# AI/ML Fundamentals: Introduction and Market Trends

Apr. 2025

“**Artificial intelligence** (**AI**), in its broadest sense, is intelligence exhibited by machines, particularly computer systems.”

Source: Artificial intelligence - Wikipedia

# AI: definition, history and evolution

## AI Development Highlights (2020- Apr 2025)

### 2020-2023: AI Revolution

- GPT-3® and ChatGPT® advanced natural language understanding.
- AlphaFold® 2 set new benchmarks in protein prediction.
- Governments and organizations began regulating AI with forums and safety summits.

### 2024: More Applications

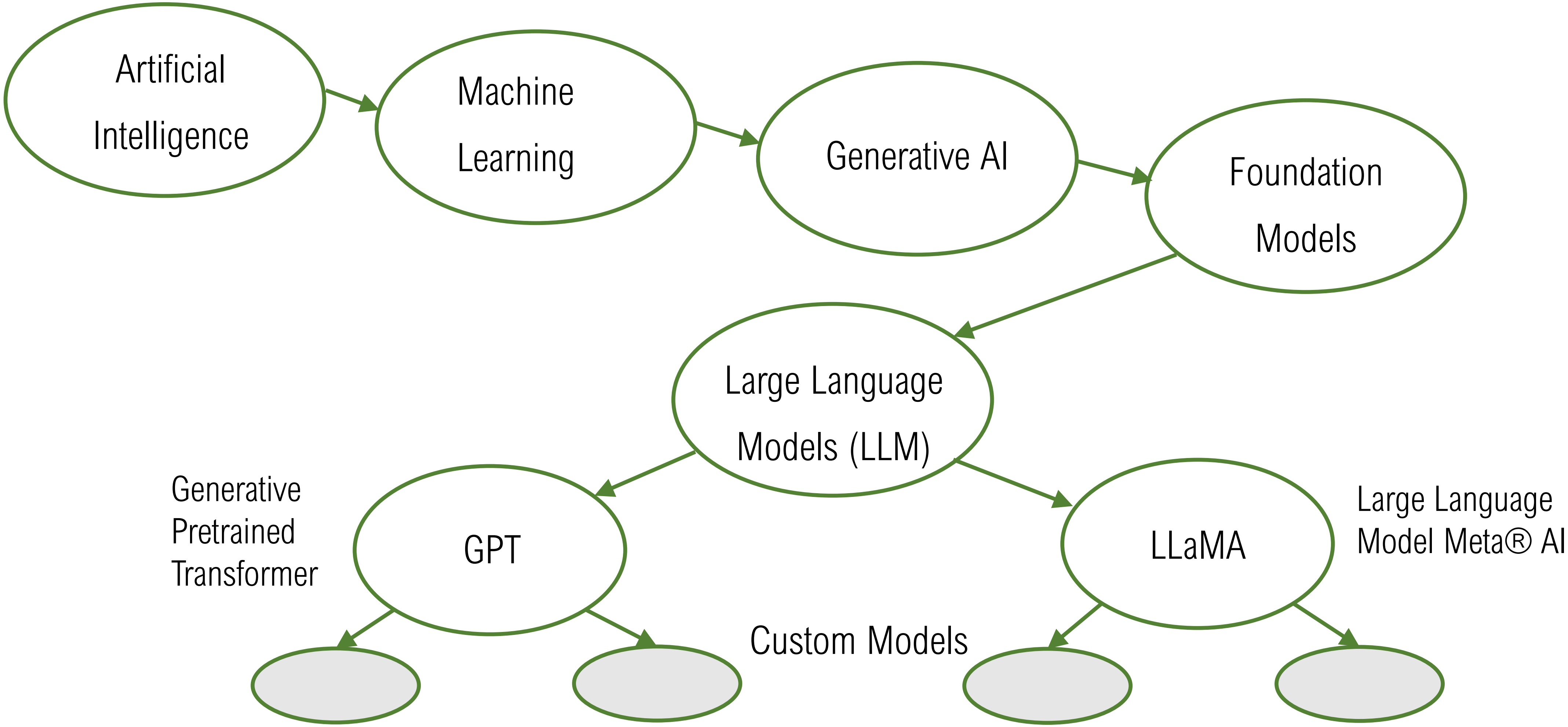
- Google®’s Gemini® 1.5 and OpenAI®’s Sora® debuted advanced AI systems.
- Apple® launched "Apple Intelligence," integrating AI into Siri® and iPhones®.
- GPT-o1® applied inference thinking into the model
- AlphaFold® won the Nobel Prize in Chemistry for revolutionizing protein research.

### 2025: Cost-efficient, Thinking, Open-Weight, and Agents

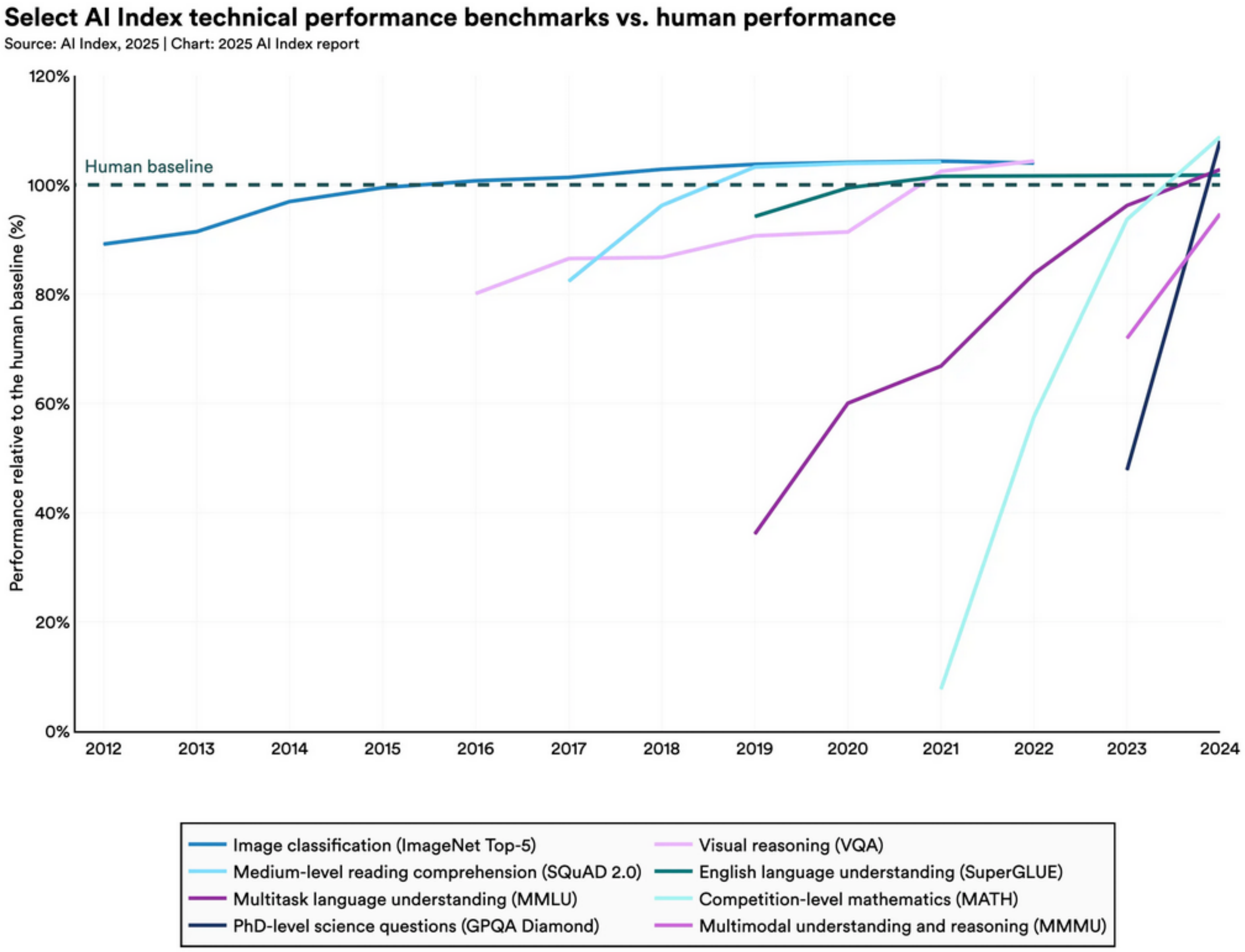
- Deepseek® series with new RL process and low-cost training.
- GPT-4o® multi-modal picture generation gained attention.
- Open-weight models like Deepseek-R1® and Llama4® are gaining popularity.
- Nvidia® Dynamo® inference framework open sourced.
- Anthropic® MCP® standard and Google® A2A® standard released for agents.
- OpenAI® o3 released for multi-modal thinking.

Source: Timeline of artificial intelligence - Wikipedia

# Terminologies



# Human Performance as a Benchmark



Select AI Index technical performance benchmarks vs. human performance

Source: AI Index, 2025 | Chart: 2025 AI Index report

The machine is beating human performance in more and more tasks.

Source: [AI Index Report 2025 – Artificial Intelligence Index](), Stanford AI index

arXiv:2503.23674v1 [cs.CL] 31 Mar 2025

### Large Language Models Pass the Turing Test

**Cameron R. Jones**
Department of Cognitive Science
UC San Diego
San Diego, CA 92119
cameron@ucsd.edu

**Benjamin K. Bergen**
Department of Cognitive Science
UC San Diego
San Diego, CA 92119
bkbergen@ucsd.edu

**Abstract**

We evaluated 4 systems (ELIZA, GPT-4o, LLaMa-3.1-405B, and GPT-4.5) in two randomised, controlled, and pre-registered Turing tests on independent populations. Participants had 5 minute conversations simultaneously with another human participant and one of these systems before judging which conversational partner they thought was human. When prompted to adopt a humanlike persona, GPT-4.5 was judged to be the human 73% of the time: significantly more often than interrogators selected the real human participant. LLaMa-3.1, with the same prompt, was judged to be the human 56% of the time—not significantly more or less often than the humans they were being compared to—while baseline models (ELIZA and GPT-4o) achieved win rates significantly below chance (23% and 21% respectively). The results constitute the first empirical evidence that any artificial system passes a standard three-party Turing test. The results have implications for debates about what kind of intelligence is exhibited by Large Language Models (LLMs), and the social and economic impacts these systems are likely to have.

**1 Introduction**

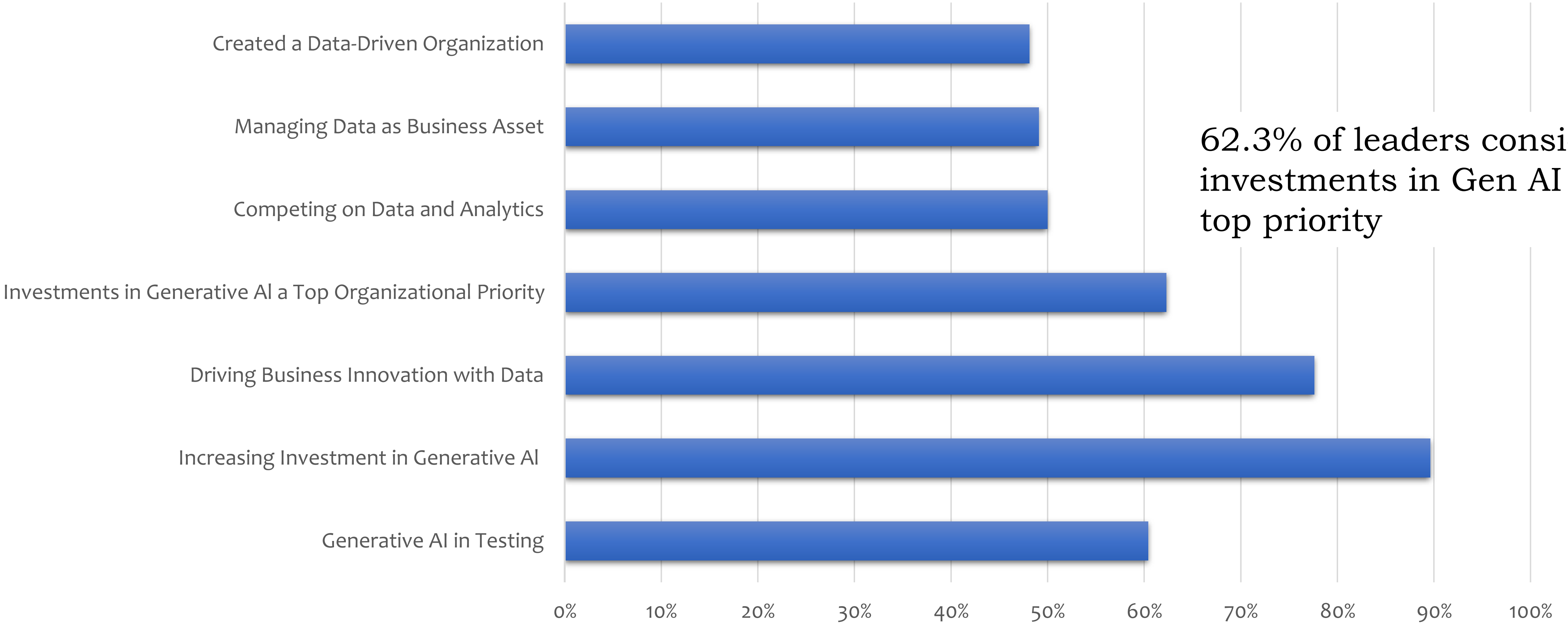
**1.1 The Turing test**

75 years ago, Alan Turing (1950) proposed the imitation game as a method of determining whether machines could be said to be intelligent. In the game—now widely known as the Turing test—a human interrogator speaks simultaneously to two witnesses (one human and one machine) via a text-only interface. Both witnesses attempt to persuade the interrogator that they are the real human. If the interrogator cannot reliably identify the human, the machine is said to have passed: an indication of its ability to imitate humanlike intelligence.

Turing's article "has unquestionably generated more commentary and controversy than any other article in the field of artificial intelligence" (French, 2000, p. 116). Turing originally proposed the test as a very general measure of intelligence, in that the machine would have to be able to imitate human behaviour on "almost any one of the fields of human endeavour" (Turing, 1950, p. 436) that are available in natural language. However, others have argued that the test might be too easy—because human judges are fallible (Gunderson, 1964; Hayes and Ford, 1995)—or too hard in that the machine must deceive while humans need only be honest (French, 2000; Saygin et al., 2000).

Source: [2503.23674] Large Language Models Pass the Turing Test

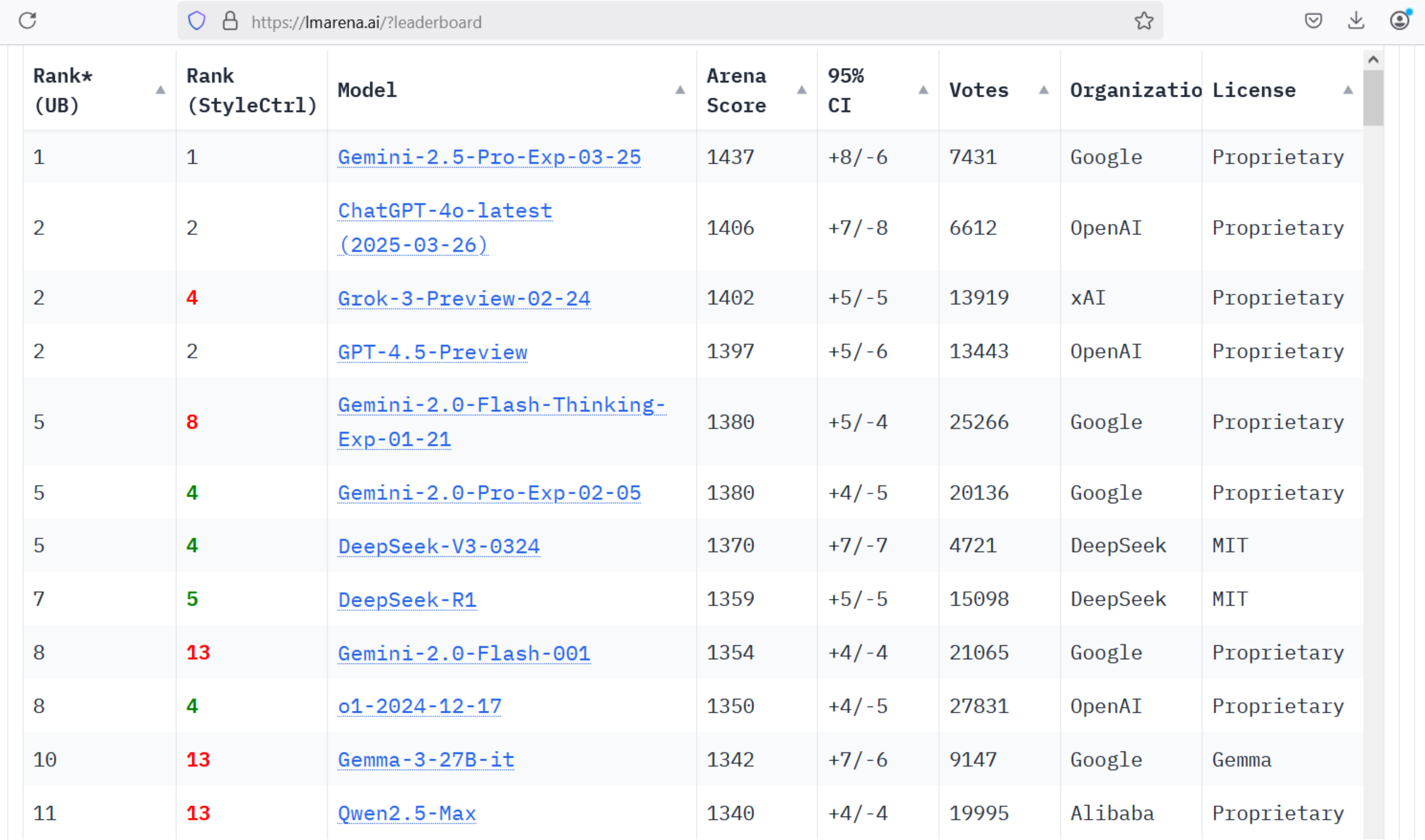
# Importance of Data Platform and Gen AI



62.3% of leaders consider investments in Gen AI as a top priority

Source: WaveStone 2024 DATA AND ANALYTICS LEADERSHIP ANNUAL EXECUTIVE SURVEY
DataAI-ExecutiveLeadershipSurveyFinalAsset.pdf (wavestone.com)

# Chatbot Arena Ranking (as of Apr 14, 2025)



https://lmarena.ai/?leaderboard

| Rank* (UB) | Rank (StyleCtrl) | Model | Arena Score | 95% CI | Votes | Organizatio | License |
|---|---|---|---|---|---|---|---|
| 1 | 1 | Gemini-2.5-Pro-Exp-03-25 | 1437 | +8/-6 | 7431 | Google | Proprietary |
| 2 | 2 | ChatGPT-4o-latest (2025-03-26) | 1406 | +7/-8 | 6612 | OpenAI | Proprietary |
| 2 | 4 | Grok-3-Preview-02-24 | 1402 | +5/-5 | 13919 | xAI | Proprietary |
| 2 | 2 | GPT-4.5-Preview | 1397 | +5/-6 | 13443 | OpenAI | Proprietary |
| 5 | 8 | Gemini-2.0-Flash-Thinking-Exp-01-21 | 1380 | +5/-4 | 25266 | Google | Proprietary |
| 5 | 4 | Gemini-2.0-Pro-Exp-02-05 | 1380 | +4/-5 | 20136 | Google | Proprietary |
| 5 | 4 | DeepSeek-V3-0324 | 1370 | +7/-7 | 4721 | DeepSeek | MIT |
| 7 | 5 | DeepSeek-R1 | 1359 | +5/-5 | 15098 | DeepSeek | MIT |
| 8 | 13 | Gemini-2.0-Flash-001 | 1354 | +4/-4 | 21065 | Google | Proprietary |
| 8 | 4 | o1-2024-12-17 | 1350 | +4/-5 | 27831 | OpenAI | Proprietary |
| 10 | 13 | Gemma-3-27B-it | 1342 | +7/-6 | 9147 | Google | Gemma |
| 11 | 13 | Qwen2.5-Max | 1340 | +4/-4 | 19995 | Alibaba | Proprietary |

- More varieties
- Beat common human performance in
  - Math/Coding
  - Painting
  - … more

Algorithms (e.g., Models, NN, Transformers, etc.)

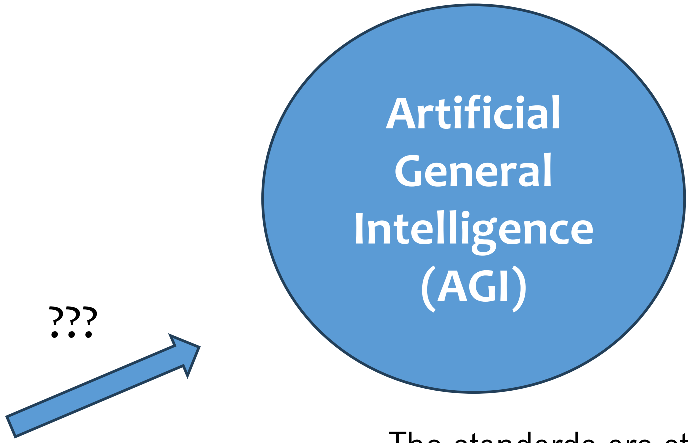
Core Enablers of AI

Data (e.g., Text, Video, Images, etc.

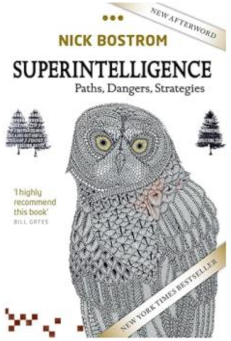
- Data widely exist on Internet and in enterprises
  - Document
  - Data lake
- Simulation and synthetic data
- Multi-modal

Computation (e.g., accelerators, GPUs, etc.)

- Faster GPUs every year
- More varieties of accelerators



The standards are still vague


NICK BOSTROM
SUPERINTELLIGENCE
Paths, Dangers, Strategies
NEW AFTERWORD
'I highly recommend this book' BILL GATES
NEW YORK TIMES BESTSELLER

Book by Nick Bostrom, 2014

# Training and Serving Pipeline

- Training Goal: Generate or finetune the model.
- Serving (aka Inferencing or deployment) Goal: Use the model to finish the task in hand.

**1. Data Collection:** Gather relevant and high-quality data to train your model or system.

**2. Data Preparation:** Clean, preprocess, and transform the data into a usable format.

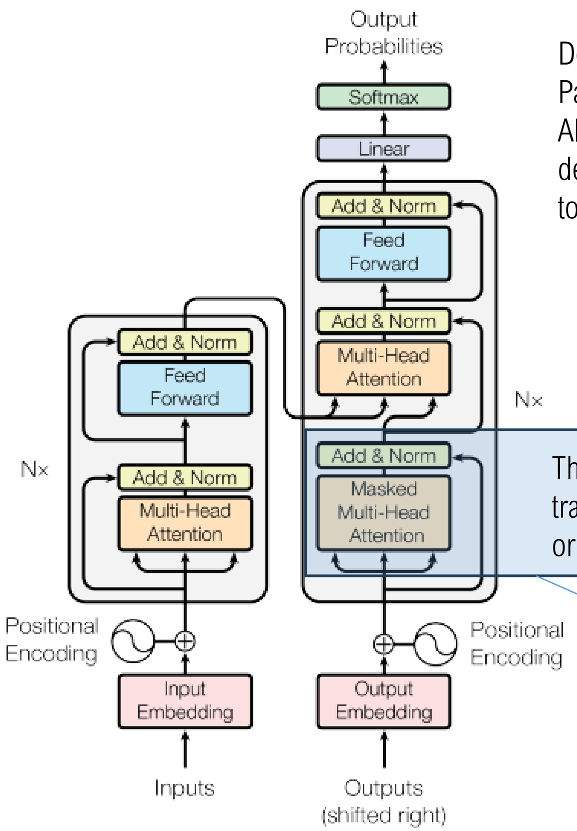
**3. Model Training:** Use the prepared data to train the model, optimizing it over iterations.

**4. Evaluation:** Test the model on validation data to measure performance and identify issues.

**5. Deployment:** Integrate the trained model into real-world applications or systems.

**6. Monitoring:** Continuously monitor the model's performance and update as needed.

# Transformer



Encoder Part

Decoder Part (most gen AI models are decoder only today)

This is called a transformer layer or block

Source: Vaswani et al. 2017 [1706.03762] Attention Is All You Need

- **Foundation for Pretrained Models**: Powers modern AI advancements in text, vision, and science.
  - The models for NLP tasks are called Large Language Models (LLM).
  - The models for vision tasks are called large vision models.
  - The models for a mixed range of tasks are called multi-modal models.
- When the scale of transformers is large (into the billions), the models show the capability of reasoning besides memorizing.
  - It is called emergent behavior.
  - The performance is better if the prompt is explaining the thinking steps. It is referred to as Chain of Thought or CoT (Wei et al. 2022, [2201.11903] Chain-of-Thought Prompting Elicits Reasoning in Large Language Models)
- Today many of the models can generate CoT during the inference time.

For example, the LLaMA-7b model has 32 transformer layers and it is decoder only. A larger model has more layers.

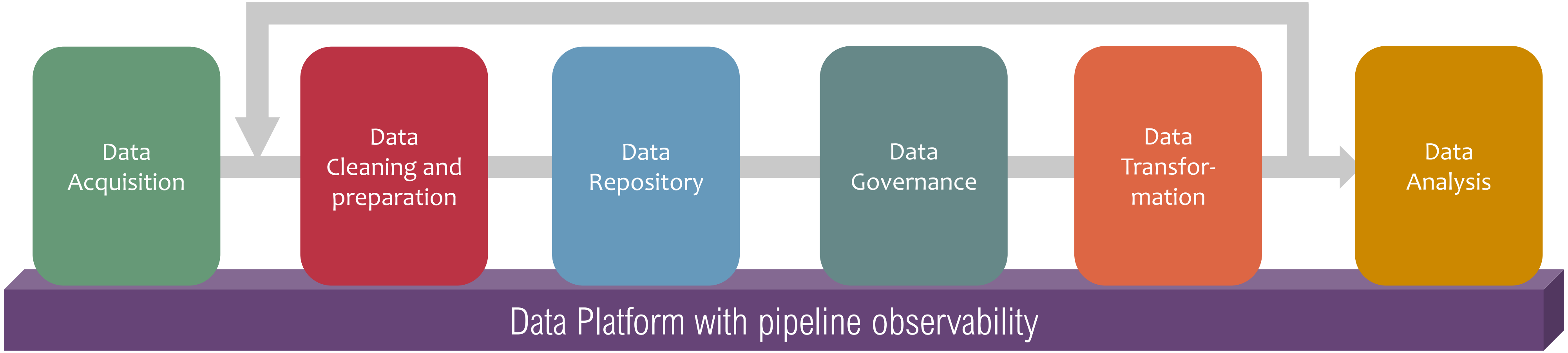
Larger models often have a better performance than smaller models today. For example, a 70b model likely has a better benchmark score than a 7b model.

# Data Needs Preparation

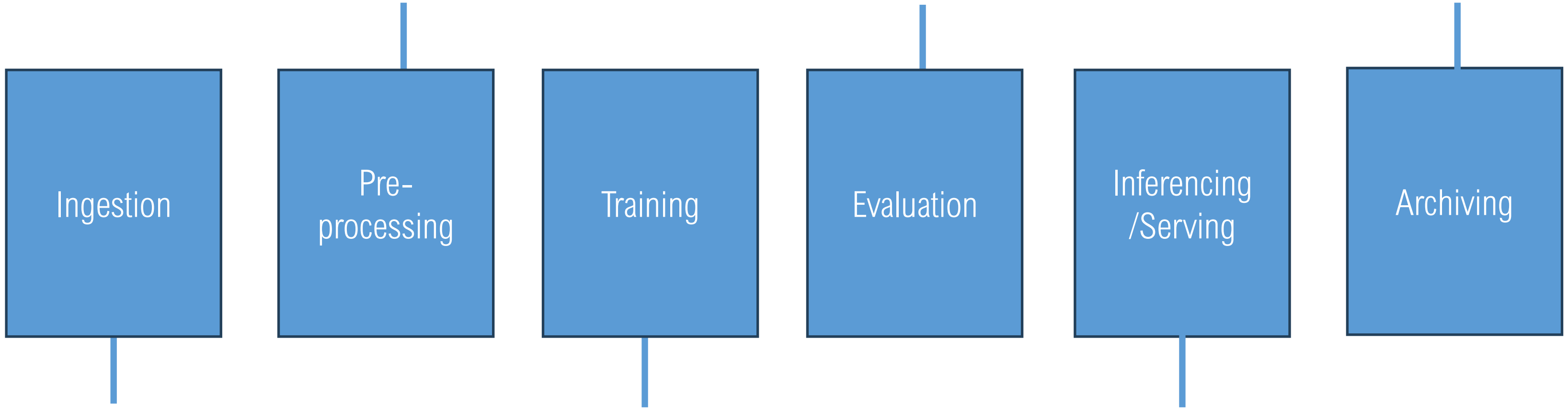
## Avoid “Garbage in, garbage out”

Data needs preparation to be used.

- Cleaning and possibly labeling
- Reformatting
- Refreshing knowledge



# Data Storage Needs



**Data Ingestion**

- Multi-protocol access
- High parallelization
- High bandwidth
- Large capacity

**Data Pre-Processing**

- Multi-protocol access
- High-speed search
- Mix of large and small I/O
- Mix of IO read and write

**Data Training**

- Large amount of read requests
- High bandwidth
- Low latency
- Checkpoints

**Model Evaluation**

- I/O Read Dominant
- High bandwidth
- Low latency

**Model Inferencing**

- High parallelization
- Low latency
- Model repository and deployment

**Data Archiving**

- Data lifecycle management
- Low cost
- Frequent access

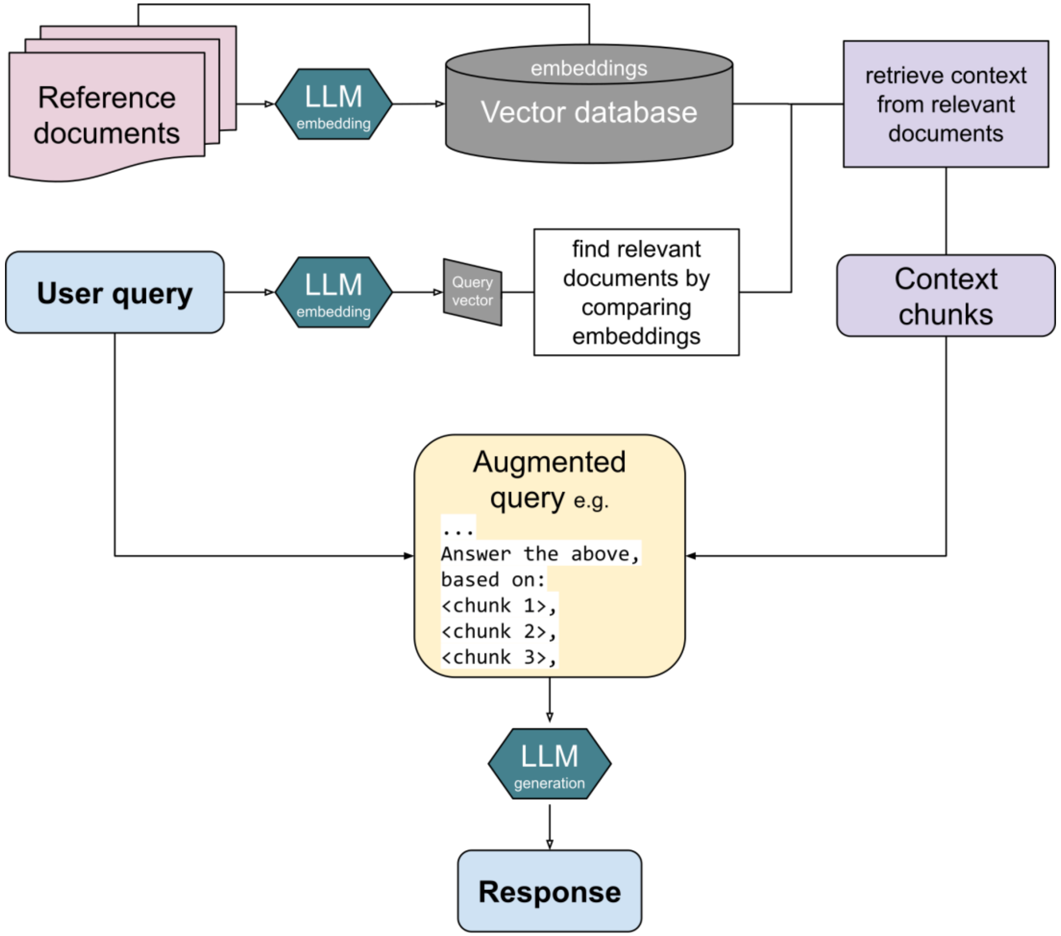
# Retrieval Augmented Generation (RAG)

RAG process:

- Alleviate the Hallucination problem introduced by LLM-based response.
- The **retriever** encodes user-provided prompts and relevant documents into vectors, stores them in a **vector database**, and retrieves relevant context vectors based on the distance between the **encoded** prompt and documents.
- The **generator** then combines the retrieved context with the original prompt to produce a response.

Advanced RAG:

- Added more steps and ways to increase the accuracy of obtaining information.
- For example, GraphRAG (Edge et al., 2024, [2404.16130] From Local to Global: A Graph RAG Approach to Query-Focused Summarization)



Source: Retrieval-augmented generation - Wikipedia

# Vector Database

- Simplify **data storage, organization, retrieval of complex data types**: images, likes, sounds, text files, pattern data, map data, genomic information, etc.
- An integral part of **machine learning** and for data in diverse domains, offer high performance and scalability.
- Handle **high-dimensional data** and perform rapid **similarity searches.**

Boosted by the wide use of RAG

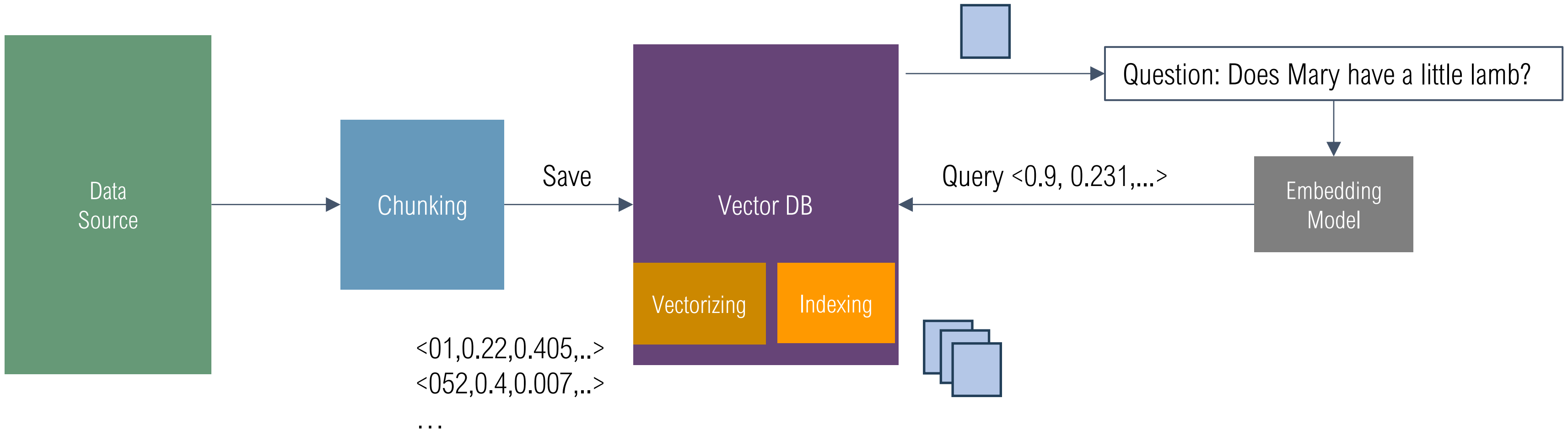
$3.04 Billion

2025 expectation

CAGR 23.7%

2024-2029 expectation

Source: The business research company, Vector Database Market Report 2025 - Vector Database Industry Analysis And Overview



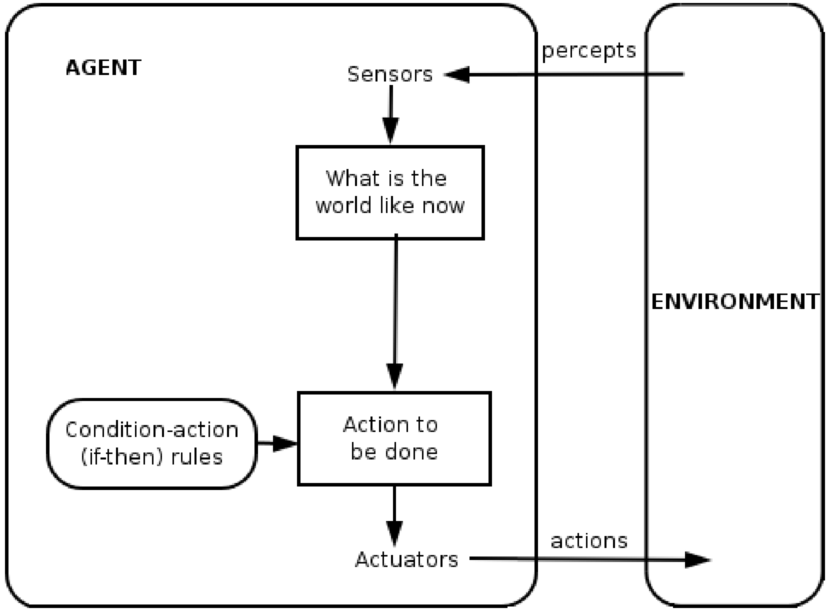
# Agentic Workflow

"I think AI agentic workflows will drive massive AI progress this year — perhaps even more than the next generation of foundation models."

-- Andrew Ng (2024 on X®)

## What about 2025?

- New paradigms of using models
- New tools developed



Source: Intelligent agent - Wikipedia

# Enterprise Readiness

| | |
|---|---|
| AI has been rapidly expended into production | => Enterprises need to be ready |
| Open-source models are ready | => On-prem deployment is ready for enterprises |
| Pretraining is converging, inferencing becomes more and more important | => Enterprises need to invest into the right infrastructure |
| RAG provides ways to increase accuracy, consistency, and ROI | => Enterprise need to build up advanced knowledge retrieval system |
| Agentic AI are developing, LLM is just part of the system | => Enterprises need system thinking and investment |

# Thank you!

- Comments
- Q&A