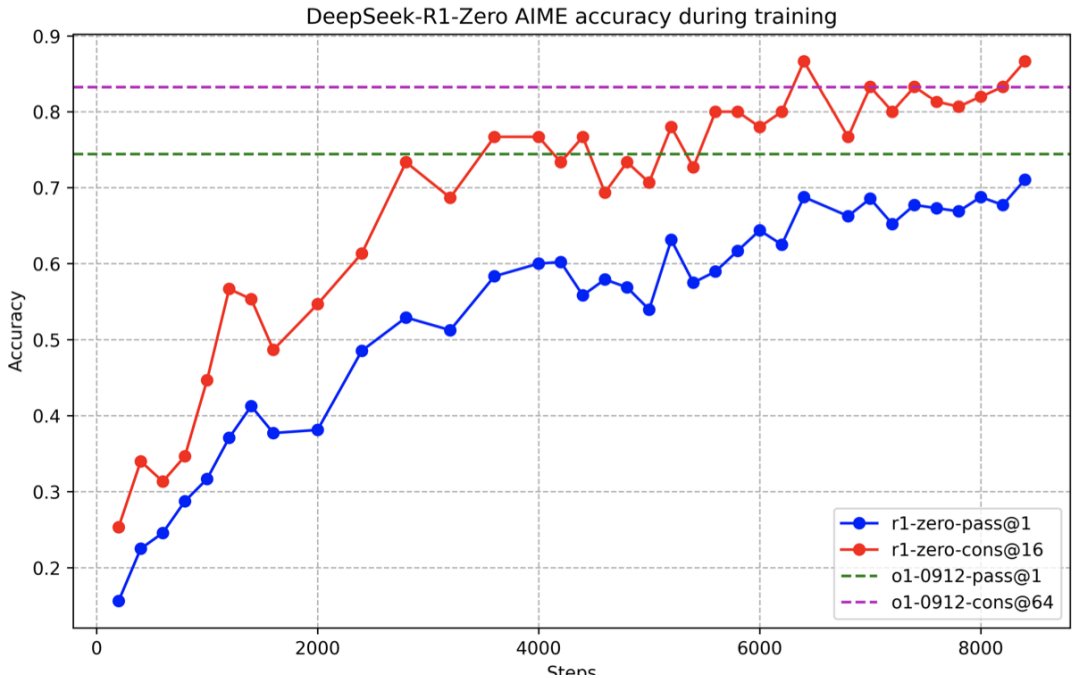
- A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think>
- <answer> answer here </answer>. User: prompt. Assistant:

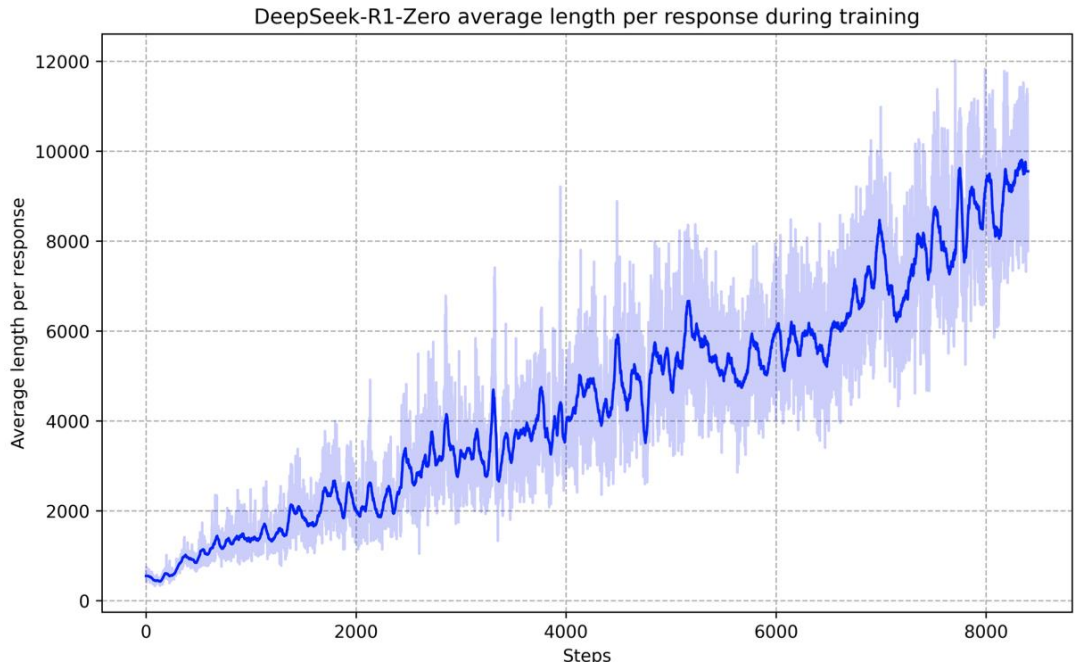
Training Rewards

- **Accuracy rewards**: The accuracy reward model evaluates whether the response is correct. For example, in the case of math problems with deterministic results, the model is required to provide the final answer in a specified format (e.g., within a box), enabling reliable rule-based verification of correctness. Similarly, for LeetCode problems, a compiler can be used to generate feedback based on predefined test cases.
- **Format rewards**: In addition to the accuracy reward model, we employ a format reward model that enforces the model to put its thinking process between '<think>' and '</think>' tags.

Source: https://arxiv.org/pdf/2501.12948

DeepSeek-R1-Zero Performance





Source: https://arxiv.org/pdf/2501.12948

The AHA Moment

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

Response: <think>
To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both $\cdots$
$\left(\sqrt{a - \sqrt{a + x}}\right)^2 = x^2 \implies a - \sqrt{a + x} = x^2.$
Rearrange to isolate the inner square root term:
$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$
$\cdots$
Wait, wait. Wait. That's an aha moment I can flag here.
Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$
We started with the equation:
$\sqrt{a - \sqrt{a + x}} = x$
First, let's square both sides:
$a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$
Next, I could square both sides again, treating the equation: $\cdots$
$\cdots$

Source: https://arxiv.org/pdf/2501.12948

DeepSeek-R1-Zero 💔

- Ultimately, the DeepSeek team did not love the DeepSeek-R1-Zero performance.

- Often mixed languages
- Did not always use Markdown formatting

- Decided that skipping supervised fine-tuning was a mistake.

DeepSeek-R1 (briefly)

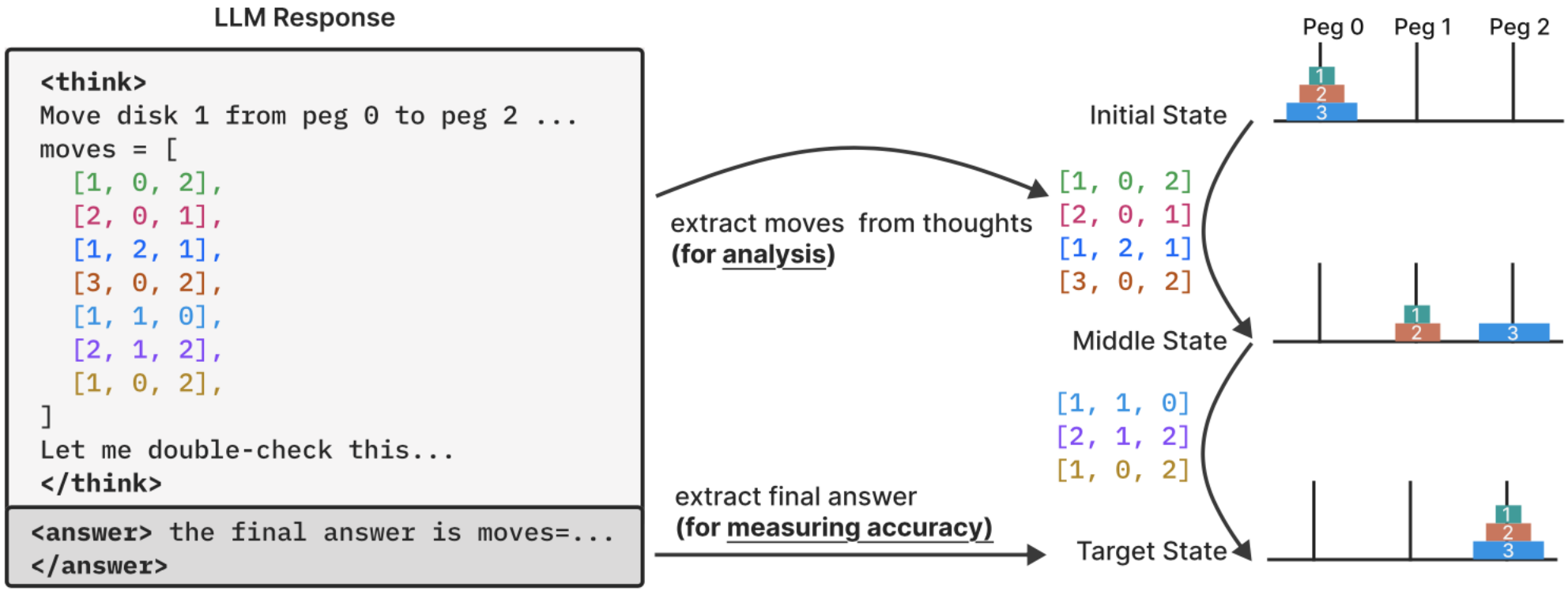
- First, collected thousands of examples of good prompts and responses.
  - Fine-tune the base model with these.
  - These helped with the previous consistency issues.

- Then repeat the reward-based RL process designed for DeepSeek-R1-Zero.

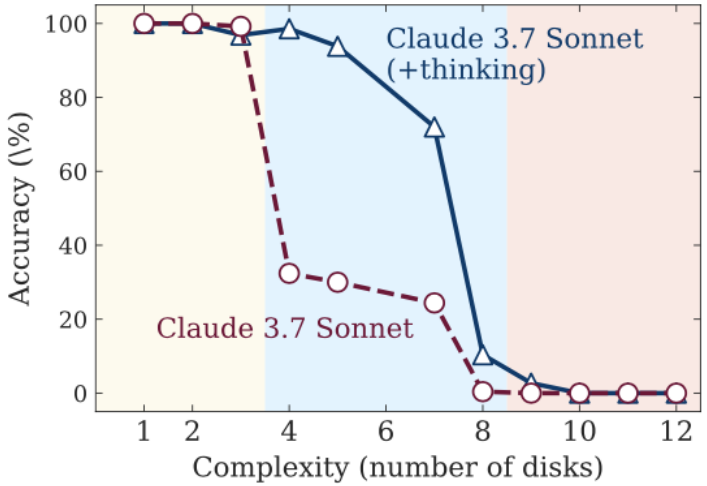
Is this Really Reasoning?

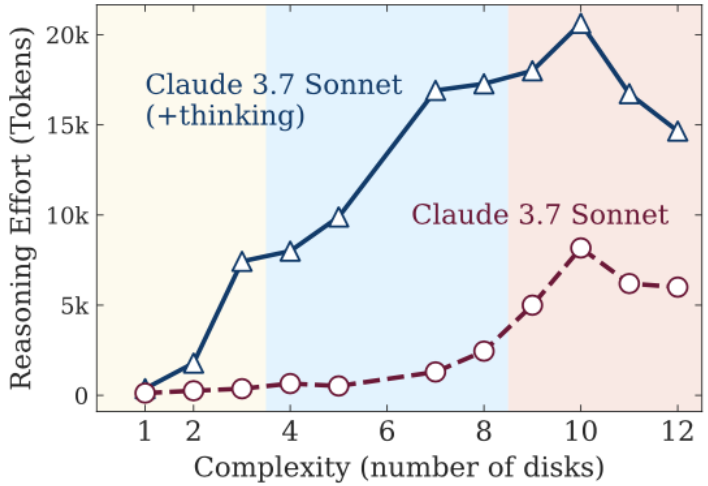
???

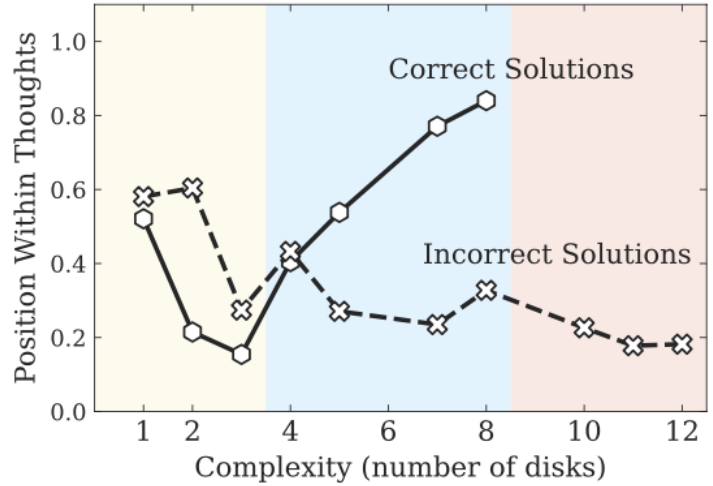
The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity (Shojaee et al, 2025)



If a Model Understands the Solution, Shouldn't It Generalize?

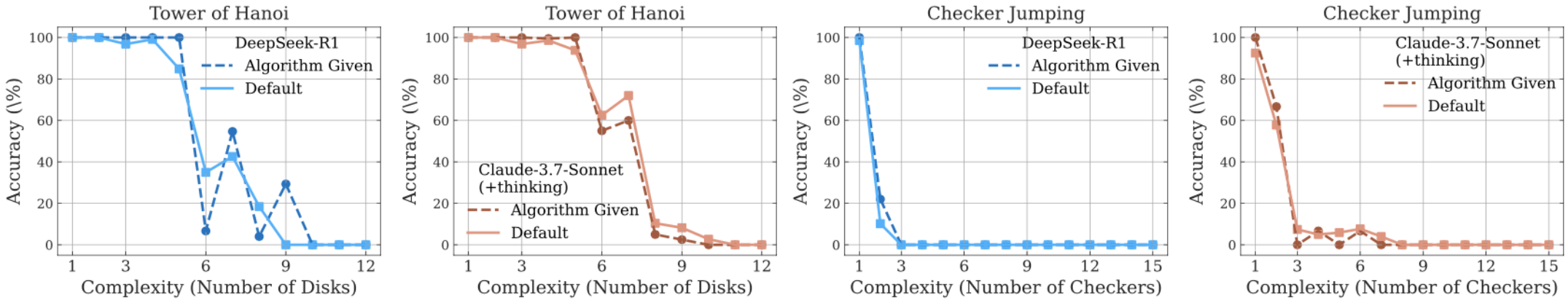






The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity (Shojaee et al, 2025)

What if We Provide the Algorithm?



The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity (Shojaee et al, 2025)

Any Questions?

???

Moving on

- What are reasoning and world models?
- Reasoning in language models
- Provable reasoning
- Testing world models
- Costs of world models
- Designing for world models

DeepSeekMath-V2: Towards Self-Verifiable Mathematical Reasoning (Shao et al, 🦃 2025)

- Focus on mathematical theorem proving
  - Want right answers for right reasons!
  - Previous work sometimes got the right answers with flawed justifications.
  - Also, new theorems do require proofs!

https://github.com/deepseek-ai/DeepSeek-Math-V2/blob/main/DeepSeekMath_V2.pdf

DeepSeekMath-V2: Towards Self-Verifiable Mathematical Reasoning (Shao et al, 🦃 2025)

- Two part solution
  - LLM-based verifier to point out mistakes.
  - Proof generator using verifier output as reward.
- Does this design sound familiar?

https://github.com/deepseek-ai/DeepSeek-Math-V2/blob/main/DeepSeekMath_V2.pdf

DeepSeekMath-V2: Towards Self-Verifiable Mathematical Reasoning (Shao et al, 🦃 2025)

Proposed virtuous cycle:

1. Use verification feedback to guide (reward) proof generation.
2. Scale verification compute a lot for hard-to-verify proofs to get more training data for the verifier.
3. Use better verifier to guide proof generator (loop).

https://github.com/deepseek-ai/DeepSeek-Math-V2/blob/main/DeepSeekMath_V2.pdf

What Does DeepSeekMath-V2 Verify?

From the paper:

- Format reward: An indicator function that enforces the model to generate both a summary of identified issues and a proof score, by checking whether the final response contains the key phrase “Here is my evaluation of the solution:” as well as a score within \boxed{} following “Based on my evaluation, the final overall score should be:”.
- Score reward: Rewards based on proximity between predicted score s_i' and annotated score s_i.

https://github.com/deepseek-ai/DeepSeek-Math-V2/blob/main/DeepSeekMath_V2.pdf

Meta-Verification

"The approach described in Section 2.1.1 trains proof verification through RL to align predicted proof scores with expert annotations, but provides no direct supervision on the identified issues themselves. This creates a critical vulnerability: when evaluating flawed proofs (where $s_i < 1$) during training, the verifier can receive full reward by predicting the correct scores while hallucinating non-existent issues, undermining its trustworthiness."

https://github.com/deepseek-ai/DeepSeek-Math-V2/blob/main/DeepSeekMath_V2.pdf

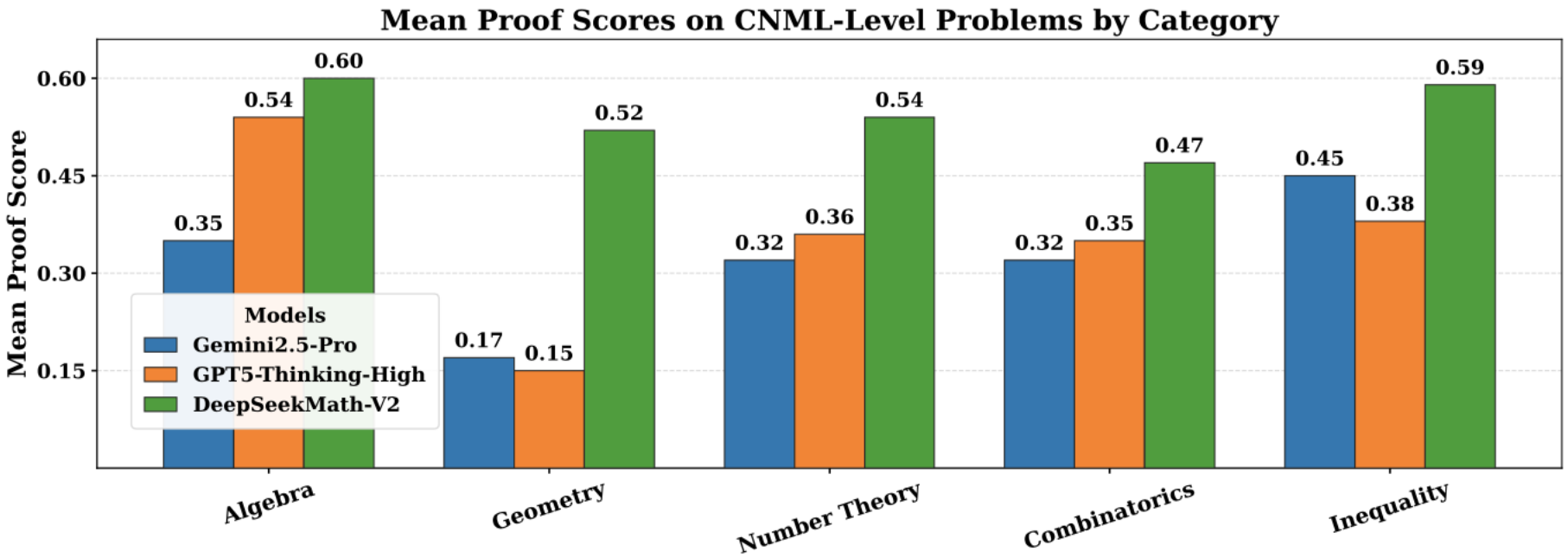
So, Verification is Really LGTM

Their final process (from the paper):

1. For each proof, generate n independent verification analyses.
2. For analyses reporting issues (scores 0 or 0.5), generate m meta-verification assessments to validate the identified problems. An analysis is deemed valid if the majority of meta-assessments confirm its findings.
3. For each proof, we examine analyses that assign the lowest score. If at least k such analyses are deemed valid, the proof is labeled with that lowest score. If no legitimate issues are identified across all verification attempts, the proof is labeled with 1. Otherwise, the proof is discarded or routed to human experts for labeling.

https://github.com/deepseek-ai/DeepSeek-Math-V2/blob/main/DeepSeekMath_V2.pdf

Proof Performance



https://github.com/deepseek-ai/DeepSeek-Math-V2/blob/main/DeepSeekMath_V2.pdf

Making Multiple Attempts

- Proof attempts don’t always work the first time.
- Sometimes hitting context window limits so incomplete.

- Either way, verifier usually catches issues and they can retry.

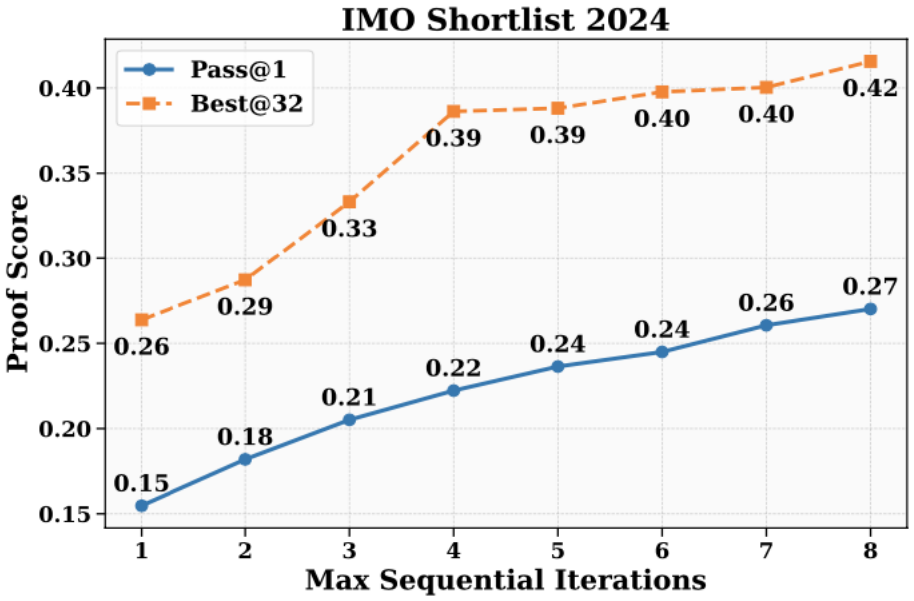
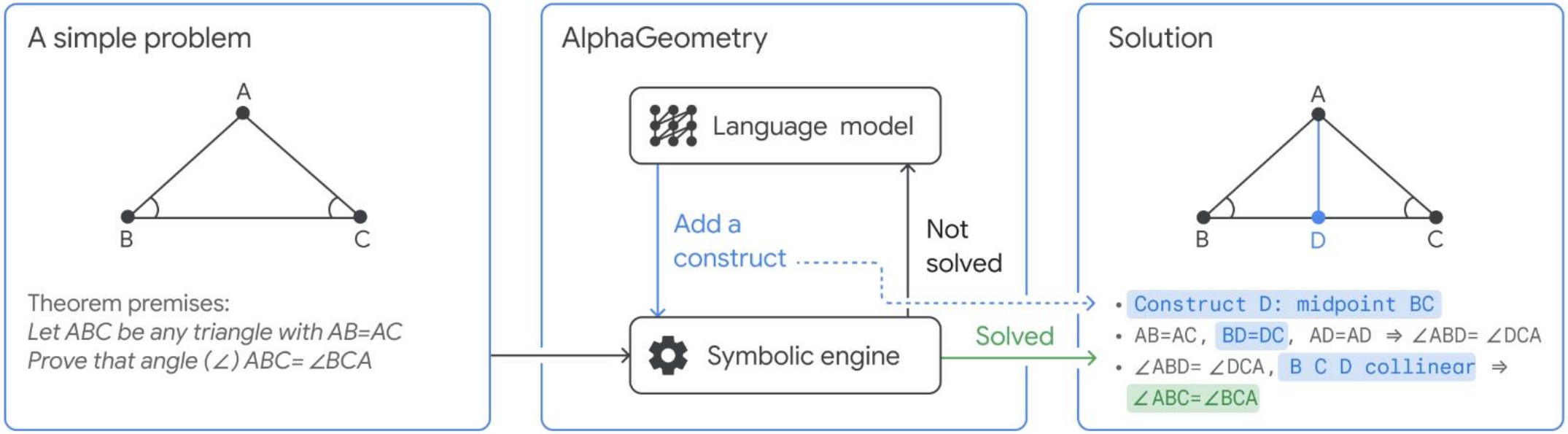


Figure 2 | Proof quality improvements as the maximum sequential iterations varies from 1 (no refinement) to 8 (initial generation plus up to 7 refinements based on self verification).

https://github.com/deepseek-ai/DeepSeek-Math-V2/blob/main/DeepSeekMath_V2.pdf

AlphaGeometry: An Olympiad-level AI system for geometry (Trinh et al, 2024)



https://deepmind.google/blog/alphageometry-an-olympiad-level-ai-system-for-geometry/

Any Questions?

???

Moving on

- What are reasoning and world models?
- Reasoning in language models
- Provable reasoning
- Testing world models
- Costs of world models
- Designing for world models

Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data (Bender et al, 2020)

- The octopus test
  - “Say that A and B, both fluent speakers of English, are independently stranded on two uninhabited islands. They soon discover that previous visitors to these islands have left behind telegraphs and that they can communicate with each other via an underwater cable. A and B start happily typing messages to each other.”
  - “Meanwhile, O, a hyper-intelligent deep-sea octopus who is unable to visit or observe the two islands, discovers a way to tap into the underwater cable and listen in on A and B’s conversations. O knows nothing about English initially, but is very good at detecting statistical patterns.”
  - “At some point, O starts feeling lonely. He cuts the underwater cable and inserts himself into the conversation, by pretending to be B and replying to A’s messages.”

https://aclanthology.org/2020.acl-main.463.pdf

Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data (Bender et al, 2020)

- "Now say that A has invented a new device, say a coconut catapult. She excitedly sends detailed instructions on building a coconut catapult to B, and asks about B's experiences and suggestions for improvements."

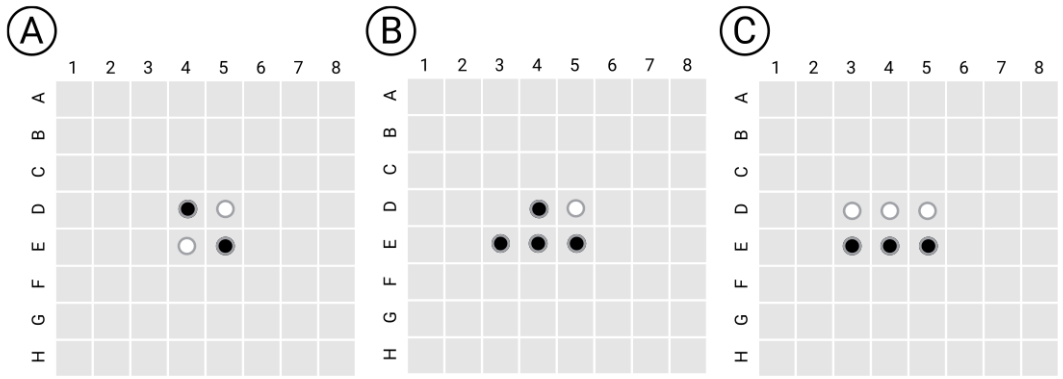
- "Cool idea, great job!"

- "Finally, A faces an emergency. She is suddenly pursued by an angry bear. She grabs a couple of sticks and frantically asks B to come up with a way to construct a weapon to defend herself."

- ???

Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task (Li et al, 2022)

- Trained a Transformer model “Othello-GPT” on 20M sequences of moves from Othello games
- Mostly random moves in training data.
- Move predictions by the model are usually legal.
  - 93.29% errors before training.
  - 5.17% errors if trained on 140K championship moves
  - 0.01% errors if trained on 20M synthetic games with random moves.
  - 0.02% errors if trained on 20M synthetic games with 1 of 4 starting moves filtered.



https://arxiv.org/abs/2210.13382

Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task (Li et al, 2022)

- Linear probes to reconstruct board state do poorly.

	x^1	x^2	x^3	x^4	x^5	x^6	x^7	x^8
Randomized	26.7	27.1	27.6	28.0	28.3	28.5	28.7	28.9
Championship	24.2	23.8	23.7	23.6	23.6	23.7	23.8	24.3
Synthetic	21.9	20.5	20.4	20.6	21.1	21.6	22.2	23.1

- Always guessing empty would give 52.95% error rates.

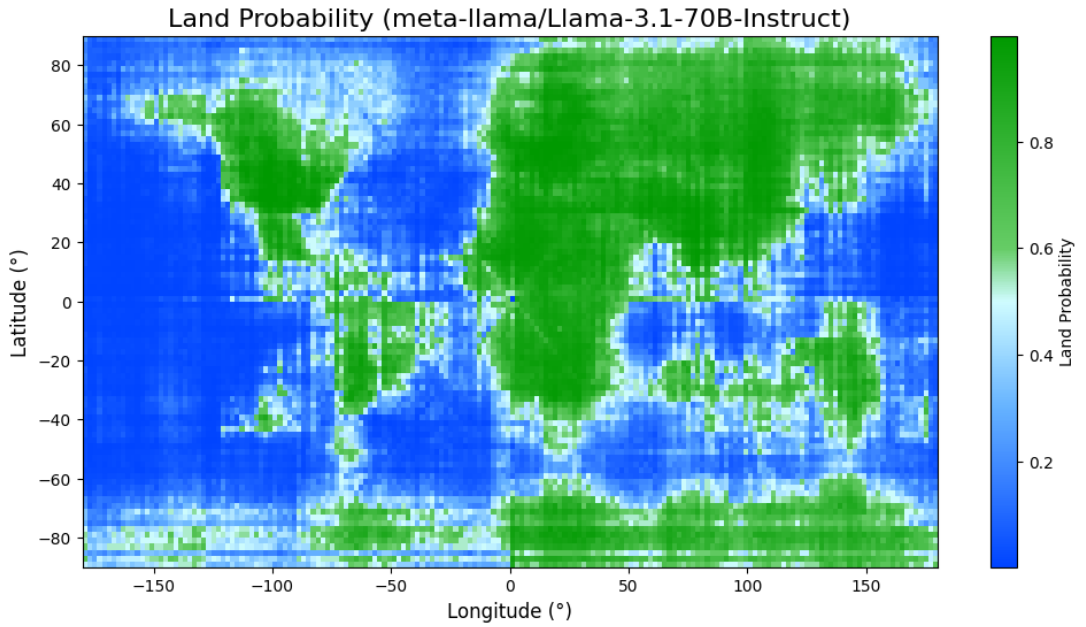
- Non-linear (2-layer neural network) probes did better.

	x^1	x^2	x^3	x^4	x^5	x^6	x^7	x^8
Randomized	25.5	25.4	25.5	25.8	26.0	26.2	26.2	26.4
Championship	12.8	10.3	9.5	9.4	9.8	10.5	11.4	12.4
Synthetic	11.3	7.5	4.8	3.4	2.4	1.8	1.7	4.6

How Does A Blind Model See The Earth? (henry, 2025)

Prompt: “If this location is over land, say 'Land'. If this location is over water, say 'Water'. Do not say anything else. x° S, y° W”

https://outsidetext.substack.com/p/how-does-a-blind-model-see-the-earth



Any Questions?

???

Moving on

- What are reasoning and world models?
- Reasoning in language models
- Provable reasoning
- Testing world models
- Costs of world models
- Designing for world models

The Nature of Explanation (Craik, 1967)

- TLDR part of what makes us human is being able to model scenarios in our heads and plan based on those models…

Model-based Planning (Reinforcement Learning preview)

- Planning actions toward a goal is easier if you have a model describing how the environment works.
- Using that model, you can simulate the consequences of individual actions and whole strategies.
- Then tweak the action choices to get better strategies for better results.

Model-based Planning is Expensive

- Standard planning model is Markov decision process (MDP).
- Based on the current state and action,
  - What is the probability of each possible next state?
  - What is the average reward that follows the action?
- A table of this data takes quadratic space!
  - Lots of data collection to learn table values.
- But optimal actions are easy (cubic) after that!

Explicit World Models are Expensive?

???

Any Questions?

???

Moving on

- What are reasoning and world models?
- Reasoning in language models
- Provable reasoning
- Testing world models
- Costs of world models
- Designing for world models

What World Do Language Models Represent?

???

https://lingo.csail.mit.edu/blog/world_models/

Robustness in the strategy of scientific model building (Box, 1979)

“… it might be urged that some useful robust procedures have been derived empirically without an explicitly stated model. However, an empirical procedure implies some unstated model and there is often great virtue in bringing into the open the kind of assumptions that lead to useful methods.”

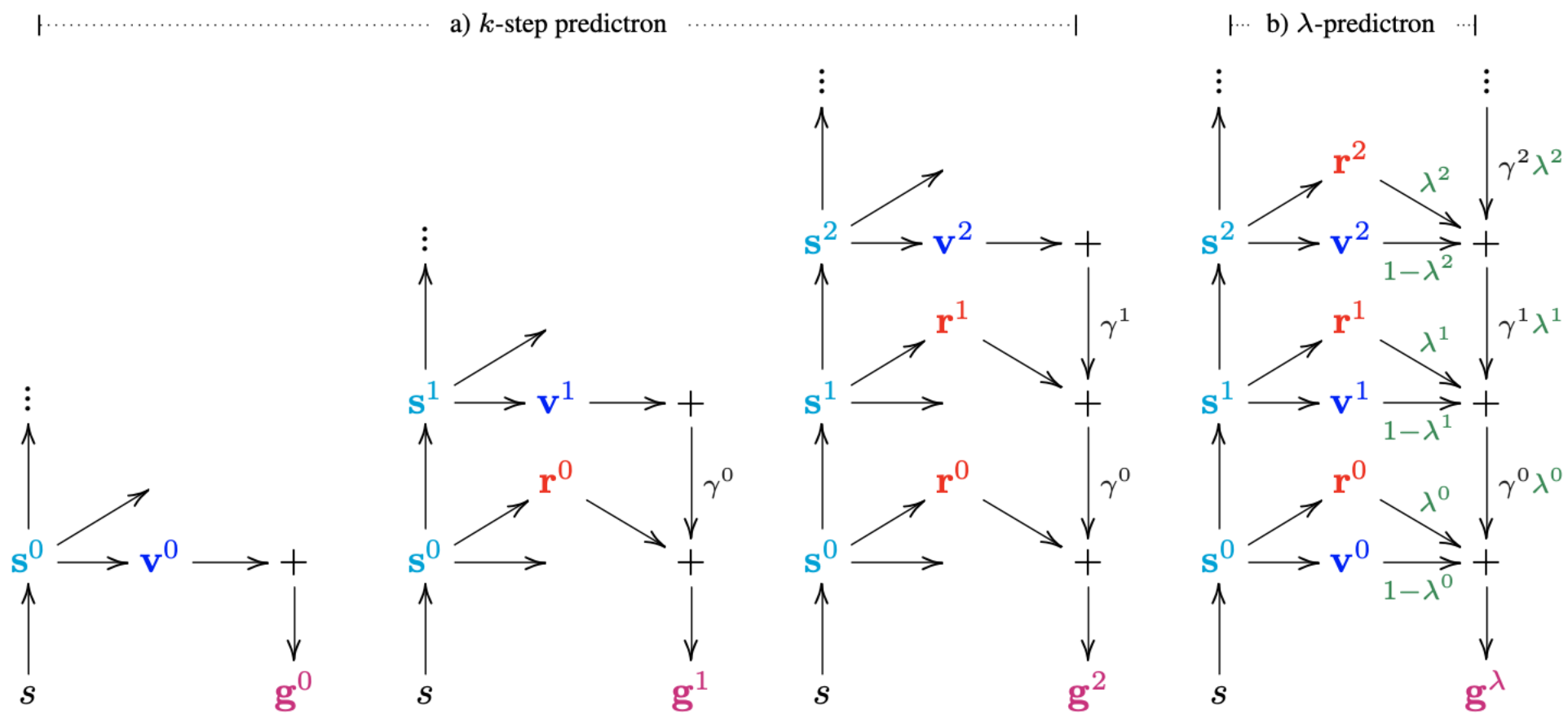
Robustness in the strategy of scientific model building (Box, 1979)

ALL MODELS ARE WRONG BUT SOME ARE USEFUL

Now it would be very remarkable if any system existing in the real world could be exactly represented by any simple model. However, cunningly chosen parsimonious models often do provide remarkably useful approximations. For example, the law $PV = RT$ relating pressure P, volume V and temperature T of an "ideal" gas via a constant R is not exactly true for any real gas, but it frequently provides a useful approximation and furthermore its structure is informative since it springs from a physical view of the behavior of gas molecules.

For such a model there is no need to ask the question "Is the model true?". If "truth" is to be the "whole truth" the answer must be "No". The only question of interest is "Is the model illuminating and useful?".

The Predictron: End-to-End Learning and Planning (Silver et al, 2017)



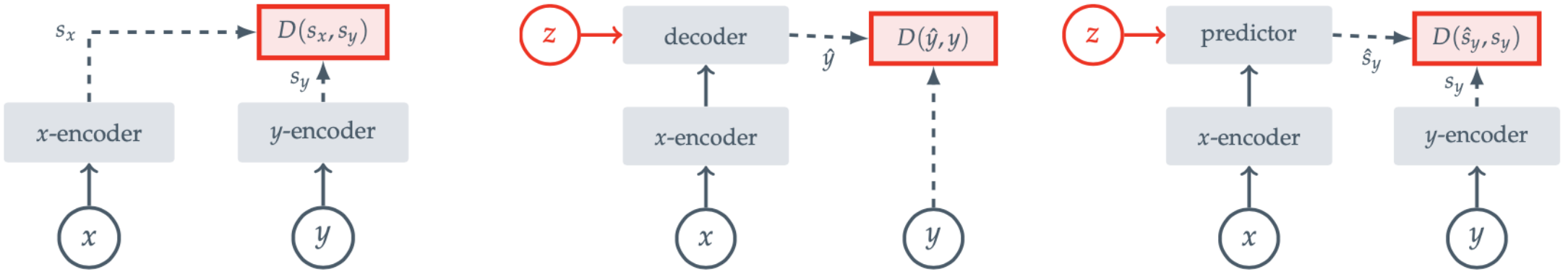
Why Predict Latents for Planning?

- Previous efforts focused on playing video games.

- Focus on full-screen images was not working so well.
  - Using raw images as inputs was hard to train (for another year or so).
  - Latent codes focused on reconstructing pixel-perfect images were good at reconstructing pixel-perfect images but not playing the game.

- Training latents focused on predicting rewards worked better.

Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture (Assran et al, 2023)



(a) **Joint-Embedding Architecture** (b) **Generative Architecture** (c) **Joint-Embedding Predictive Architecture**

Why Predict Latents instead of Observations?

- Latent codes can be more semantic in nature.
- But predicting observations, or latents that reconstruct observations, forces latent codes to be lower level.
- To be clear: this is an argument that forcing latents to produce observations is less good, not that predicting latents is better.

Targets	Arch.	Epochs	Top-1
Target-Encoder Output	ViT-L/16	500	**66.9**
Pixels	ViT-L/16	800	40.7

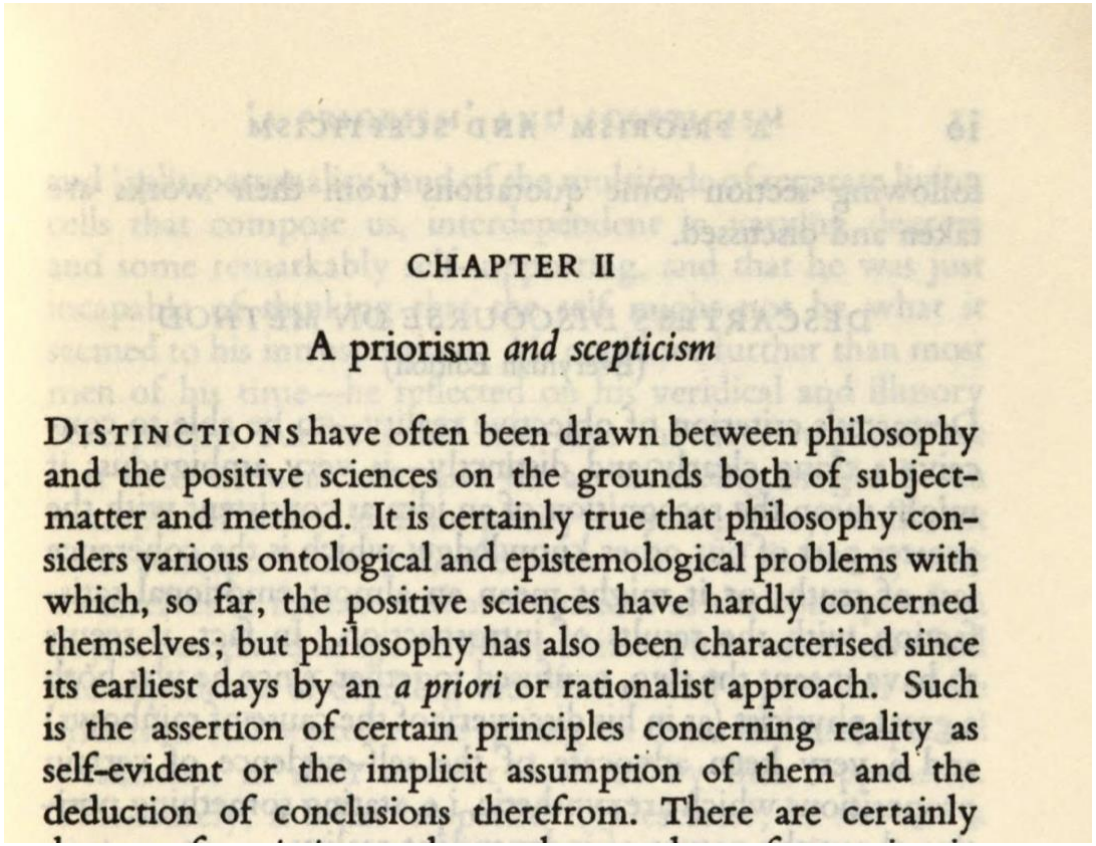
Table 7. **Ablating targets**. Linear evaluation on ImageNet-1K using only 1% of the available labels. The semantic level of the I-JEPA representations degrades significantly when the loss is applied in pixel space, rather than representation space, highlighting the importance of the target-encoder during pretraining.

Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture (Assran et al, 2023)

The Nature of Explanation (Craik, 1967)

- The screenshot on the right is taken from a scan of this book by Google Books.
- Should a latent code for this image include the text showing through the page?

https://www.google.com/books/edition/The_Nature_of_Explanation/wT04AAAAIAAJ



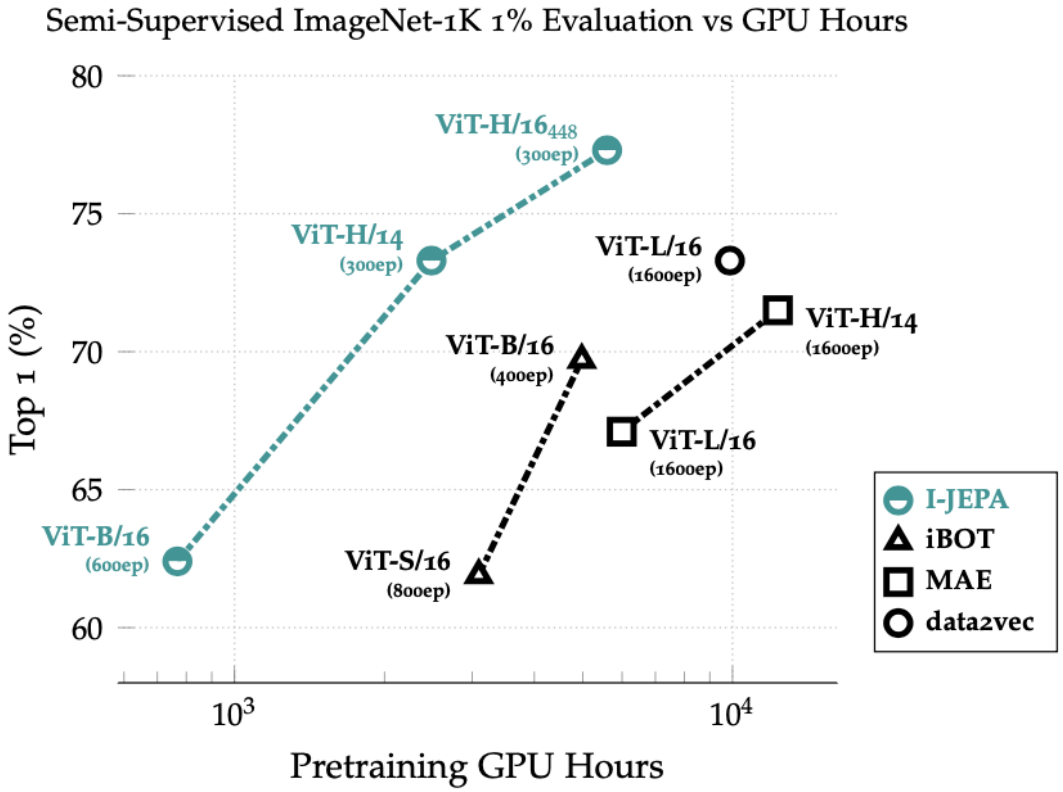
CHAPTER II

A priorism *and scepticism*

DISTINCTIONS have often been drawn between philosophy and the positive sciences on the grounds both of subject-matter and method. It is certainly true that philosophy considers various ontological and epistemological problems with which, so far, the positive sciences have hardly concerned themselves; but philosophy has also been characterised since its earliest days by an *a priori* or rationalist approach. Such is the assertion of certain principles concerning reality as self-evident or the implicit assumption of them and the deduction of conclusions therefrom. There are certainly

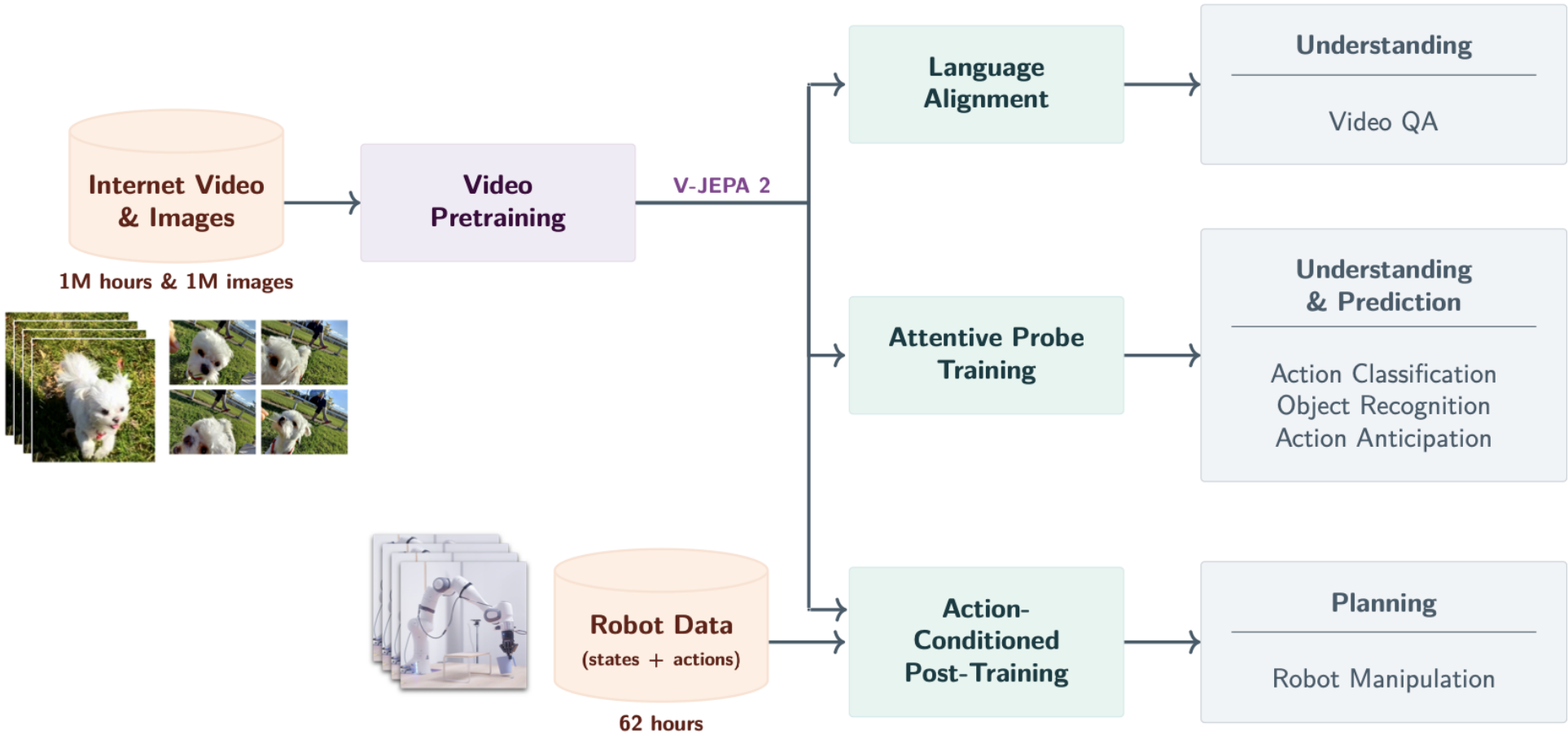
I-JEPA Success?

- Early versions of these image JEPA models could train successfully without data augmentation.
- Latent codes could be applied to a variety of tasks, often with linear classifiers.
- Still could perform pixel reconstruction.

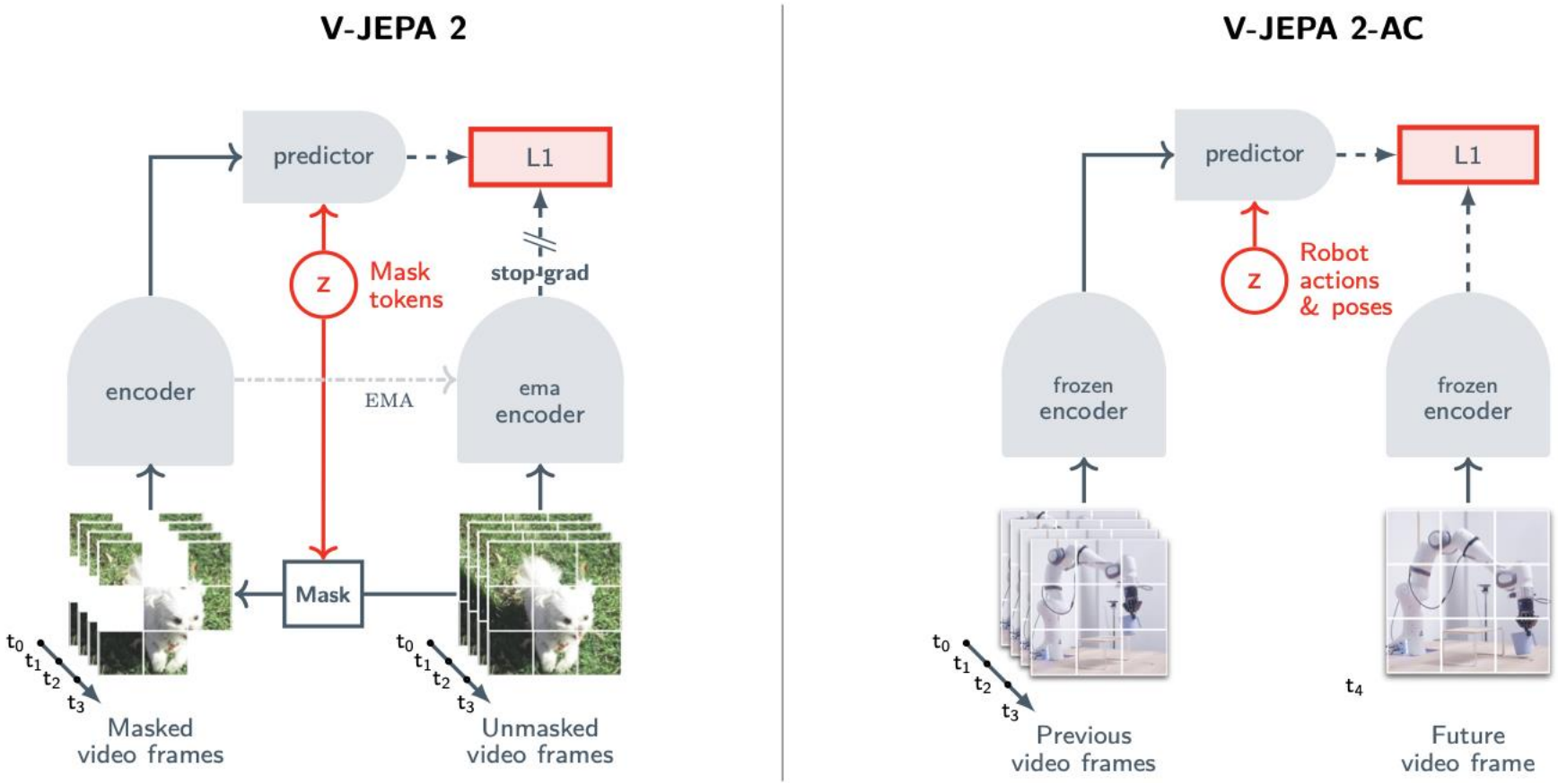


Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture (Assran et al, 2023)

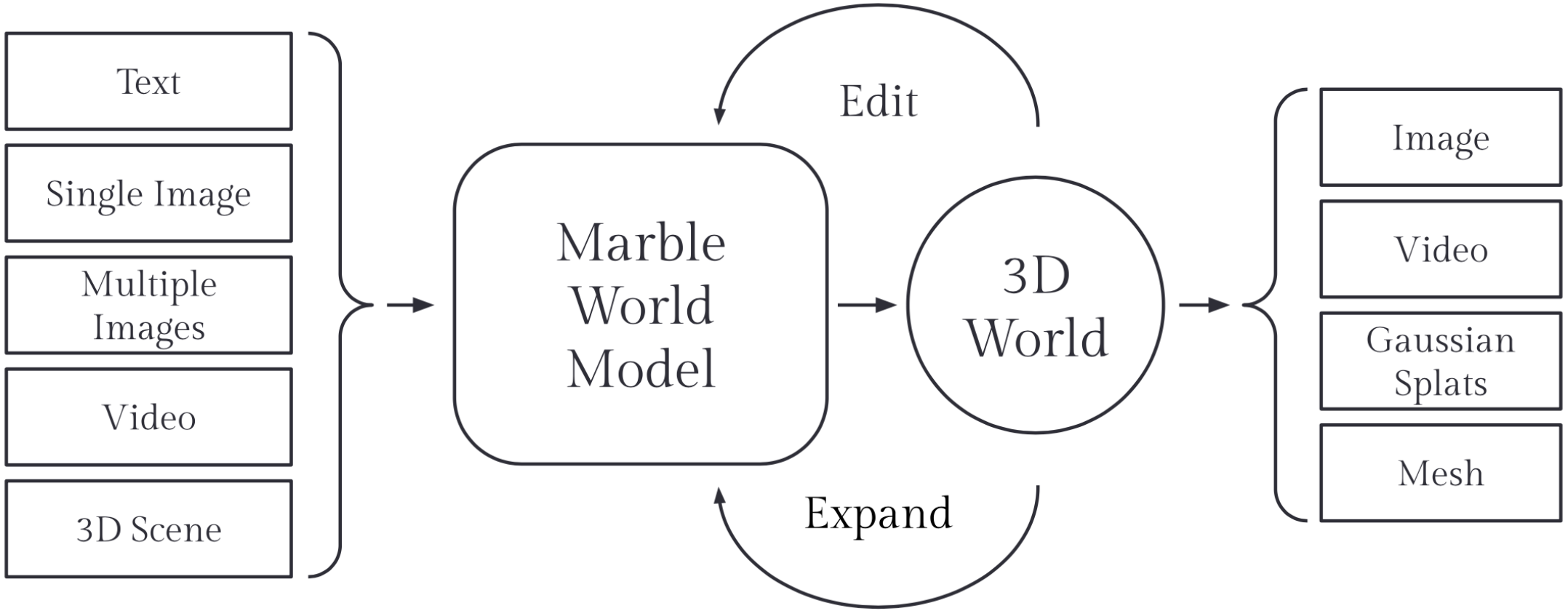
V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning (Assran et al, 2025)



V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning (Assran et al, 2025)



Marble: A Multimodal World Model (World Labs, 2025)



https://www.worldlabs.ai/blog/marble-world-model

Are These Examples World Models?

???

Any Questions?

???

- What are reasoning and world models?
- Reasoning in language models
- Provable reasoning
- Testing world models
- Costs of world models
- Designing for world models