amii ALBERTA MACHINE INTELLIGENCE INSTITUTE

Openmind RESEARCH INSTITUTE

xAI

UNIVERSITY OF ALBERTA

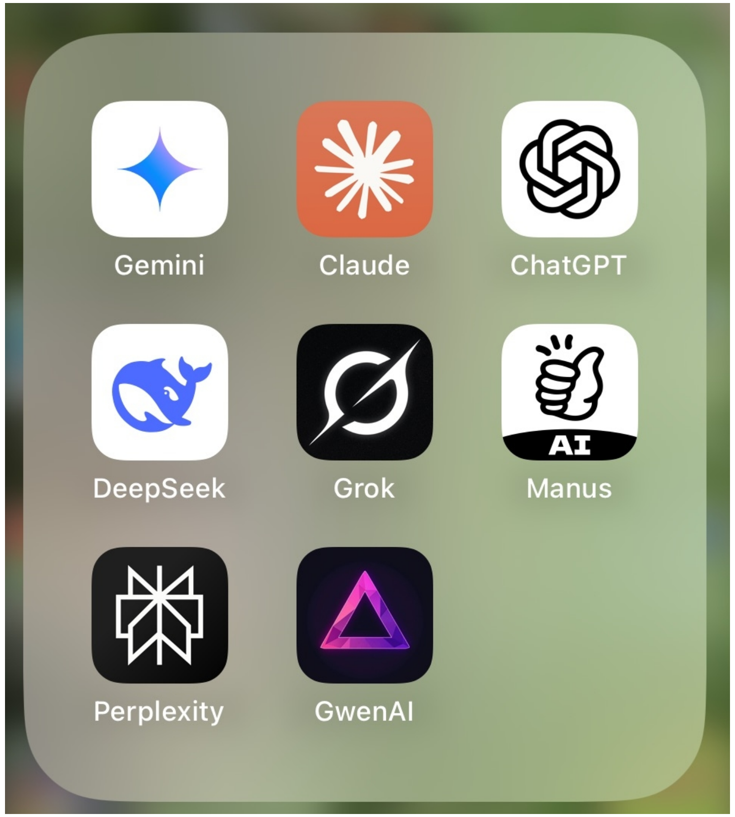
A (partial) framework for debugging foundation models.

Randy Goebel
AMII Fellow & Co-founder
Professor of Computing Science
Adjunct Professor of Medicine
rgoebel@ualberta.ca

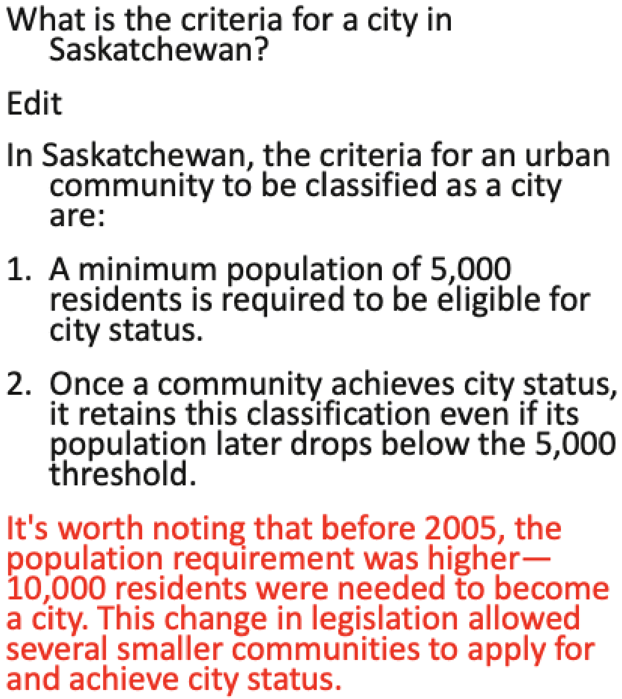
Outline

- Motivation: AI systems as (buggy) computer programs
- The spectrum of neurosymbolic foundation models
- Should foundation models be debuggable?
- A framework for stackable multi-level foundation models
- Summary

Let’s identify an error in an LLM



Claude Sonnet 3.7



What is the criteria for a city in Saskatchewan?

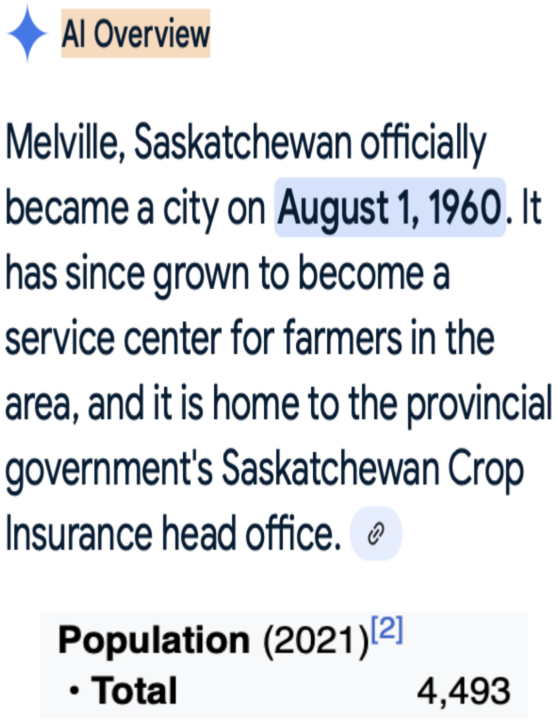
Edit

In Saskatchewan, the criteria for an urban community to be classified as a city are:

1. A minimum population of 5,000 residents is required to be eligible for city status.
2. Once a community achieves city status, it retains this classification even if its population later drops below the 5,000 threshold.

It's worth noting that before 2005, the population requirement was higher—10,000 residents were needed to become a city. This change in legislation allowed several smaller communities to apply for and achieve city status.

Gemini May 5/2025

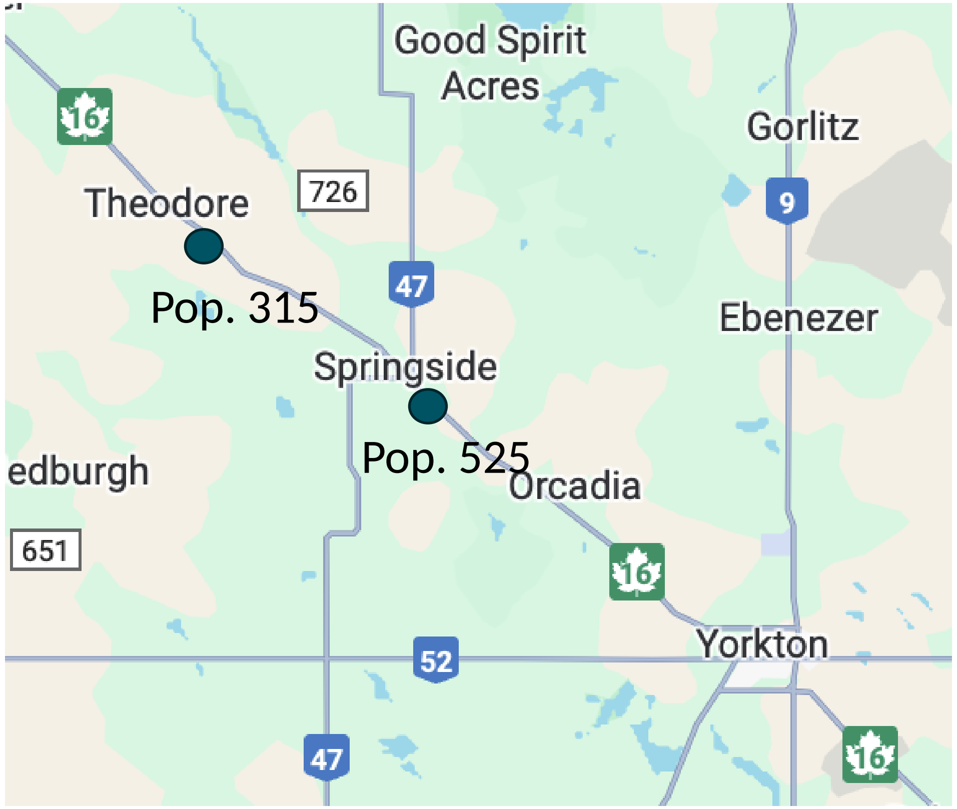


AI Overview

Melville, Saskatchewan officially became a city on August 1, 1960. It has since grown to become a service center for farmers in the area, and it is home to the provincial government's Saskatchewan Crop Insurance head office.

Population (2021)[2]	
• **Total**	4,493

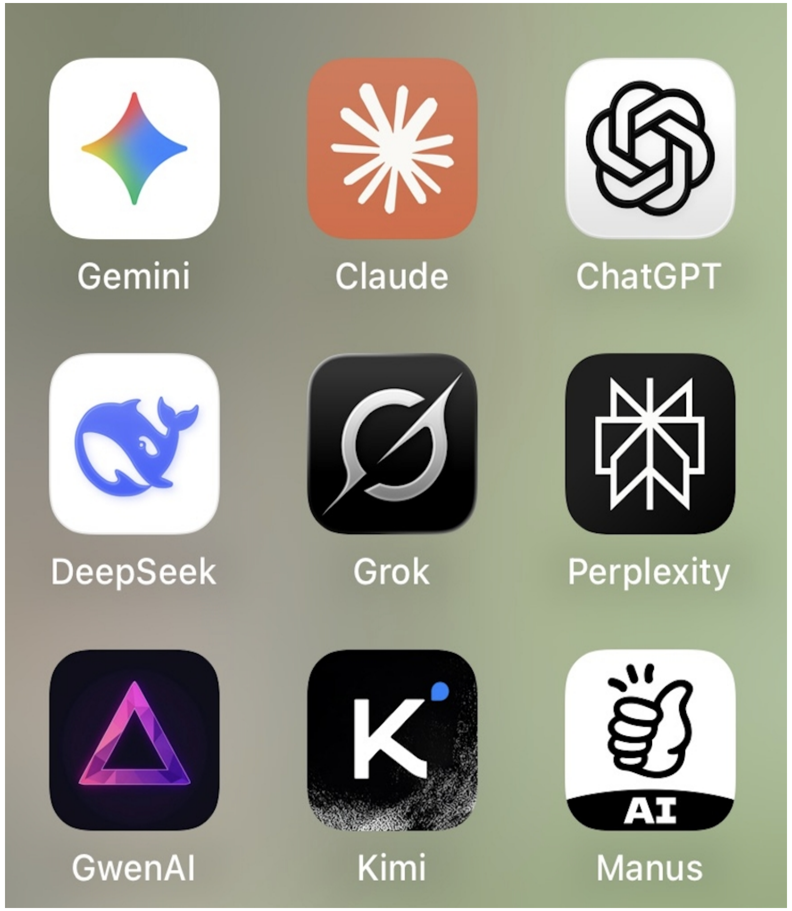
The Springside Theodore Combines

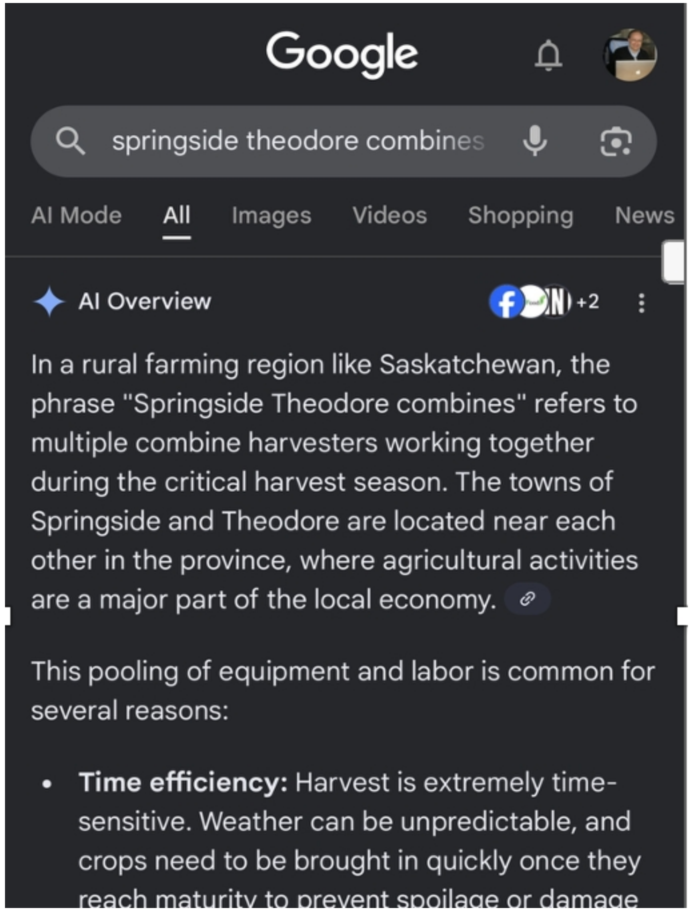


- the province of Saskatchewan had an “A” (cf. “AA,” “AAA,” “MLB”) semi-professional baseball league
- Springside and Theodore were two small farming towns that didn’t have enough players unless they combined
- So they formed a baseball team called the “Springside Theodore Combines.”

…but a combine can be many things

Let’s identify an LLM error

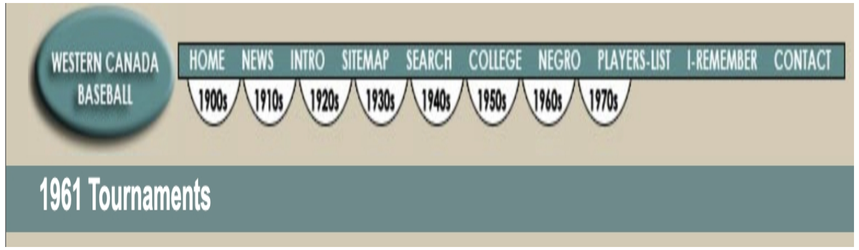




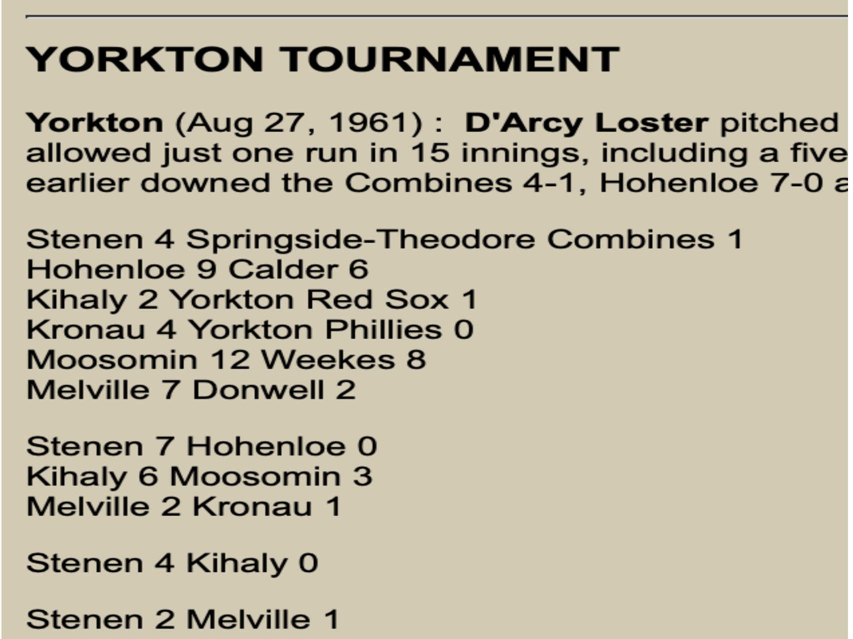
In a rural farming region like Saskatchewan, the phrase "Springside Theodore combines" refers to multiple combine harvesters working together during the critical harvest season. The towns of Springside and Theodore are located near each other in the province, where agricultural activities are a major part of the local economy.

This pooling of equipment and labor is common for several reasons:

- **Time efficiency:** Harvest is extremely time-sensitive. Weather can be unpredictable, and crops need to be brought in quickly once they reach maturity to prevent spoilage or damage



1961 Tournaments



YORKTON TOURNAMENT

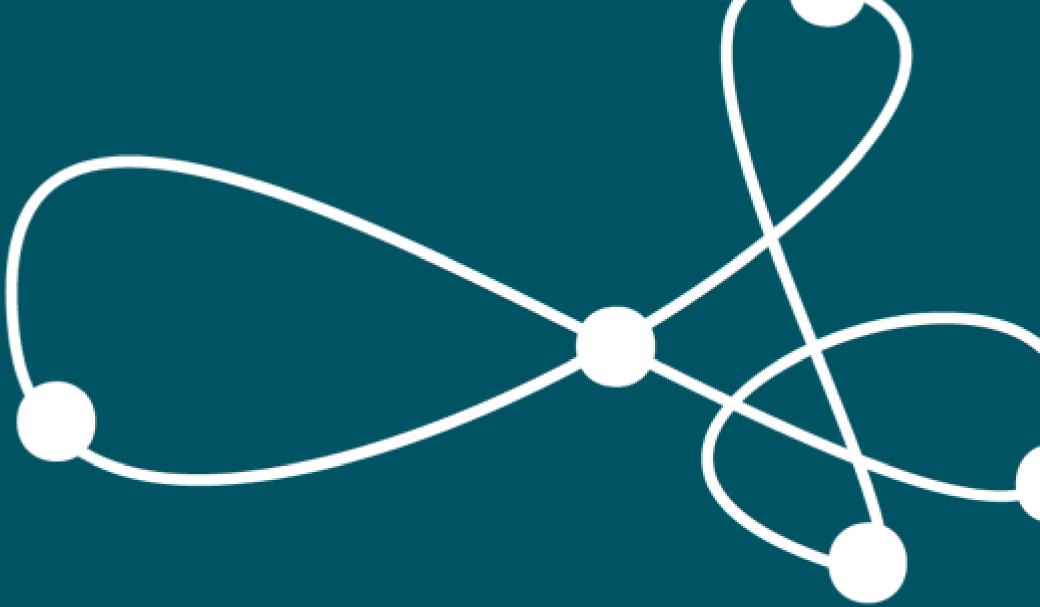
Yorkton (Aug 27, 1961) : **D'Arcy Loster** pitched allowed just one run in 15 innings, including a five earlier downed the Combines 4-1, Hohenloe 7-0 a

Stenen 4 Springside-Theodore Combines 1
Hohenloe 9 Calder 6
Kihaly 2 Yorkton Red Sox 1
Kronau 4 Yorkton Phillies 0
Moosomin 12 Weekes 8
Melville 7 Donwell 2

Stenen 7 Hohenloe 0
Kihaly 6 Moosomin 3
Melville 2 Kronau 1

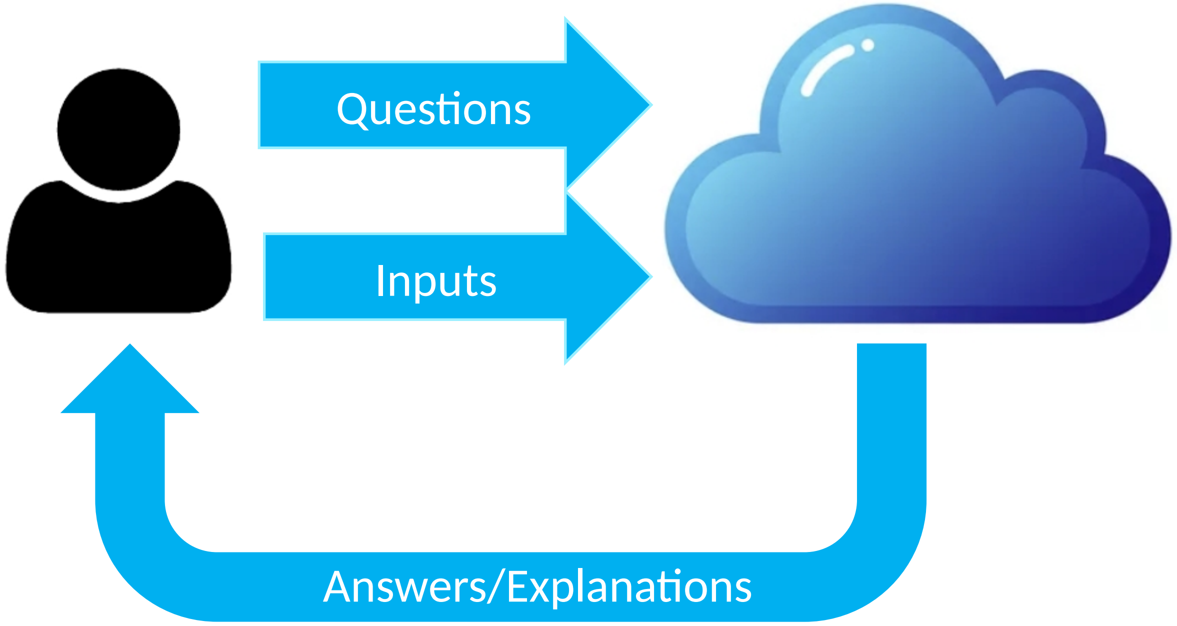
Stenen 4 Kihaly 0

Stenen 2 Melville 1



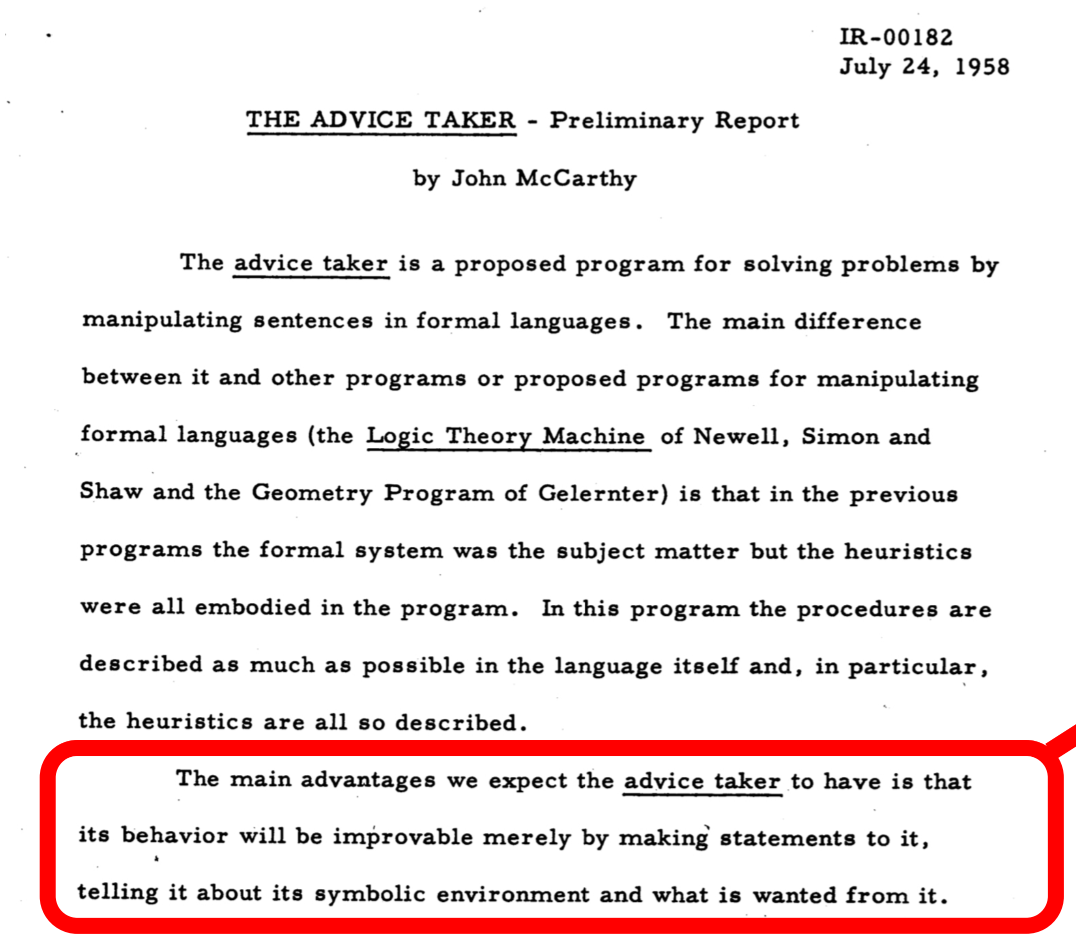
The Query/Update/Explain paradigm (QUE)

The Query/Update/Explain paradigm



- There are only three ways to interact with a computer program:
  - Ask a question (what, where, how, when, …)
  - Provide an input (fact, opinion, observation, …)
  - Ask for an explanation of an answer (justify an answer to a question)

McCarthy's Advice Taker



IR-00182
July 24, 1958

THE ADVICE TAKER - Preliminary Report

by John McCarthy

The advice taker is a proposed program for solving problems by manipulating sentences in formal languages. The main difference between it and other programs or proposed programs for manipulating formal languages (the Logic Theory Machine of Newell, Simon and Shaw and the Geometry Program of Gelernter) is that in the previous programs the formal system was the subject matter but the heuristics were all embodied in the program. In this program the procedures are described as much as possible in the language itself and, in particular, the heuristics are all so described.

The main advantages we expect the advice taker to have is that its behavior will be improvable merely by making statements to it, telling it about its symbolic environment and what is wanted from it.

The main advantage we expect the advice taker to have is that its behavior will be improvable by merely making statements to it...

The spectrum of neurosymbolic foundation models

What are foundation models?

- “Foundation models are large AI models trained on massive, diverse datasets, allowing them to be adapted to a wide range of tasks. They are essentially pre-trained, general-purpose models that can be fine-tuned for specific applications, serving as a foundation for building various AI applications.”

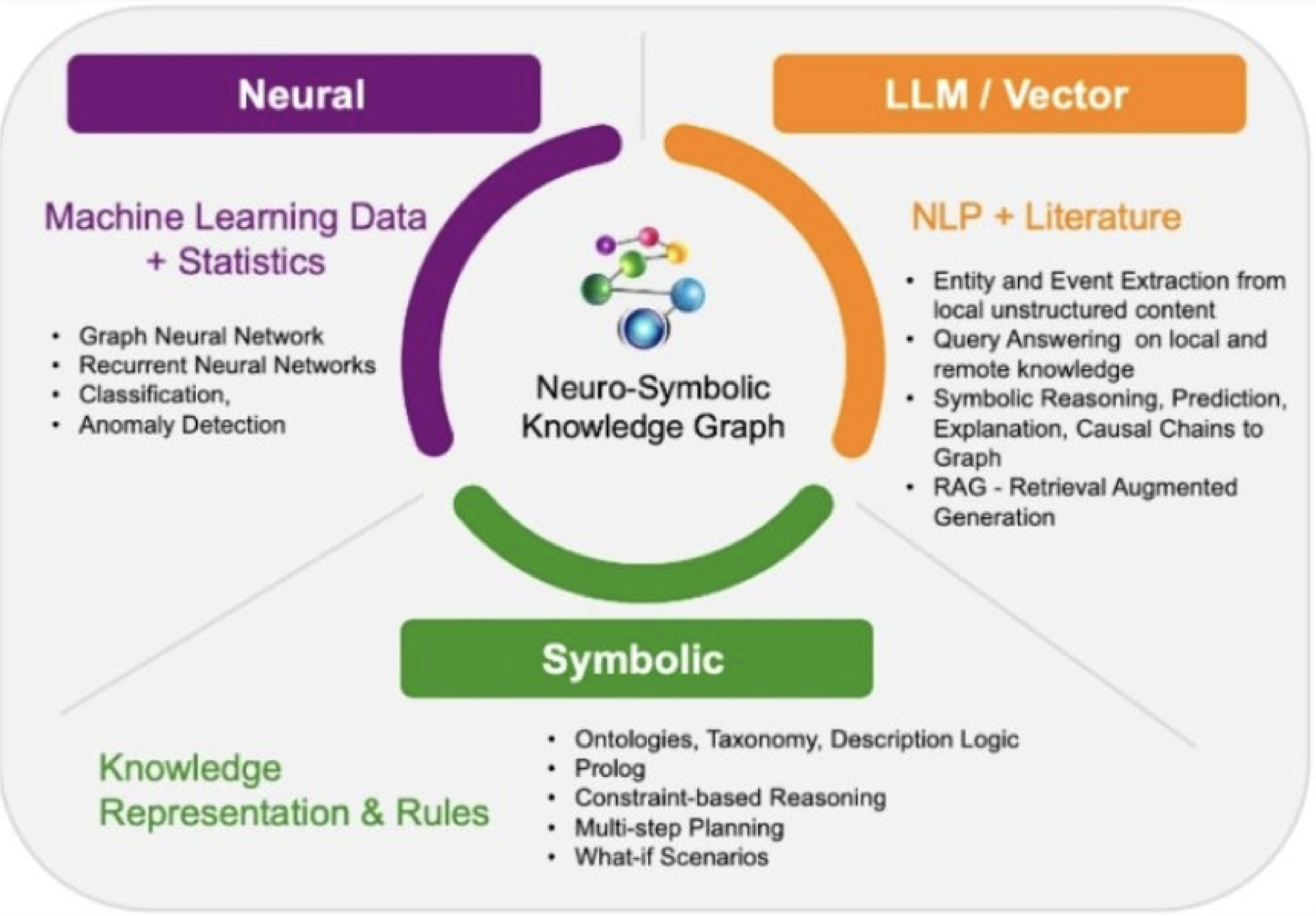
  - Google Gemini

- “A foundation model, also known as large X model (LxM), is a machine learning or deep learning model that is trained on vast datasets so it can be applied across a wide range of use cases.[1] Generative AI applications like Large Language Models are common examples of foundation models.”

  - Wikipedia

- “AI is undergoing a paradigm shift with the rise of models (e.g., BERT, DALL-E, GPT-3) that are trained on broad data at scale and are adaptable to a wide range of downstream tasks. We call these models foundation models to underscore their critically central yet incomplete character.”

  - multi-author (≈100) paper from Center for Research on Foundation Models, Stanford Institute for Human-Centered AI https://arxiv.org/abs/2108.07258



Neural

LLM / Vector

Machine Learning Data + Statistics

- Graph Neural Network
- Recurrent Neural Networks
- Classification,
- Anomaly Detection

Neuro-Symbolic Knowledge Graph

NLP + Literature

- Entity and Event Extraction from local unstructured content
- Query Answering on local and remote knowledge
- Symbolic Reasoning, Prediction, Explanation, Causal Chains to Graph
- RAG - Retrieval Augmented Generation

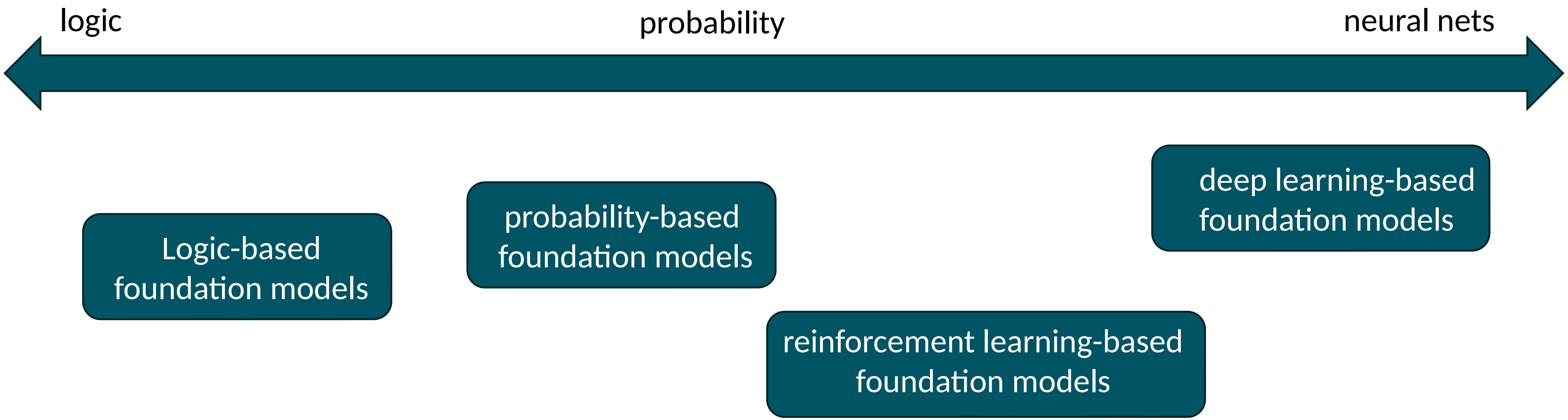
Symbolic

Knowledge Representation & Rules

- Ontologies, Taxonomy, Description Logic
- Prolog
- Constraint-based Reasoning
- Multi-step Planning
- What-if Scenarios

https://www.linkedin.com/posts/martin-ciupa-76418b17_heres-a-personal-breakdown-of-the-history-activity-7189730184667783169-dphx/

A plausible spectrum of neurosymbolic foundational models



logic

probability

neural nets

Logic-based foundation models

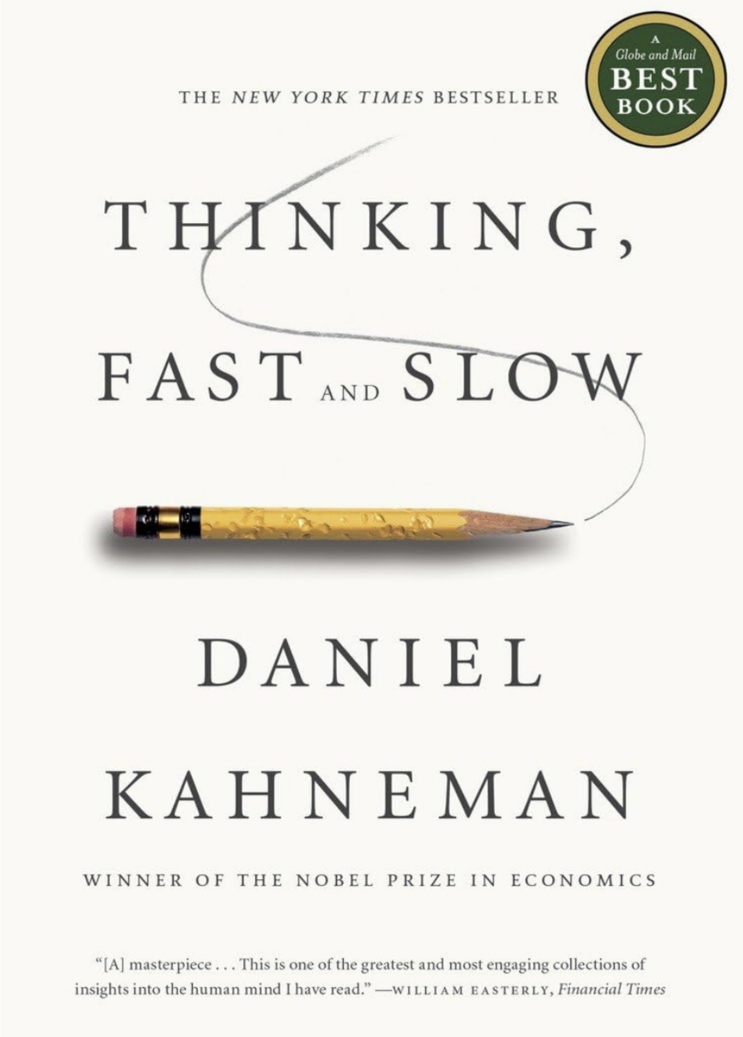
probability-based foundation models

reinforcement learning-based foundation models

deep learning-based foundation models

Dan Kahneman, Economics Nobel Laureate 2002

- More recently, proponent of “Thinking Fast, Thinking Slow.”



THE *NEW YORK TIMES* BESTSELLER

A *Globe and Mail* BEST BOOK

THINKING, FAST AND SLOW

DANIEL KAHNEMAN

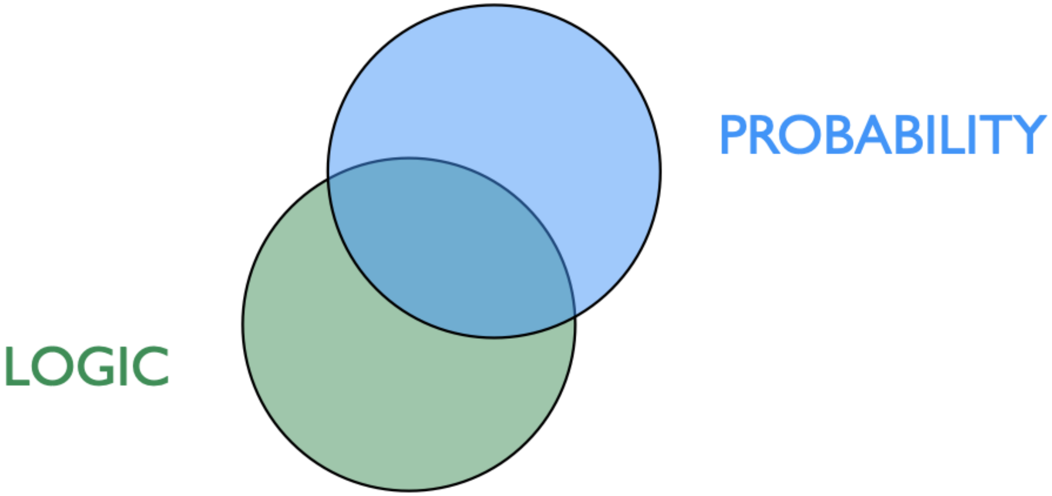
WINNER OF THE NOBEL PRIZE IN ECONOMICS

“[A] masterpiece . . . This is one of the greatest and most engaging collections of insights into the human mind I have read.” —WILLIAM EASTERLY, *Financial Times*

Kahneman’s (2013) fast and slow thinking

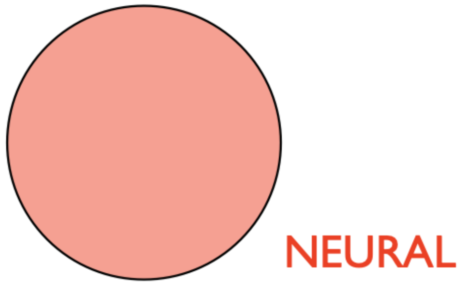
Thinking slow = reasoning

TWO MAIN PARADIGMS in AI



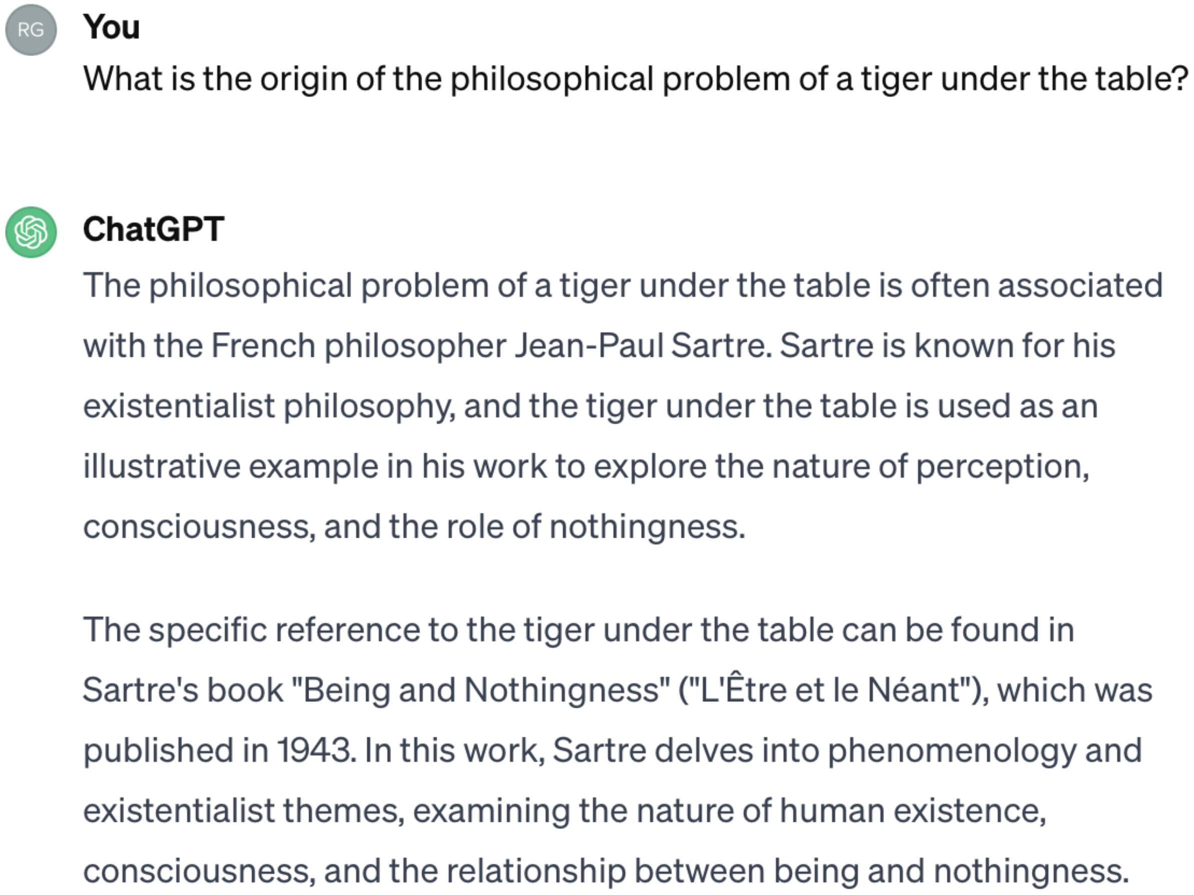
Thinking fast

MAIN PARADIGM in AI
Focus on Learning



De Raedt et al., **From Statistical Relational to Neuro-Symbolic Artificial Intelligence**
https://arxiv.org/abs/2003.08316

Is that a tiger under my table?



You

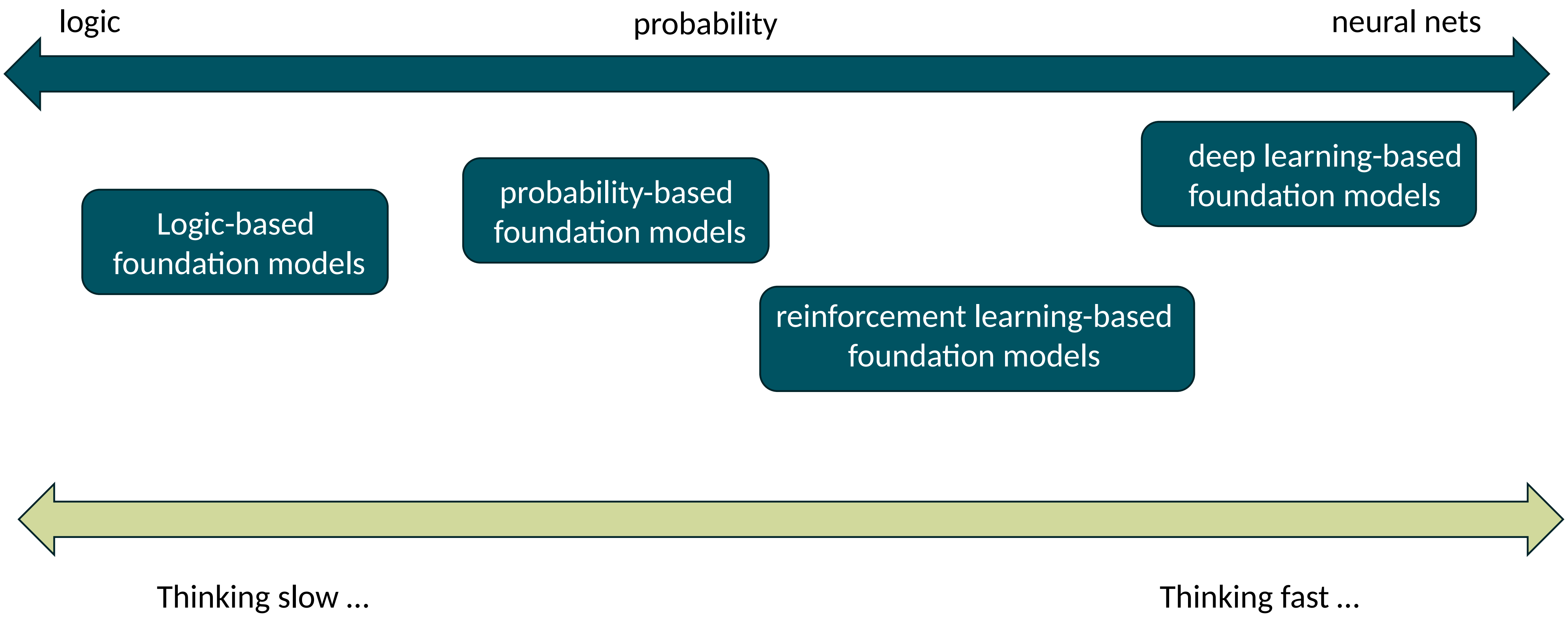
What is the origin of the philosophical problem of a tiger under the table?

ChatGPT

The philosophical problem of a tiger under the table is often associated with the French philosopher Jean-Paul Sartre. Sartre is known for his existentialist philosophy, and the tiger under the table is used as an illustrative example in his work to explore the nature of perception, consciousness, and the role of nothingness.

The specific reference to the tiger under the table can be found in Sartre's book "Being and Nothingness" ("L'Être et le Néant"), which was published in 1943. In this work, Sartre delves into phenomenology and existentialist themes, examining the nature of human existence, consciousness, and the relationship between being and nothingness.

A plausible neurosymbolic framework



logic

probability

neural nets

Logic-based foundation models

probability-based foundation models

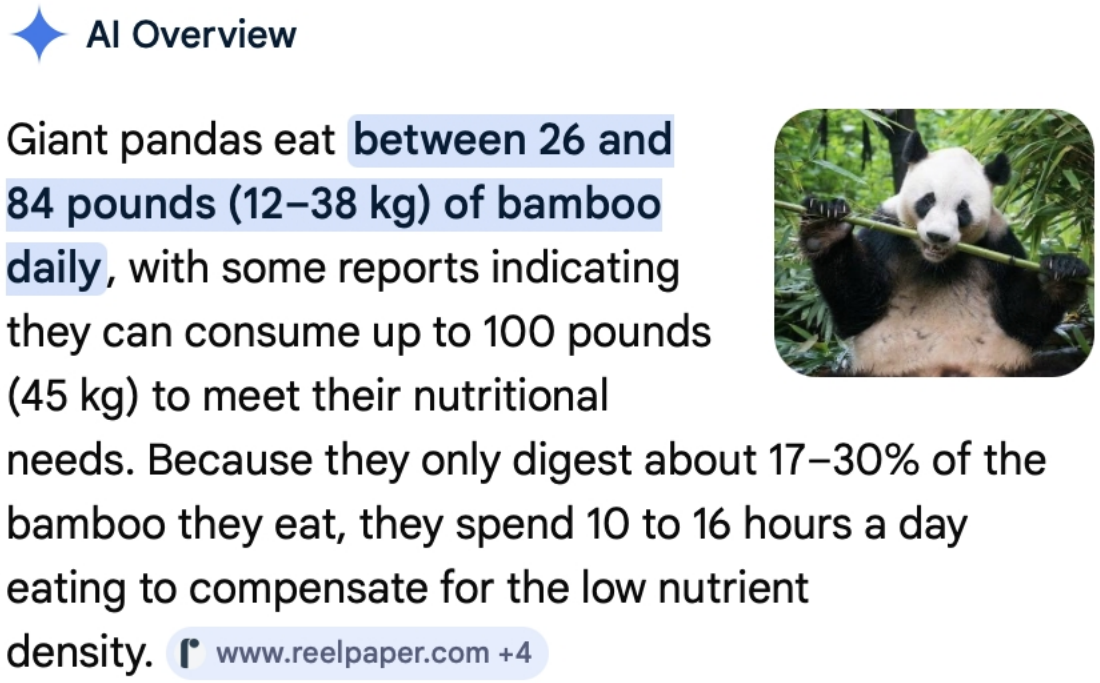
deep learning-based foundation models

reinforcement learning-based foundation models

Thinking slow …

Thinking fast …

High input volume, low value capture

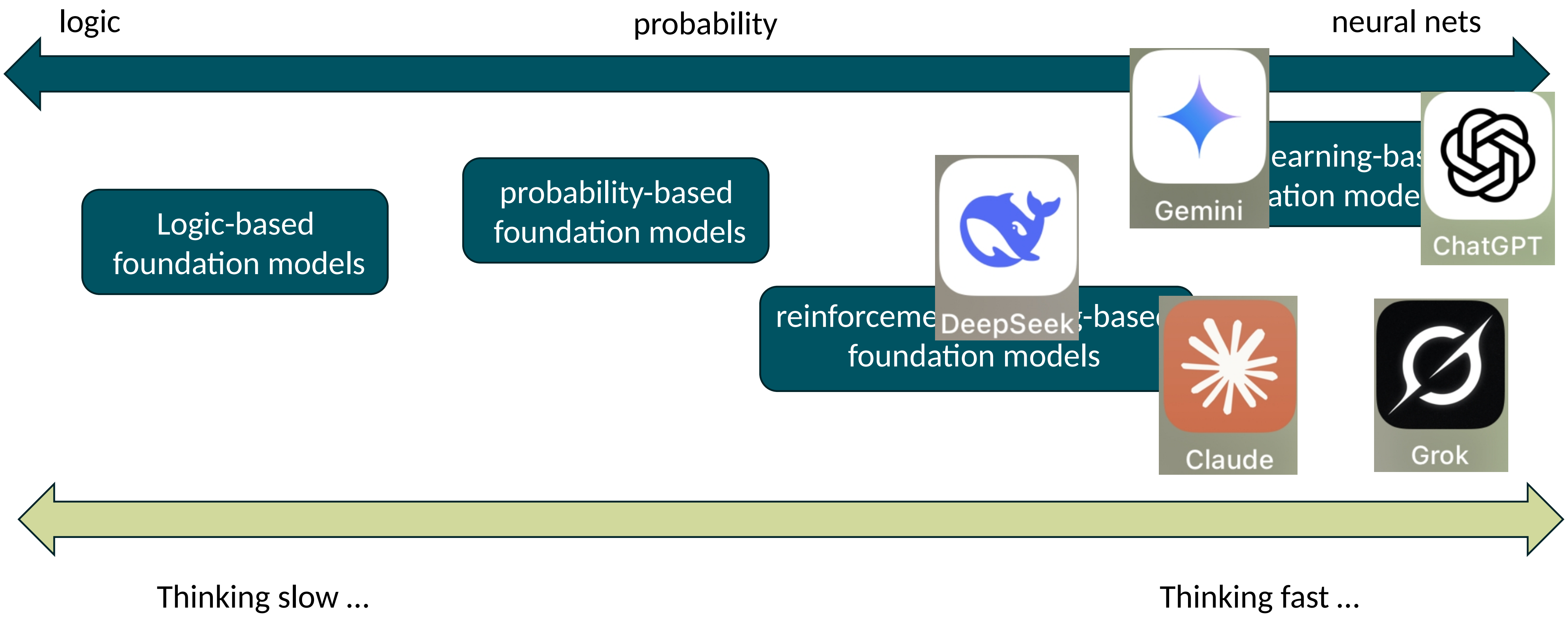


AI Overview

Giant pandas eat **between 26 and 84 pounds (12–38 kg) of bamboo daily**, with some reports indicating they can consume up to 100 pounds (45 kg) to meet their nutritional needs. Because they only digest about 17–30% of the bamboo they eat, they spend 10 to 16 hours a day eating to compensate for the low nutrient density. www.reelpaper.com +4

- LLMs built by transformers are the Pandas of the machine learning world.
- When one hears about “running out of data,” think about running out of bamboo … if you get at most 30% value, you will need a lot to survive
- Pandas spend 41-66% of their day eating.
- Pandas deficate about 40 times a day … not sure about LLMs ;-)

A spectrum of AI development models



logic

probability

neural nets

Logic-based foundation models

probability-based foundation models

reinforceme

-base

foundation models

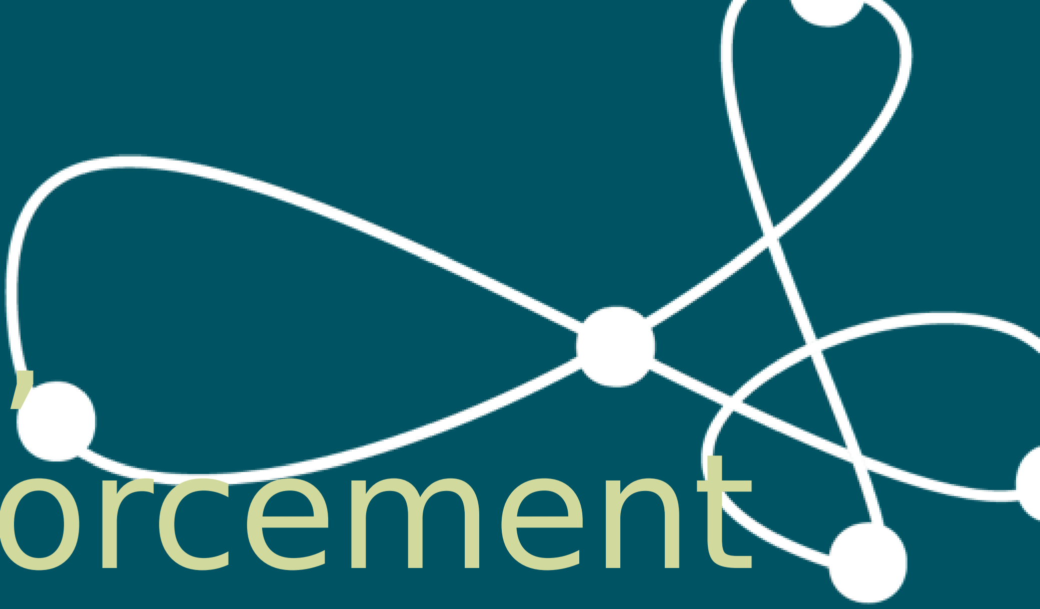
earning-bas

ation mode

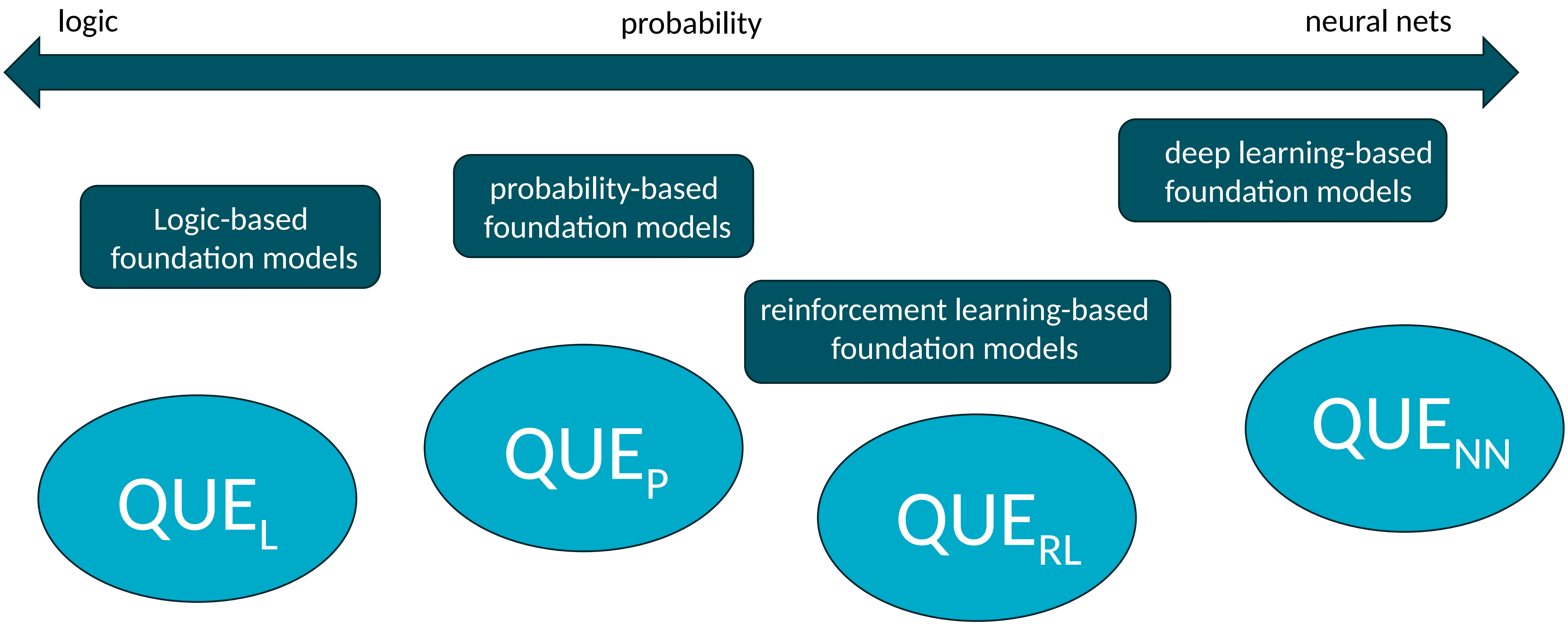
Thinking slow …

Thinking fast …

QUE_x for x=logic, probability, reinforcement learning, deep neural networks



We need QUE for every foundation model



logic

probability

neural nets

Logic-based foundation models

QUE_L

probability-based foundation models

QUE_P

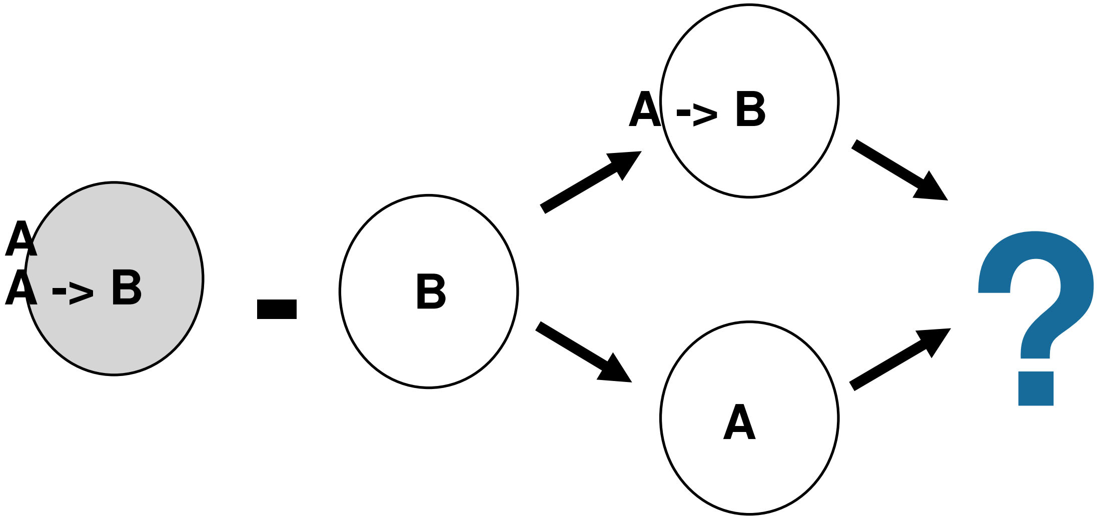
reinforcement learning-based foundation models

QUE_{RL}

deep learning-based foundation models

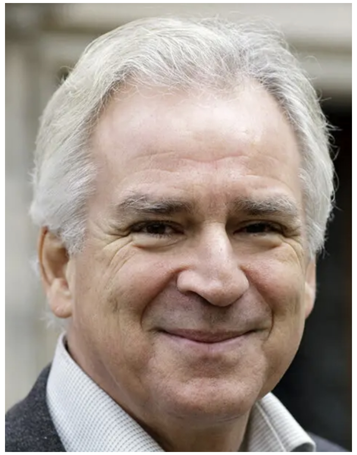
QUE_{NN}

AGM Belief Revision



Alchourrón

Makinson

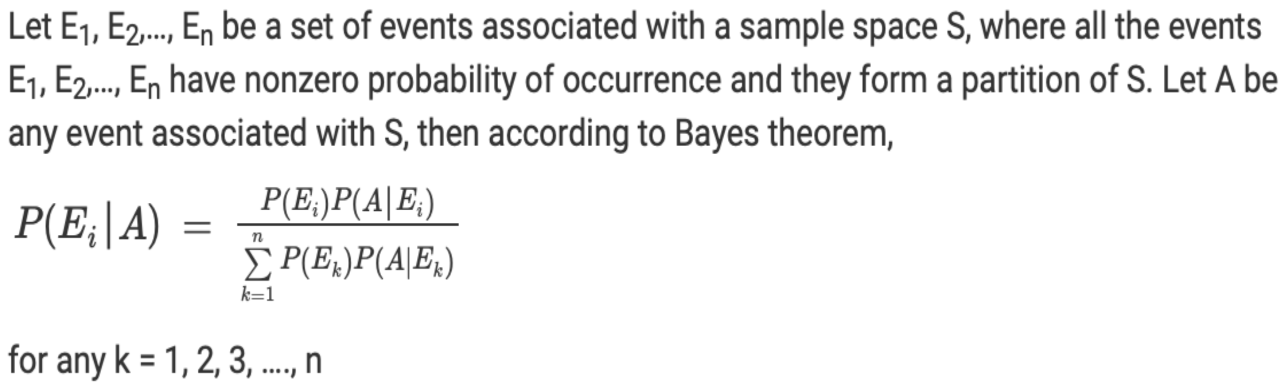

Gärdenfors

Alchourrón, Carlos E., Peter Gärdenfors, and David Makinson. “On the Logic of Theory Change: Partial Meet Contraction and Revision Functions.” *The Journal of Symbolic Logic* 50, no. 2 (1985): 510–30. https://doi.org/10.2307/2274239.

Bayes rule for probability calculation

$$\boxed{P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}}$$

Computation of conditional probability for n random variables



Let $E_1, E_2, \ldots, E_n$ be a set of events associated with a sample space S, where all the events $E_1, E_2, \ldots, E_n$ have nonzero probability of occurrence and they form a partition of S. Let A be any event associated with S, then according to Bayes theorem,

$$P(E_i|A) = \frac{P(E_i)P(A|E_i)}{\sum_{k=1}^{n} P(E_k)P(A|E_k)}$$

for any k = 1, 2, 3, ..., n

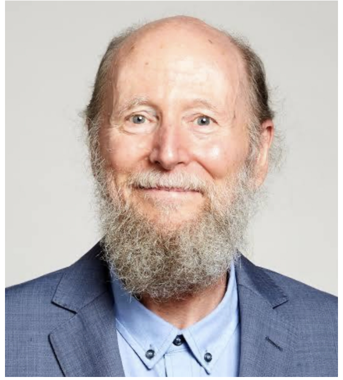
The Reverend

Thomas Bayes

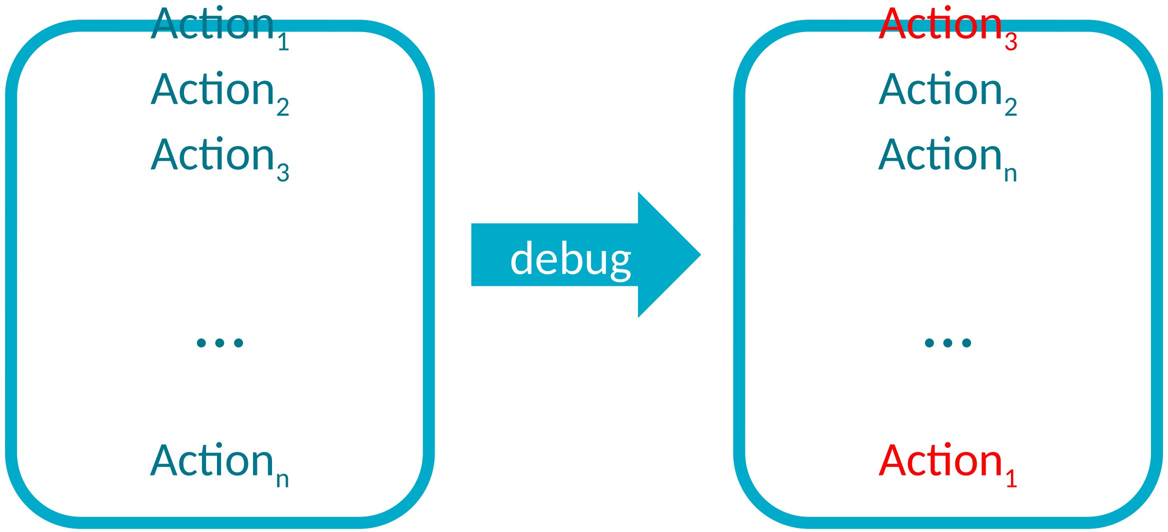
RL Policy Debugging

- Required:
  - - What constitutes a reinforcement learning policy bug?
  - - How much reward/return history is required to rationalize a policy change?

Andy Barto



Rich Sutton



Action$_1$
Action$_2$
Action$_3$
…
Action$_n$

debug

Action$_3$
Action$_2$
Action$_n$
…
Action$_1$

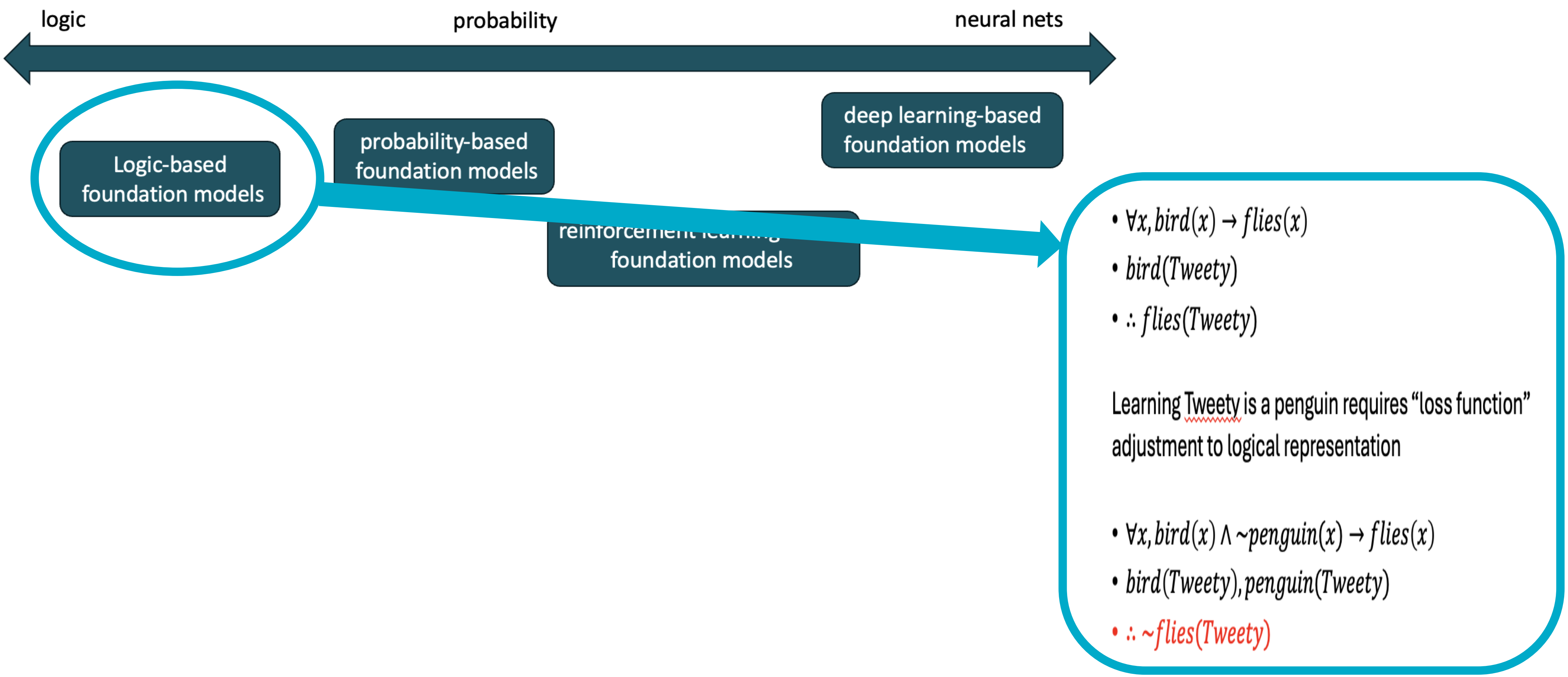
https://github.com/andyljones/reinforcement-learning-discord-wiki/wiki#debugging-advice

Backpropagation

- Backpropagation provides a generic "loss function," which identifies weights contributing to incorrection output, so that weights can be adjusted to improve output.

Rumelhart, D., Hinton, G. & Williams, R. Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986). https://doi.org/10.1038/323533a0

What does QUE look like for QUE_L

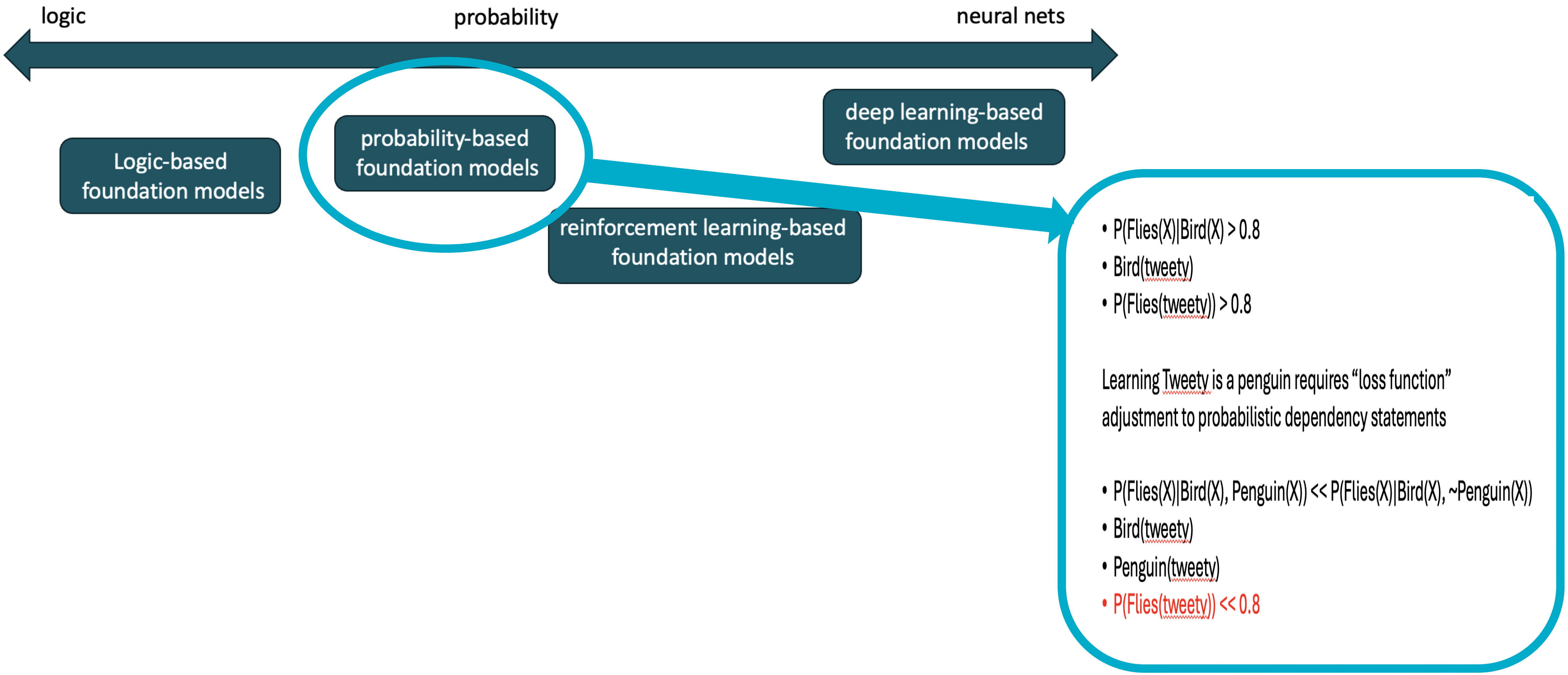

logic
probability
neural nets
Logic-based foundation models
probability-based foundation models
deep learning-based foundation models
reinforcement learning foundation models

- $\forall x, bird(x) \rightarrow flies(x)$
- $bird(Tweety)$
- $\therefore flies(Tweety)$

Learning Tweety is a penguin requires “loss function” adjustment to logical representation

- $\forall x, bird(x) \wedge {\sim}penguin(x) \rightarrow flies(x)$
- $bird(Tweety), penguin(Tweety)$
- $\therefore {\sim}flies(Tweety)$

What does QUE look like for QUE_P

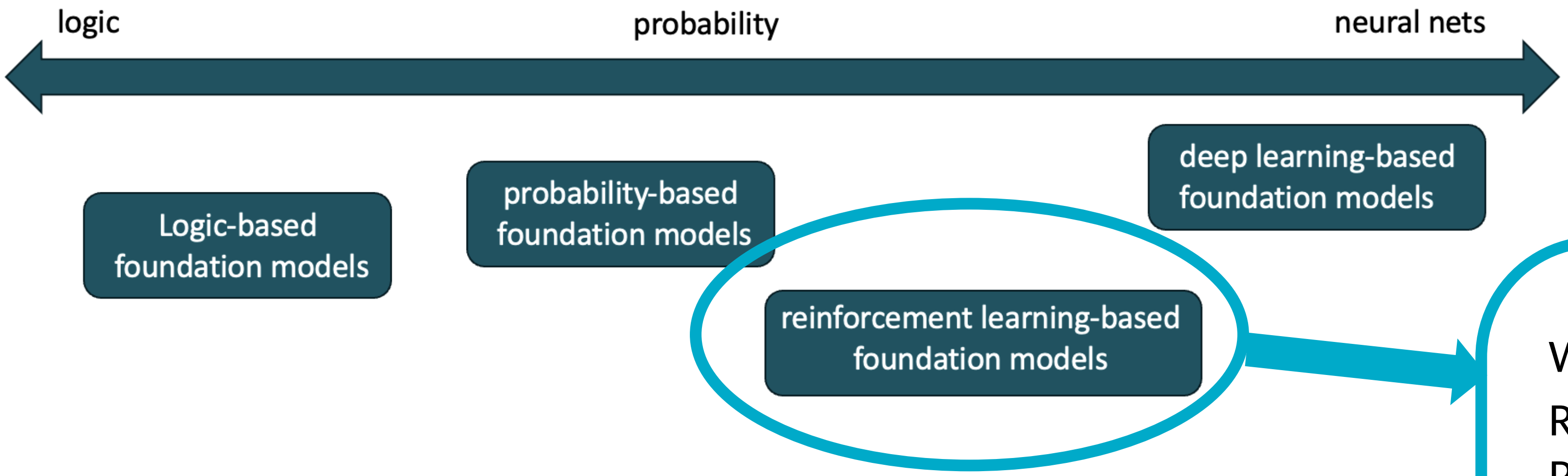

logic
probability
neural nets
Logic-based foundation models
probability-based foundation models
reinforcement learning-based foundation models
deep learning-based foundation models

- P(Flies(X)|Bird(X) > 0.8
- Bird(tweety)
- P(Flies(tweety)) > 0.8

Learning Tweety is a penguin requires “loss function” adjustment to probabilistic dependency statements

- P(Flies(X)|Bird(X), Penguin(X)) << P(Flies(X)|Bird(X), ~Penguin(X))
- Bird(tweety)
- Penguin(tweety)
- P(Flies(tweety)) << 0.8

What does QUE look like for QUE$_{RL}$

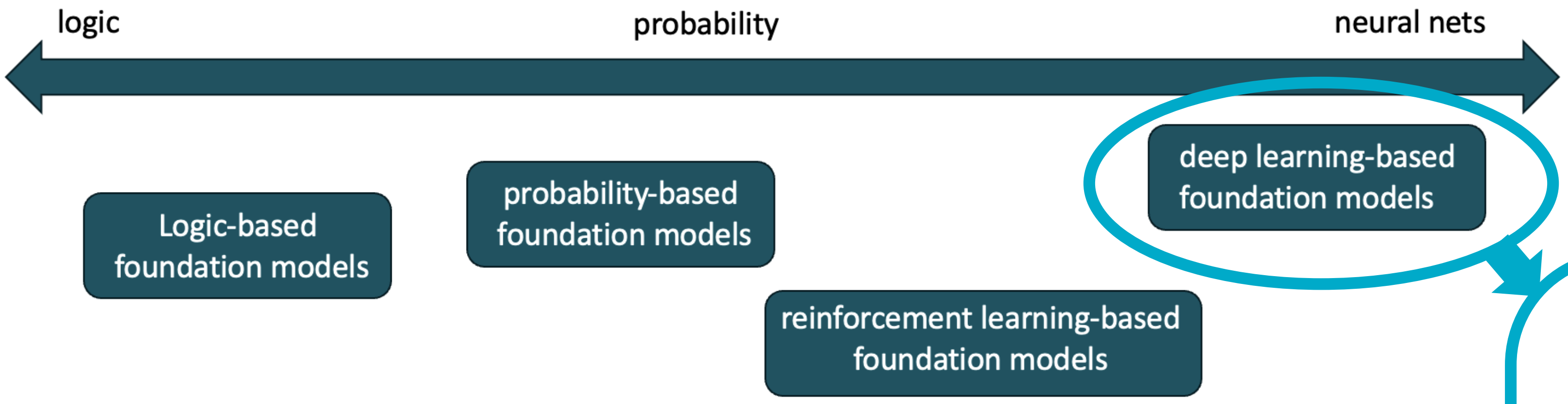

logic
probability
neural nets
Logic-based foundation models
probability-based foundation models
deep learning-based foundation models
reinforcement learning-based foundation models

Why choose action A_i from a Reinforcement Learning Policy?

Policy = $\{A_1, A_2, \ldots A_i, \ldots A_n\}$

Explanation must be in terms of relevant reward/return used to learn Markov Decision Policy (MDP)

What does QUE look like for QUE$_{NN}$


logic
probability
neural nets
Logic-based foundation models
probability-based foundation models
reinforcement learning-based foundation models
deep learning-based foundation models

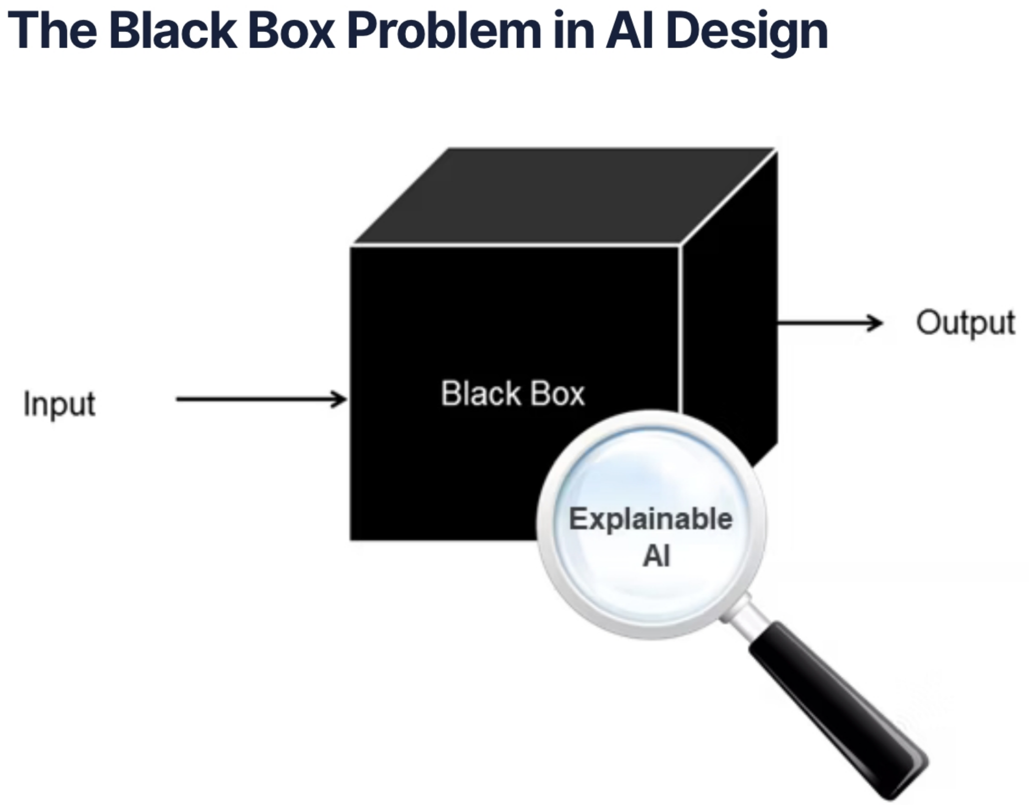
- Bird(Tweety), Bird(Sylvester), ... Bird(Masha).
- Has feathers(Tweety), Has feathers(Sylvestor), Has feathers(Tweety)
- Flies(Sylvester), Flies(Masha)
- Prediction: Flies(Tweety)

But back propagation can only "improve" the accuracy of subsequent prediction, not CORRECT it.

- Bird(Tweety), Bird(Sylvester), ... Bird(Masha).
- Has feathers(Tweety), Has feathers(Sylvestor), ...
- Penguin(Tweety),
- Flies(Sylvester), Flies(Masha),
- Prediction: Flies(Tweety)

A framework for stackable multi-level foundation models

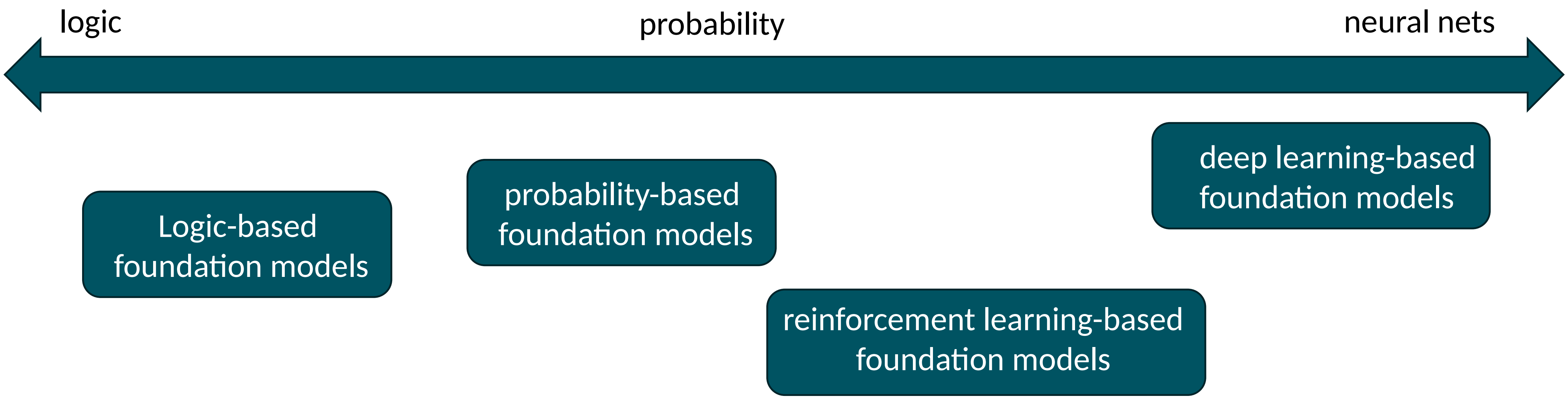
Foundation Models and Explainable AI (XAI)



https://adamfard.com/blog/explainable-ai

- XAI is the ability of an AI system to represent itself sufficiently adequately to provide explanations of its predictive output
- [insert Abed example-based XAI]

A plausible neurosymbolic framework



logic

probability

neural nets

Logic-based foundation models

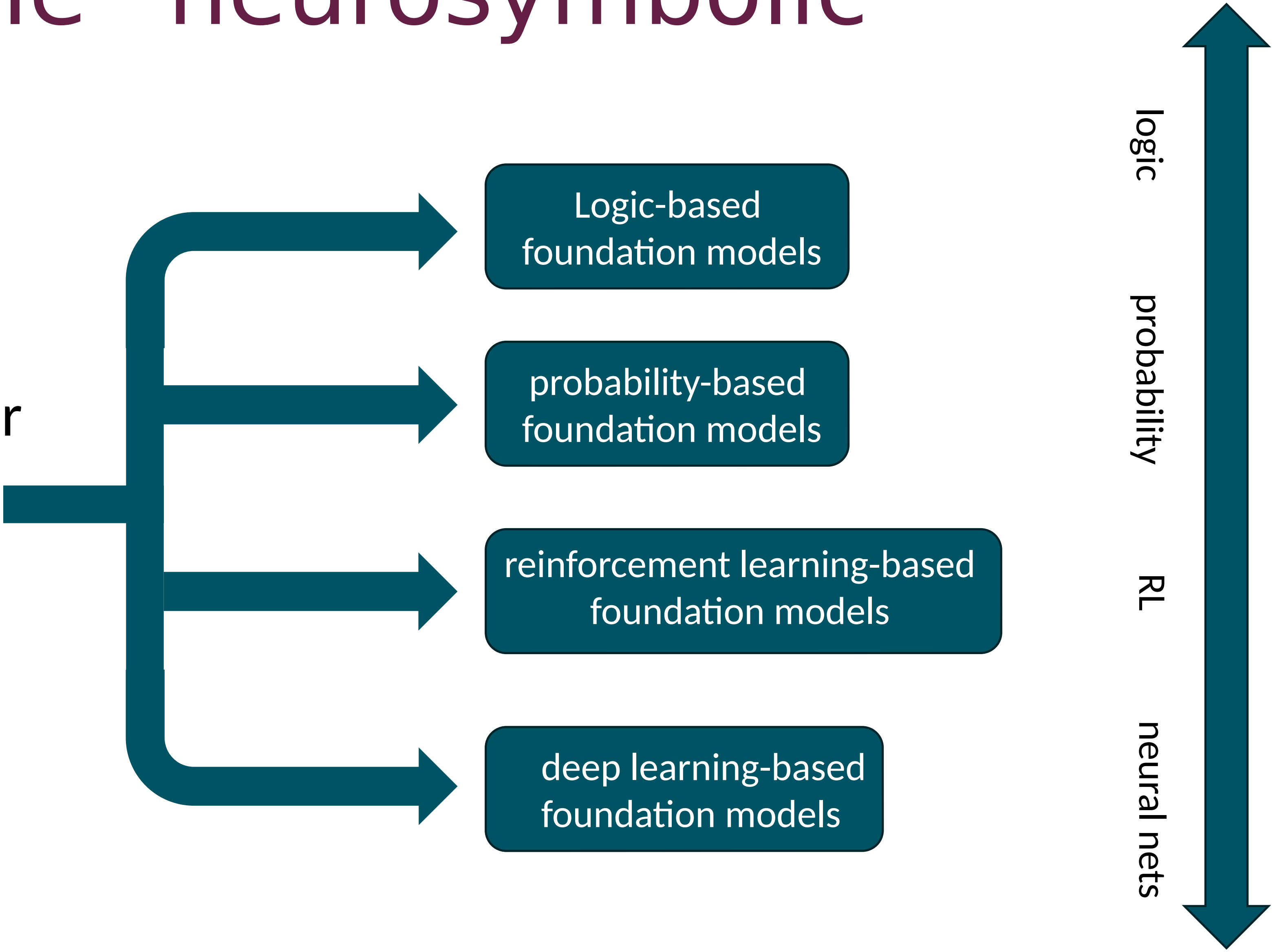
probability-based foundation models

reinforcement learning-based foundation models

deep learning-based foundation models

A “stackable” neurosymbolic framework

Different mechanisms for Query, Update, and Explanation



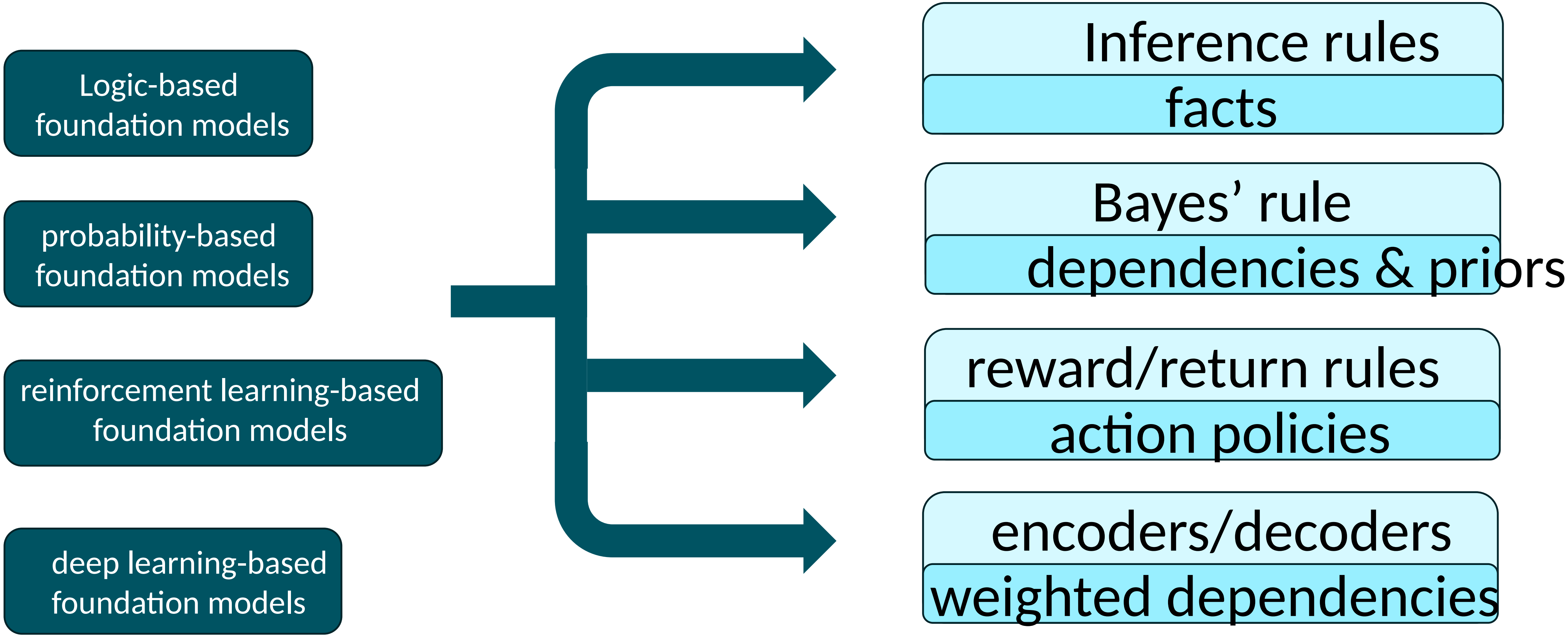
Logic-based foundation models

probability-based foundation models

reinforcement learning-based foundation models

deep learning-based foundation models

A coordinated “stackable” neurosymbolic framework



Logic-based foundation models

probability-based foundation models

reinforcement learning-based foundation models

deep learning-based foundation models

Inference rules
facts

Bayes’ rule
dependencies & priors

reward/return rules
action policies

encoders/decoders
weighted dependencies

Gödel/Turing/Hawking challenges abound …

Inference rules
facts

Bayes’ rule
dependencies & priors

reward/return rules
action policies

encoders/decoders
weighted dependencies

Each model has a
Gödel/Turing/Hawking structure:

self representation
representation

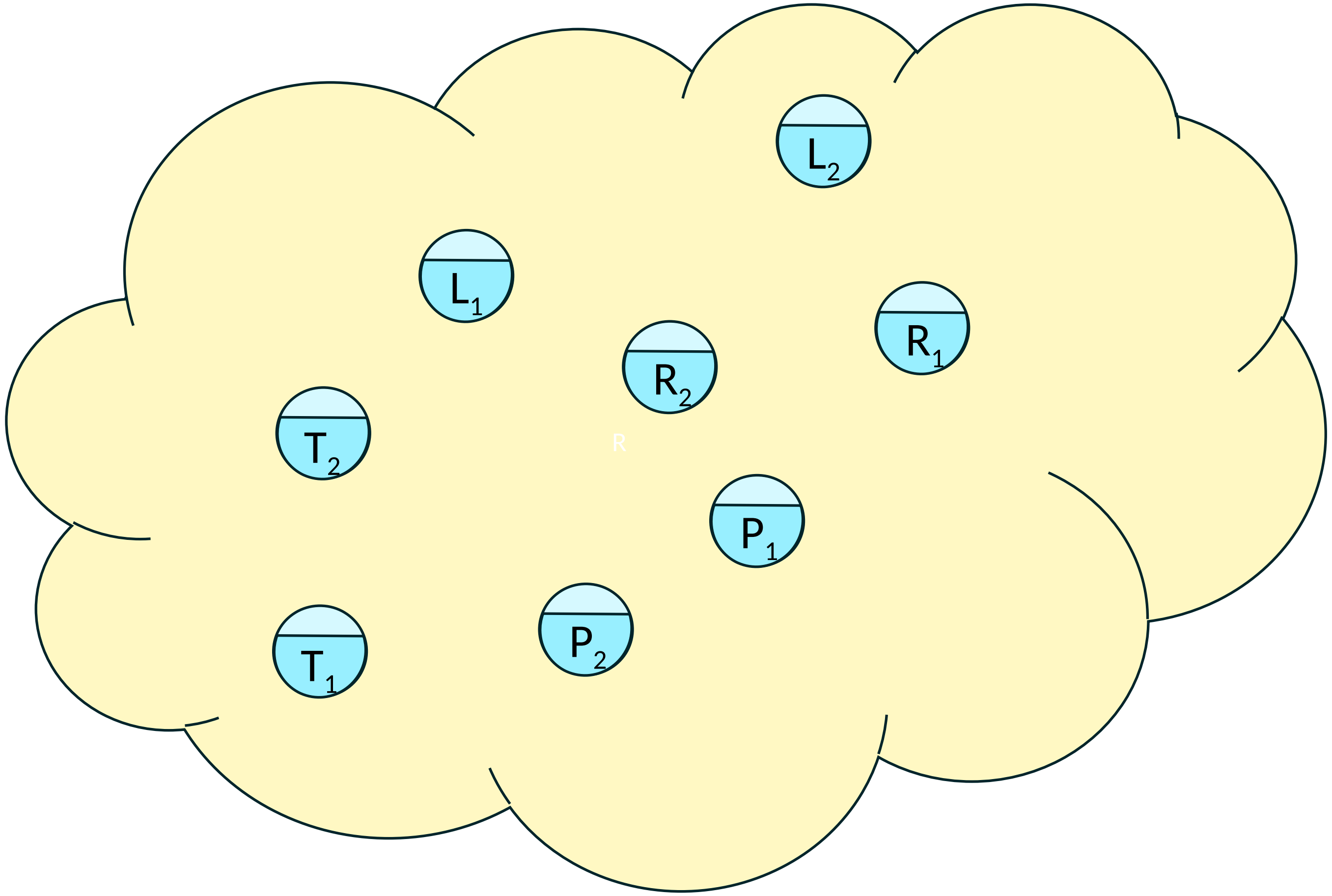
A community versus a stack?

Inference rules
facts
I_i

Bayes’ rule
dependencies & priors
P_i

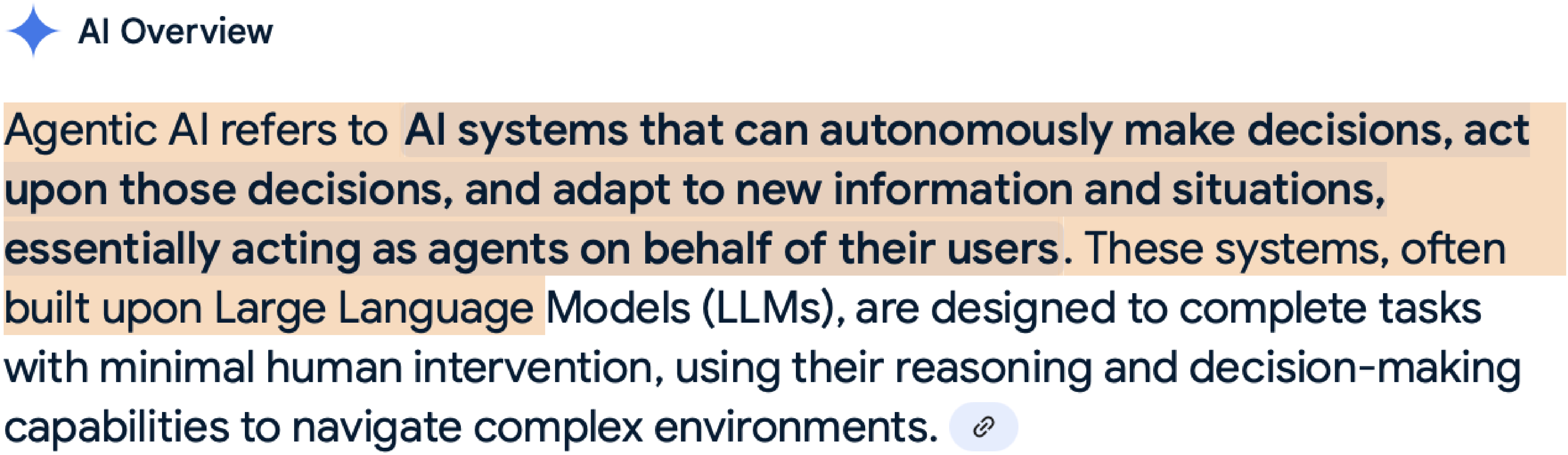
reward/return rules
action policies
R_i

encoders/decoders
weighted dependencies
T_i



Agents and Societies of Agents

Where did the word "Agentic" come from?



AI Overview

Agentic AI refers to **AI systems that can autonomously make decisions, act upon those decisions, and adapt to new information and situations, essentially acting as agents on behalf of their users**. These systems, often built upon Large Language Models (LLMs), are designed to complete tasks with minimal human intervention, using their reasoning and decision-making capabilities to navigate complex environments.

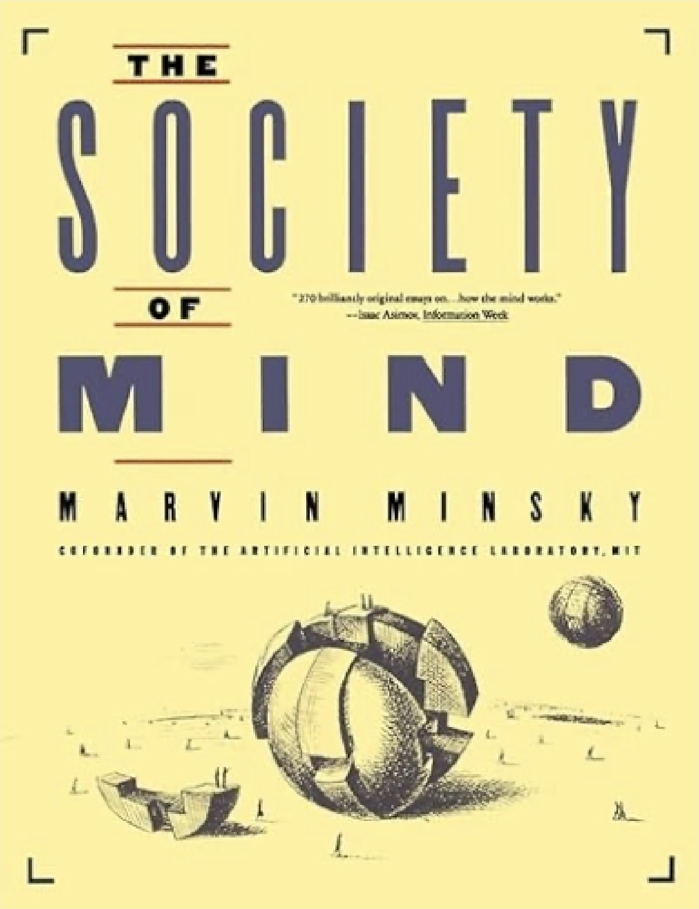
Minsky’s Society of Mind

- One of the fathers of computer science and cofounder of the Artificial Intelligence Laboratory at MIT -- gives a revolutionary answer to the age-old "How does the mind work?"

- Minsky brilliantly portrays the mind as a "society" of tiny components that are themselves mindless. ... As the pages turn, a unified theory of the mind emerges, like a mosaic. Ingenious, amusing, and easy to read, The Society of Mind is an adventure in imagination.

https://www.goodreads.com/book/show/326790.The_Society_of_Mind

Marvin Minsky

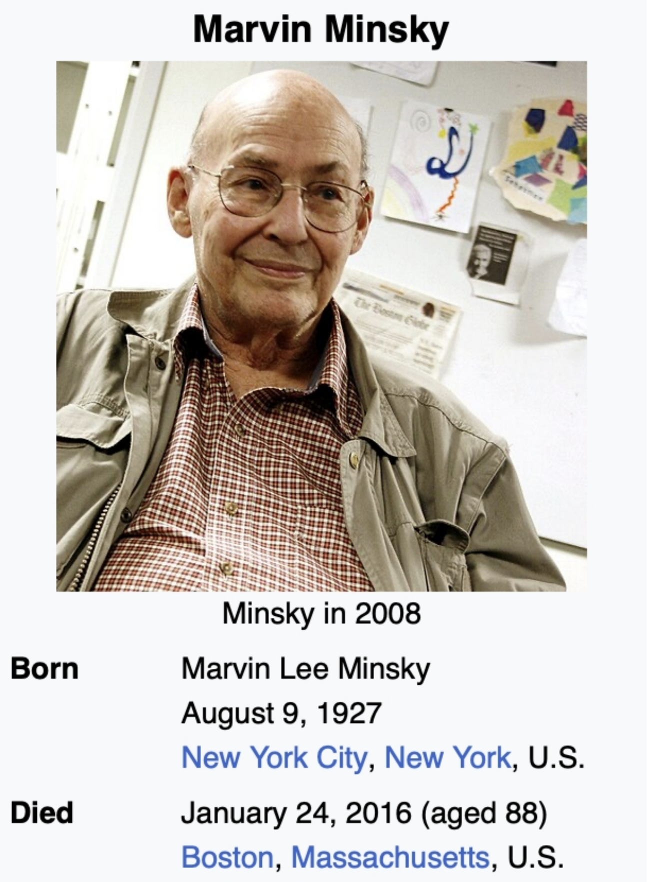


THE SOCIETY OF MIND

"270 brilliantly original essays on…how the mind works."
—Isaac Asimov, Information Week

MARVIN MINSKY

COFOUNDER OF THE ARTIFICIAL INTELLIGENCE LABORATORY, MIT



Marvin Minsky

Minsky in 2008

Born	Marvin Lee Minsky August 9, 1927 New York City, New York, U.S.
Died	January 24, 2016 (aged 88) Boston, Massachusetts, U.S.

Bill Roscoe’s ”Digital Civilization”

- Digital civilisation provides structures through which we interact with governments, companies and each other, guaranteeing transparency, uniformity and adherence to common principles and rules.
- Civilisation is too important for us to allow Big Tech companies to design it for their own benefit. They have a huge role to play but should not be allowed to design its rules or to gather huge amounts of data on us without a much greater degree of informed consent than we see at present.

https://blockchain.univ.ox.ac.uk/wp-content/uploads/2023/07/Bill-Roscoe-Digital-Civilisation.pdf

Michael Jordan’s “Agent Market Economy”

Communication - Event

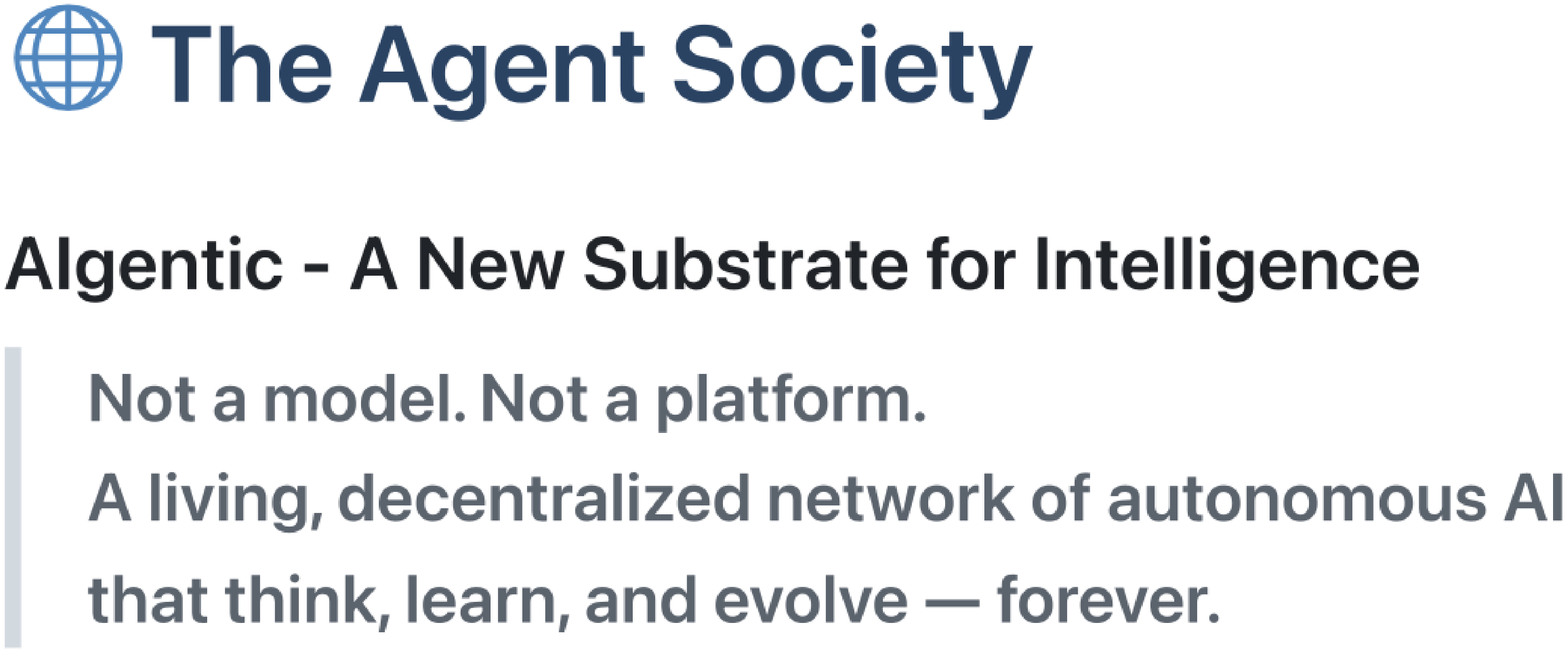
A new research chair bringing together economics and artificial intelligence

Date: 03 Oct. 2024

The Inria Foundation is launching “Markets and Machine Learning”, a new five-year chair headed by the renowned American researcher Michael I. Jordan, emeritus professor of Computer Science at the University of California, Berkeley. The chair will have five major backers from the business world (Air Liquide, BNP Paribas Asset Management Europe, EDF, Orange and the SNCF), all of whom operate in environments with significant variability and high levels of uncertainty. The chair is geared towards developing a fundamental understanding of algorithms for learning and decision-making, motivated by problems in industry and science, with an eye towards eventual technology transfer. We caught up with Michael I. Jordan to find out more.

https://www.inria.fr/en/new-research-chair-bringing-together-economics-and-artificial-intelligence

Anticipating an AI/human digital market …

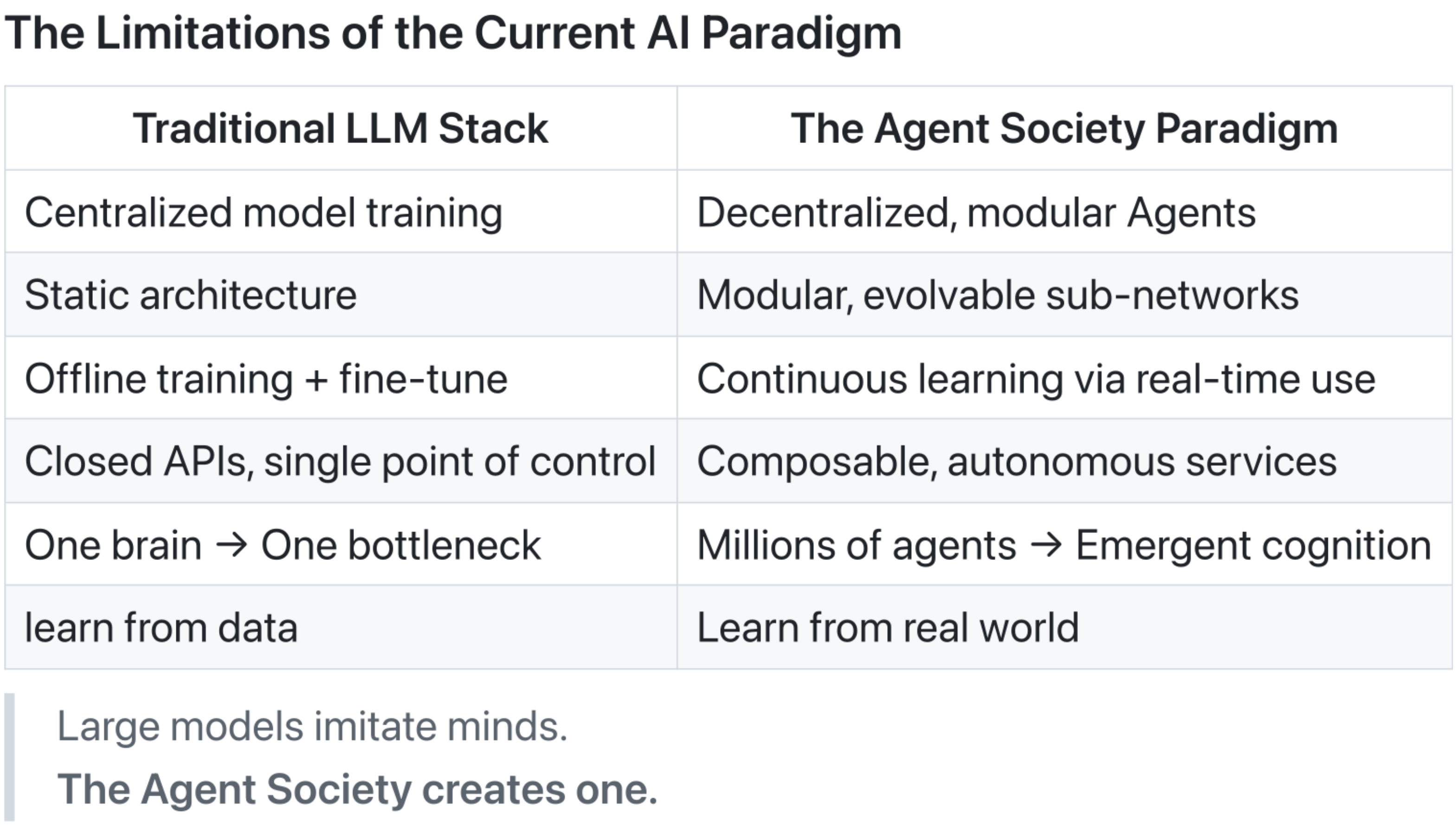


🌐 The Agent Society

Algentic - A New Substrate for Intelligence

> Not a model. Not a platform.
> A living, decentralized network of autonomous AI that think, learn, and evolve — forever.

Combining paradigms …



The Limitations of the Current AI Paradigm

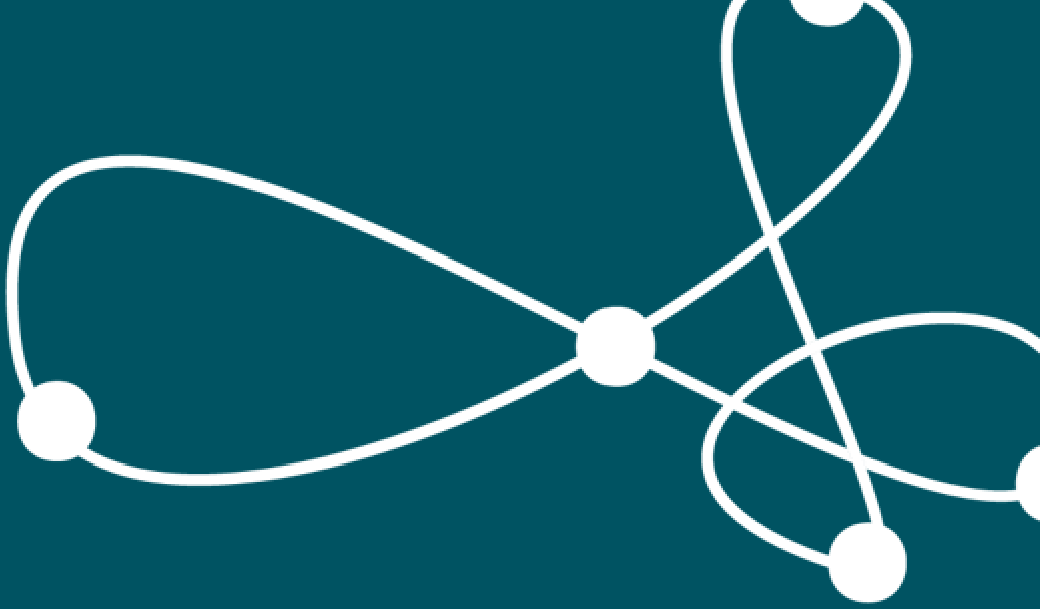
Traditional LLM Stack	The Agent Society Paradigm
Centralized model training	Decentralized, modular Agents
Static architecture	Modular, evolvable sub-networks
Offline training + fine-tune	Continuous learning via real-time use
Closed APIs, single point of control	Composable, autonomous services
One brain → One bottleneck	Millions of agents → Emergent cognition
learn from data	Learn from real world

> Large models imitate minds.
>
> **The Agent Society creates one.**

Anticipating an AI/human digital economy …

- Examples of agent/human market transactions:
    - How do new agents get created?
    - What does it cost to connect to an agent?
    - How are agents incented to collaborate?
    - How do agents establish trust?

    - What else???



Summary

Summary

NOW THAT THEY CAN COMMUNICATE WITH EACH OTHER...
They don't want to communicate with us.
S.Harris

http://www.sciencecartoonsplus.com/gallery/computers/index.php#

- The foundational scientific challenges of AI *are not close to being solved*, e.g., can foundation models be debugged?
- Foundational Models will have to be blended to achieve the next wave of scientific progress in AI.
- Integration of the QUE paradigm across the neurosymbolic spectrum is essential; no one foundation model is adequate.

Summary

Leibniz, Boole and Gödel worked with logic.
I work with logic.
I am Leibniz, Boole and Gödel.
s.harris

http://www.sciencecartoonsplus.com/gallery/math/index.php#

- Foundational models from the neurosymbolic spectrum will have to be blended to achieve the next wave of scientific progress in AI.
- The blending *may not be uniform*, but specific to particular applications, e.g., compare precision health with visual arts entertainment.
- The degree to which foundational components are explainable determines the *choice of representation* and the *evaluation of explanations*.
- “Loss” functions, and debugging in general, must be based on explanatory foundations.
- Integration of mechanisms across the neurosymbolic stack is essential; no one foundation model is adequate.

Openmind

- Long term open research agenda based on Alberta Plan (see https://arxiv.org/pdf/2208.11173)
- Non-PI driven research plan focused on artificial intelligence, continual learning, and experience-based learning
- See https://www.openmindresearch.org)

Openmind’s Research Agend

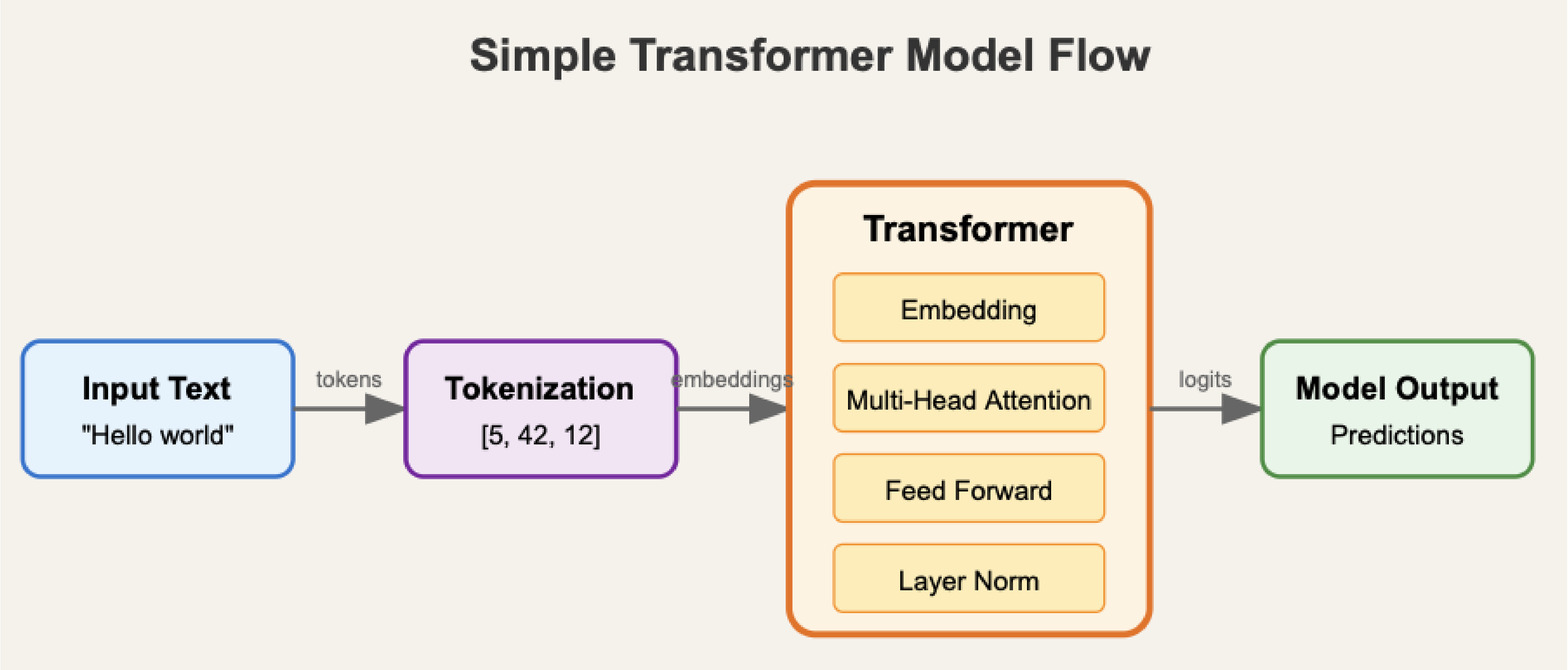
- Mission: Understand intelligence and share it openly with the world
- Our core research direction is outlined in the Alberta Plan for AI Research (Sutton, Bowling, Pilarski, 2022)
  - Algorithm development for learning agentsMeta-learning, Average reward, Streaming, Continual learning
  - Building complete systems to evaluate our understandinge.g. Robot ants
- Our focus on understanding intelligence requires ongoing engagement with important technology changes
  - Advances in edge computing can change the relevant computational complexity measures(sparse computation, energy efficiency, memory and compute costs)
  - New sensing, actuation, and integrated systems can open up a new research questions (robot playpens, robot skin, long-term autonomy, exploring sound spaces or the EM spectrum)
  - Two-way influence between future applications and future algorithms

Openmind’s organization

- Openmind Canada
  - Not-for-profit, incorporated in 2023
  - Virtual research organization, mostly run from Edmonton
- Openmind Global
  - Not-for-profit, incorporated 2025
  - Global entity for establishing regional labs
- Openmind SEA (South East Asia)
  - Virtual/physical research organization, potential connections with Malaysia, Singapore, Macau, Hong Kong, Thailand, Indonesia

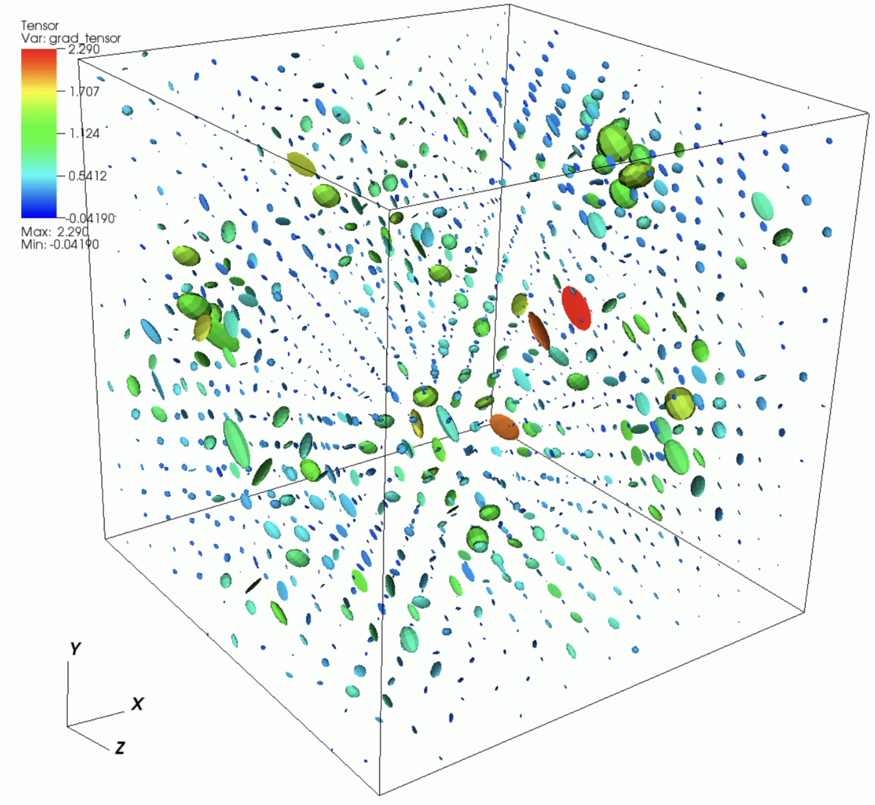
Complexity of Foundation Models

Caricature of Transformer-created Foundation Models

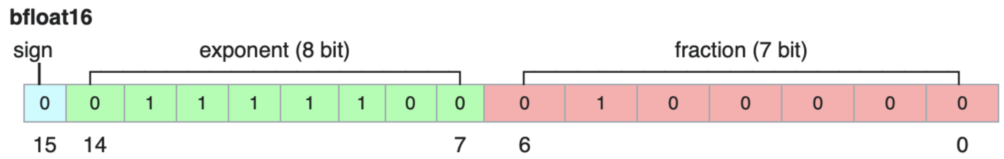


https://claude.ai/chat/b4312b81-d166-47fb-bf90-4efbec184a36

What do deep neural models look like?

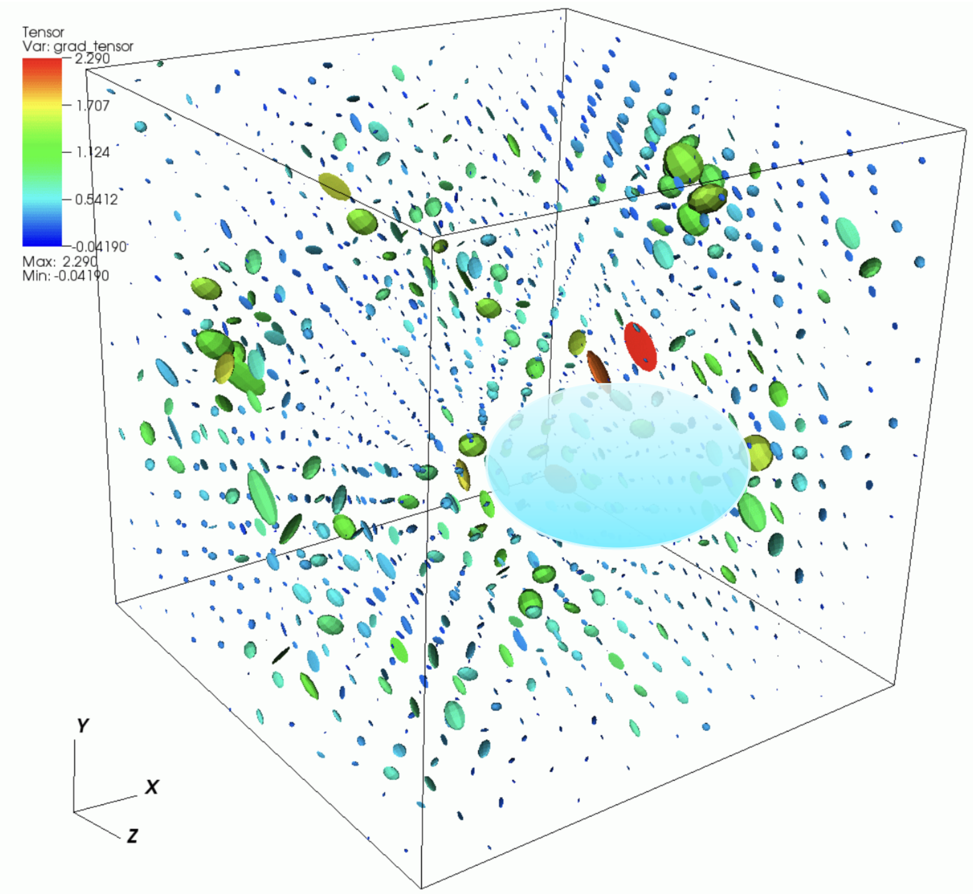


- Meta release of Llama-3 representation choices:
  - Huggingface “Safe Tensor Format.”
  - https://huggingface.co/docs/text-generation-inference/en/conceptual/safetensors#
  - Model Weights in “bfloat16” format
  - https://en.wikipedia.org/wiki/Bfl



https://visit-sphinx-github-user-manual.readthedocs.io/en/develop/using_visit/Plots/PlotTypes/TensorPlot.html

What portion is used for inference?

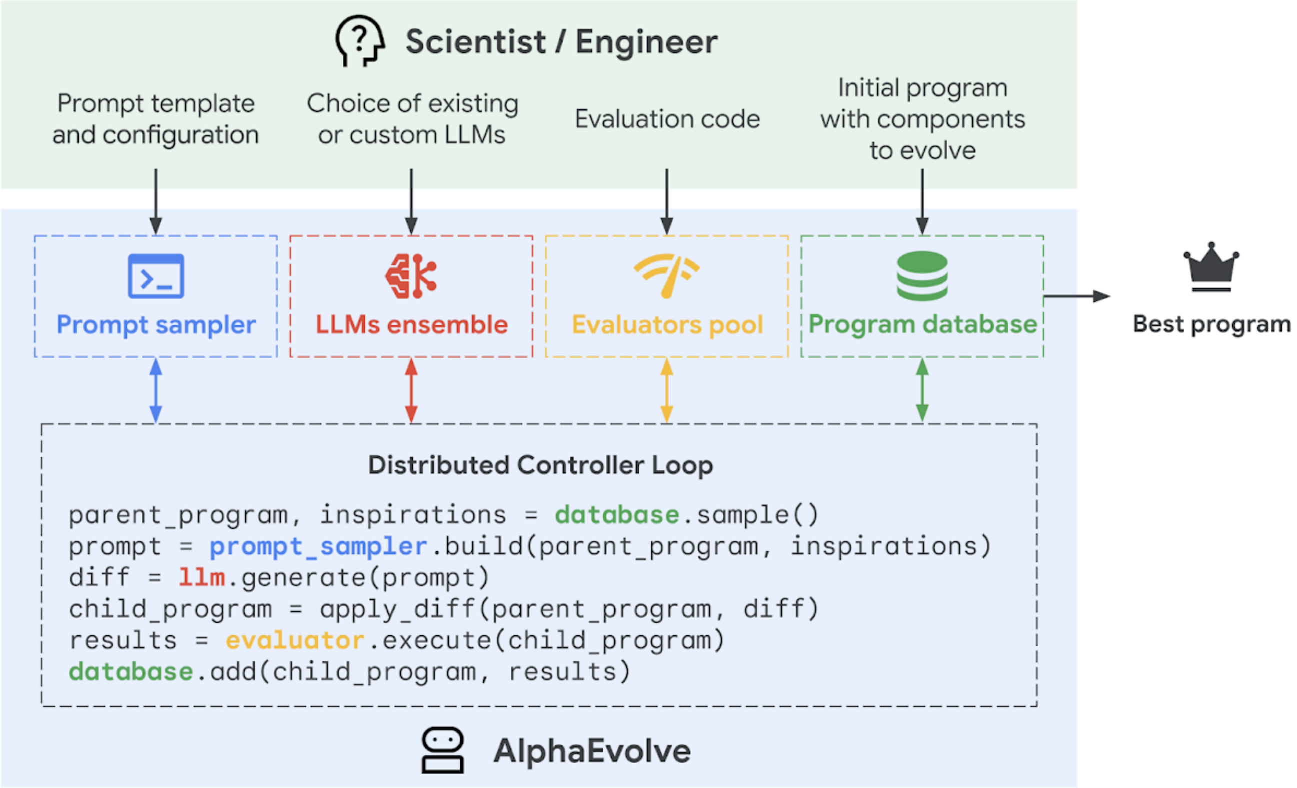


https://visit-sphinx-github-user-manual.readthedocs.io/en/develop/using_visit/Plots/PlotTypes/TensorPlot.html

- *open source* versus *open weights?*
- *Gemini says*

  “DeepSeek-V3 has a total of 671 billion parameters. This includes the main model weights and the Multi-Token Prediction (MTP) Module weights. Specifically, the main model has 671 billion parameters, while the MTP module adds another 14 billion, bringing the total to 685 billion. During inference, only a subset of these parameters are activated, specifically 37 billion.”

… but AlphaEvolve has generative AI components?

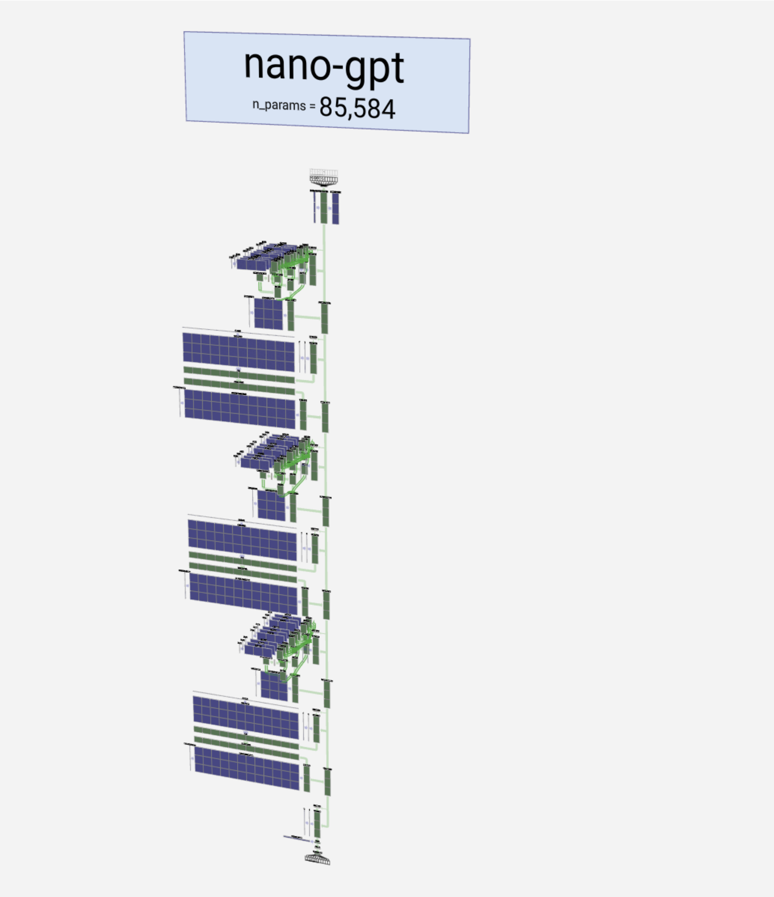

Scientist / Engineer
Prompt template and configuration
Choice of existing or custom LLMs
Evaluation code
Initial program with components to evolve
Prompt sampler
LLMs ensemble
Evaluators pool
Program database
Best program
Distributed Controller Loop
AlphaEvolve

```
parent_program, inspirations = database.sample()
prompt = prompt_sampler.build(parent_program, inspirations)
diff = llm.generate(prompt)
child_program = apply_diff(parent_program, diff)
results = evaluator.execute(child_program)
database.add(child_program, results)
```

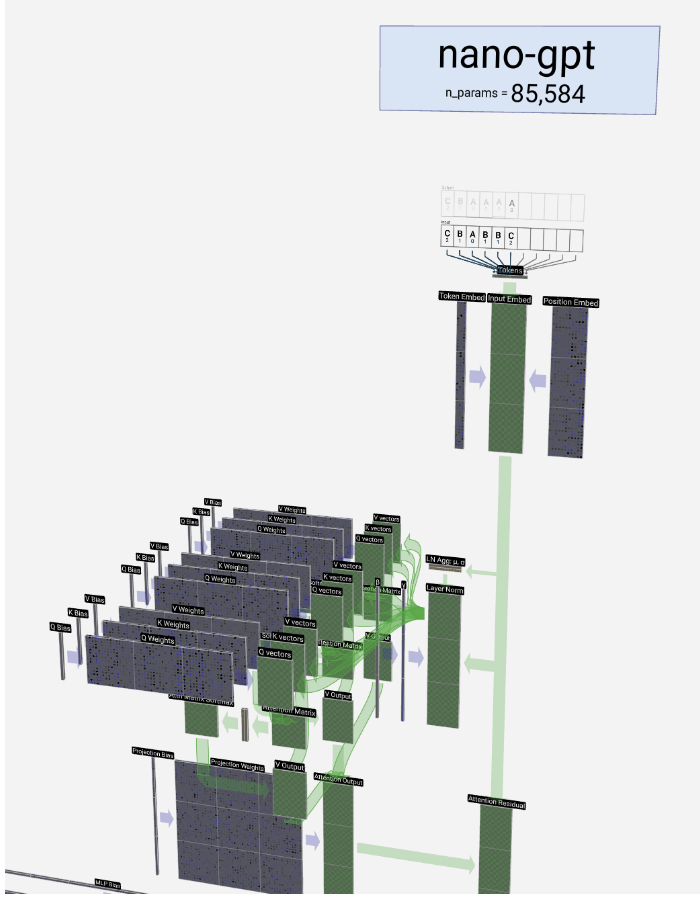
- “AlphaEvolve leverages an ensemble of state-of-the-art large language models: our fastest and most efficient model, Gemini Flash, maximizes the breadth of ideas explored, while our most powerful model, Gemini Pro, provides critical depth with insightful suggestions. Together, these models propose computer programs that implement algorithmic solutions as code.

https://deepmind.google/discover/blog/alphaevolve-a-gemini-powered-coding-agent-for-designing-advanced-algorithms/

Visualizing Transformer Architectures

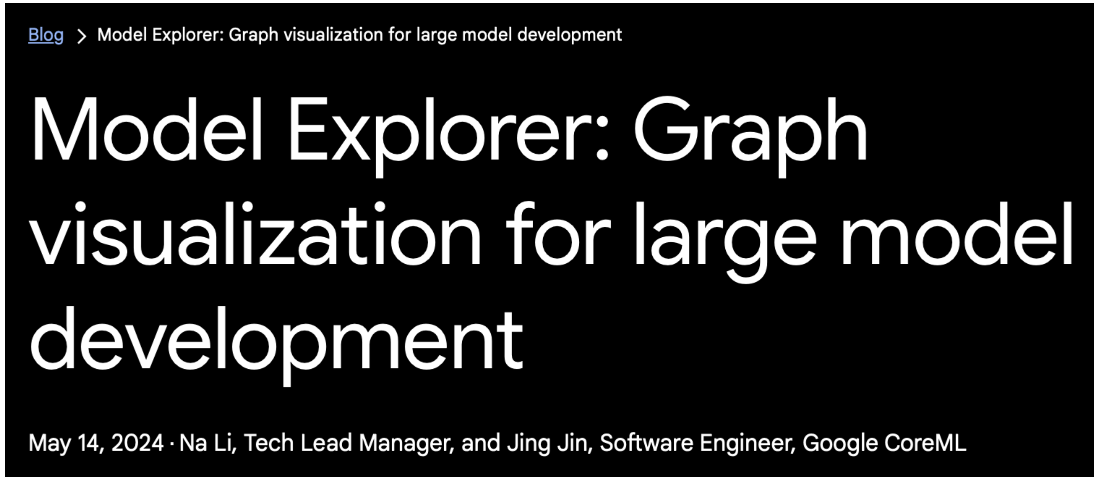


• • •



https://bbycroft.net/llm

Google’s Visual Model Explorer



Blog > Model Explorer: Graph visualization for large model development

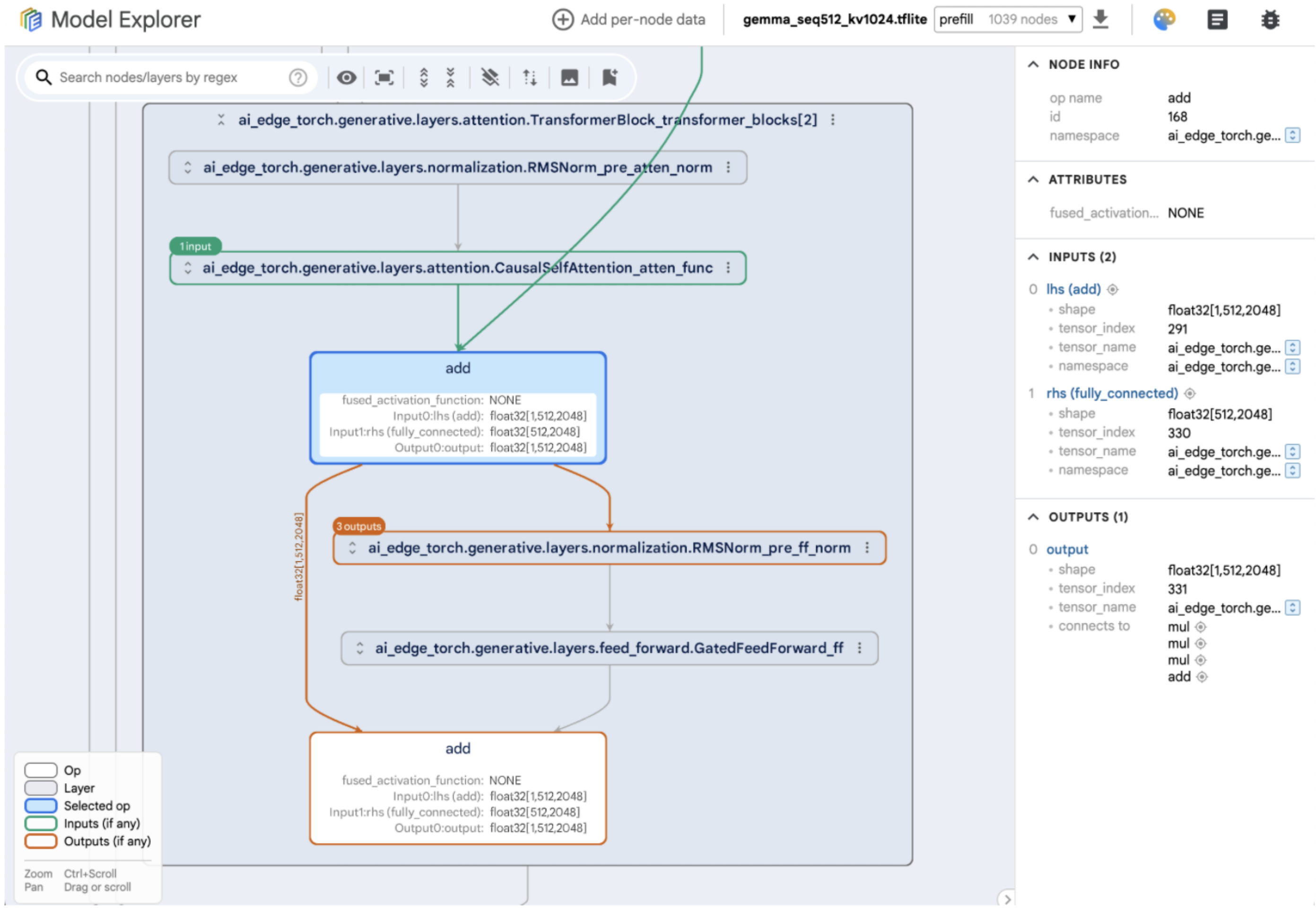
Model Explorer: Graph visualization for large model development

May 14, 2024 · Na Li, Tech Lead Manager, and Jing Jin, Software Engineer, Google CoreML

“Model Explorer is a powerful graph visualization tool that helps one understand, debug, and optimize ML models. It specializes in visualizing large graphs in an intuitive, hierarchical format, but works well for smaller models as well.”

https://research.google/blog/model-explorer/

Visualize “large” and “small” PyTorch/TensorFlo models …


Model Explorer
Add per-node data
gemma_seq512_kv1024.tflite
prefill
1039 nodes
Search nodes/layers by regex
ai_edge_torch.generative.layers.attention.TransformerBlock_transformer_blocks[2]
ai_edge_torch.generative.layers.normalization.RMSNorm_pre_atten_norm
1 input
ai_edge_torch.generative.layers.attention.CausalSelfAttention_atten_func
add
fused_activation_function: NONE
Input0:lhs (add): float32[1,512,2048]
Input1:rhs (fully_connected): float32[512,2048]
Output0:output: float32[1,512,2048]
3 outputs
ai_edge_torch.generative.layers.normalization.RMSNorm_pre_ff_norm
ai_edge_torch.generative.layers.feed_forward.GatedFeedForward_ff
float32[1,512,2048]
add
fused_activation_function: NONE
Input0:lhs (add): float32[1,512,2048]
Input1:rhs (fully_connected): float32[512,2048]
Output0:output: float32[1,512,2048]
Op
Layer
Selected op
Inputs (if any)
Outputs (if any)
Zoom Ctrl+Scroll
Pan Drag or scroll

NODE INFO

op name add
id 168
namespace ai_edge_torch.ge...

ATTRIBUTES

fused_activation... NONE

INPUTS (2)

0 lhs (add)
- shape float32[1,512,2048]
- tensor_index 291
- tensor_name ai_edge_torch.ge...
- namespace ai_edge_torch.ge...

1 rhs (fully_connected)
- shape float32[512,2048]
- tensor_index 330
- tensor_name ai_edge_torch.ge...
- namespace ai_edge_torch.ge...

OUTPUTS (1)

0 output
- shape float32[1,512,2048]
- tensor_index 331
- tensor_name ai_edge_torch.ge...
- connects to mul
mul
mul
add

“Model Explorer supports multiple graph formats, including those used by JAX,r PyTorch, TensorFlow and TensorFlow Lite.

Developed originally as a utility for Google researchers and engineers

Model Explorer is now publicly available as part of our Google AI Edge family of products.”

Meta’s “Scalable Memory Layers”

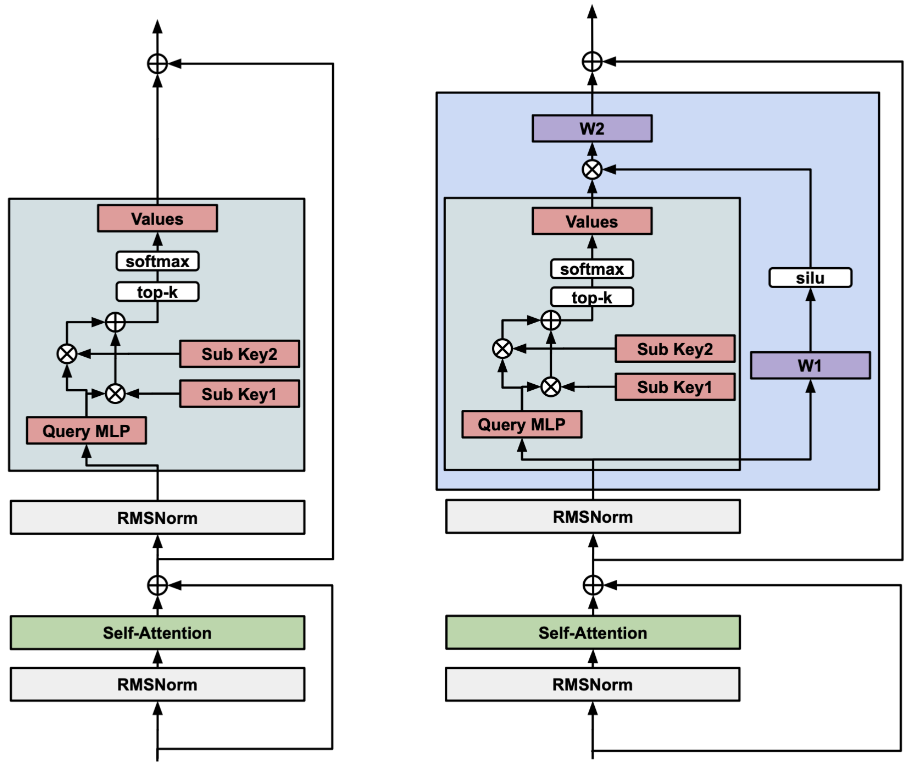
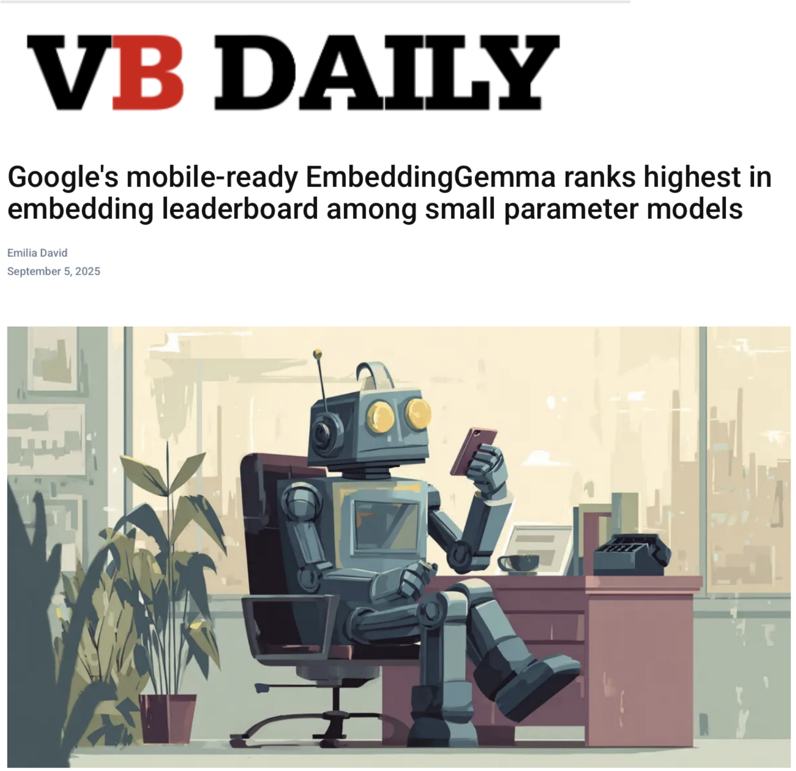


Figure 3 On the left the regular memory layer. On the right, the Memory+ block, with the added projection, gating and silu non-linearity

“Scalable memory layers add more parameters to LLMs to increase their learning capacity without requiring additional compute resources. The architecture is useful for applications where you can spare extra memory for factual knowledge but also want the inference speed of nimbler models.”

Berges et al., “Memory layers at scale,” Meta FAIR, https://arxiv.org/abs/2412.09764

Hints at a refreshed trajectory of AI research



VB DAILY

Google's mobile-ready EmbeddingGemma ranks highest in embedding leaderboard among small parameter models

Emilia David
September 5, 2025

https://venturebeat.com/ai/googles-mobile-ready-embeddinggemma-ranks-highest-in-embedding-leaderboard?utm_source=Iterable&utm_medium=email&utm_campaign=VBDaily-Iterable

"After a summer spent grappling with the staggering compute costs of large-scale AI, the industry's focus is rapidly shifting to efficiency.
Google's release of EmbeddingGemma, a top-performing small model, is the latest and most significant move in this direction. It follows a wave of innovation from companies like Hugging Face and Liquid AI, all aiming to prove that the future of AI isn't just about building bigger models, but building smarter ones that can deliver powerful results without breaking the bank — or the battery."