AI/ML Fundamentals: Introduction and Market Trends

Jan. 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 (2000- Jan 2025)

2000-2009: Foundations
- Roomba® launched as an autonomous vacuum (2002).
- **ImageNet** established as a benchmark for image recognition research.

2010-2019: Deep Learning Era
- AlexNet transformed image recognition.
- DeepMind®'s AlphaGo defeated world champions in Go
- AlphaFold® predicted protein structures

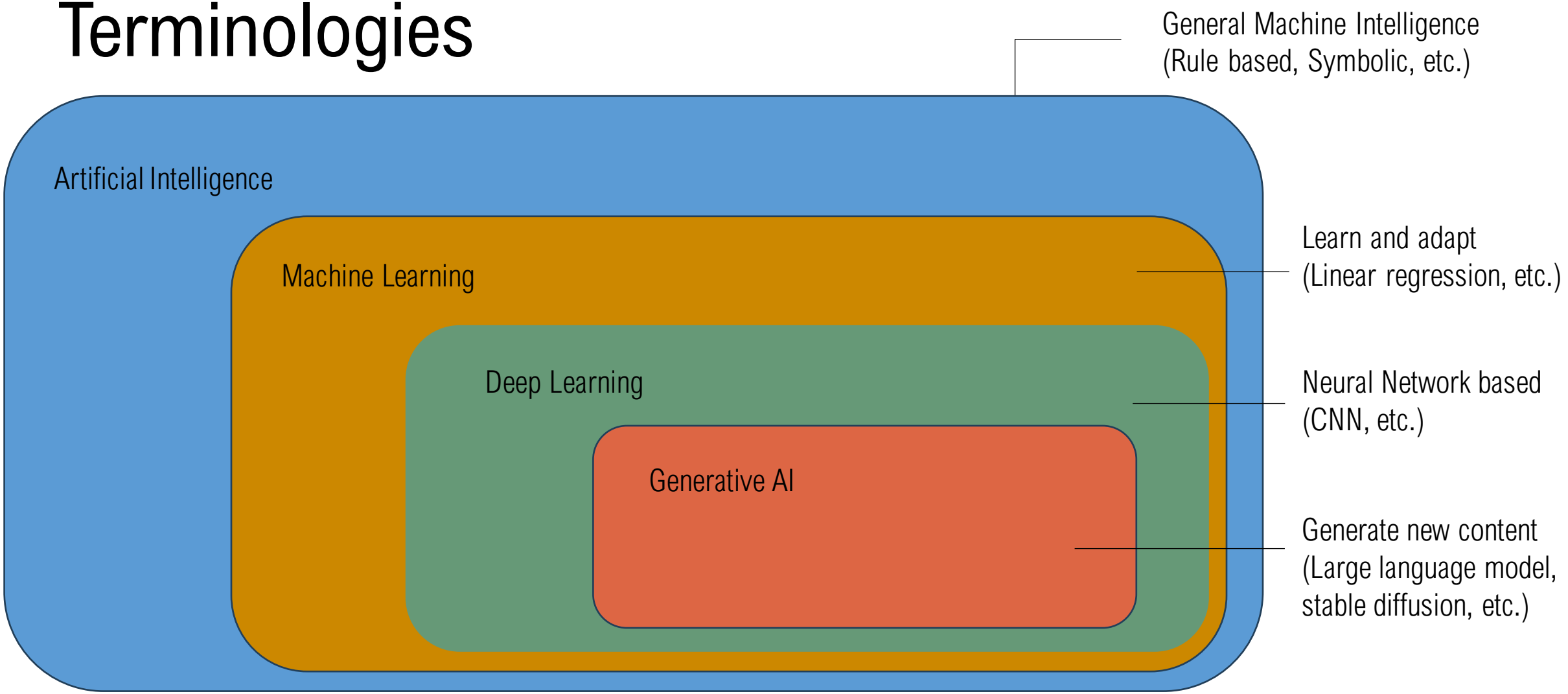
2020-2023: AI Revolution
- GPT-3® and ChatGPT® advanced natural language understanding.
- AlphaFold® 2 set new benchmarks in protein prediction.
- Governments and organizations began regulating AI with forums and safety summits.

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.

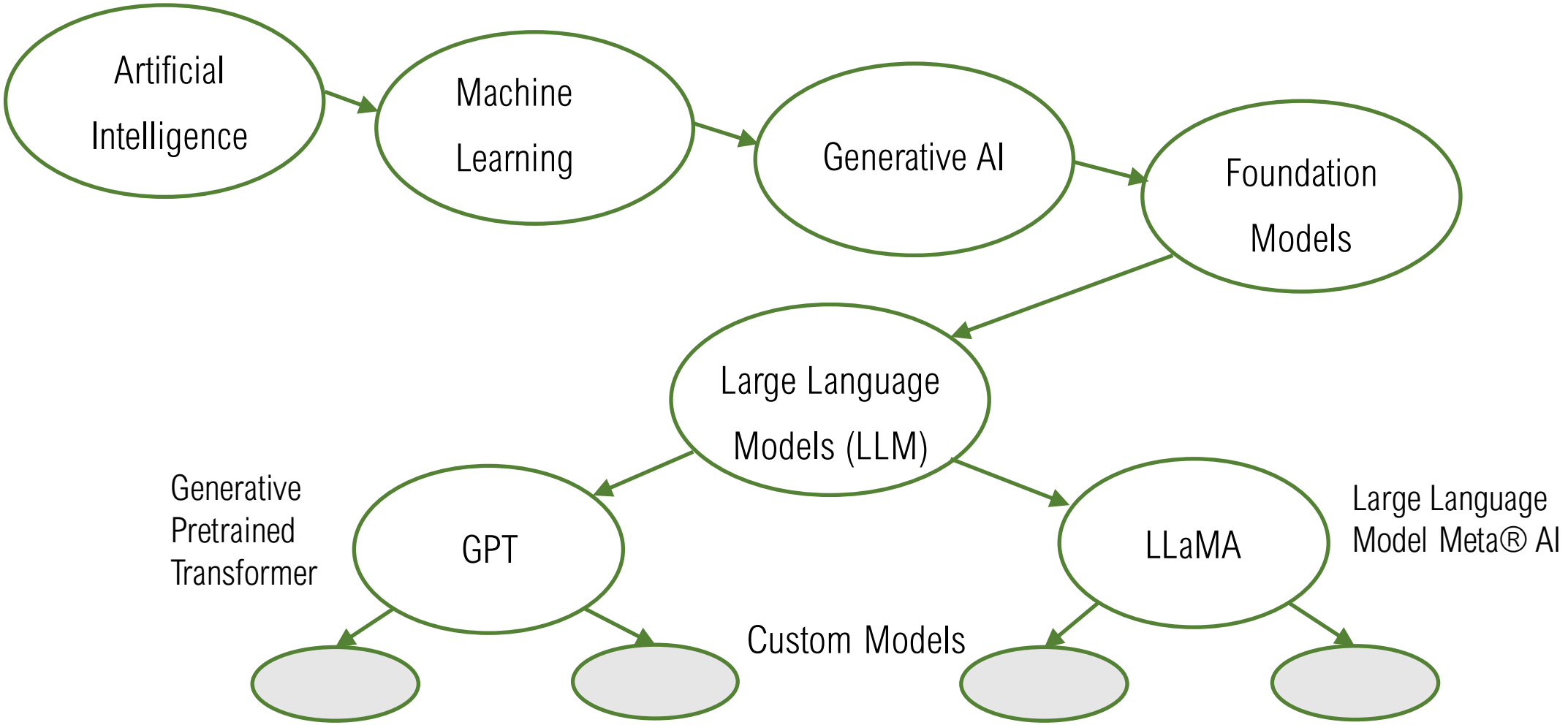
2025: Cost-efficient, Open-Source, and Beyond
- Deepseek® series with new RL process and low-cost training.

Source: Timeline of artificial intelligence - Wikipedia

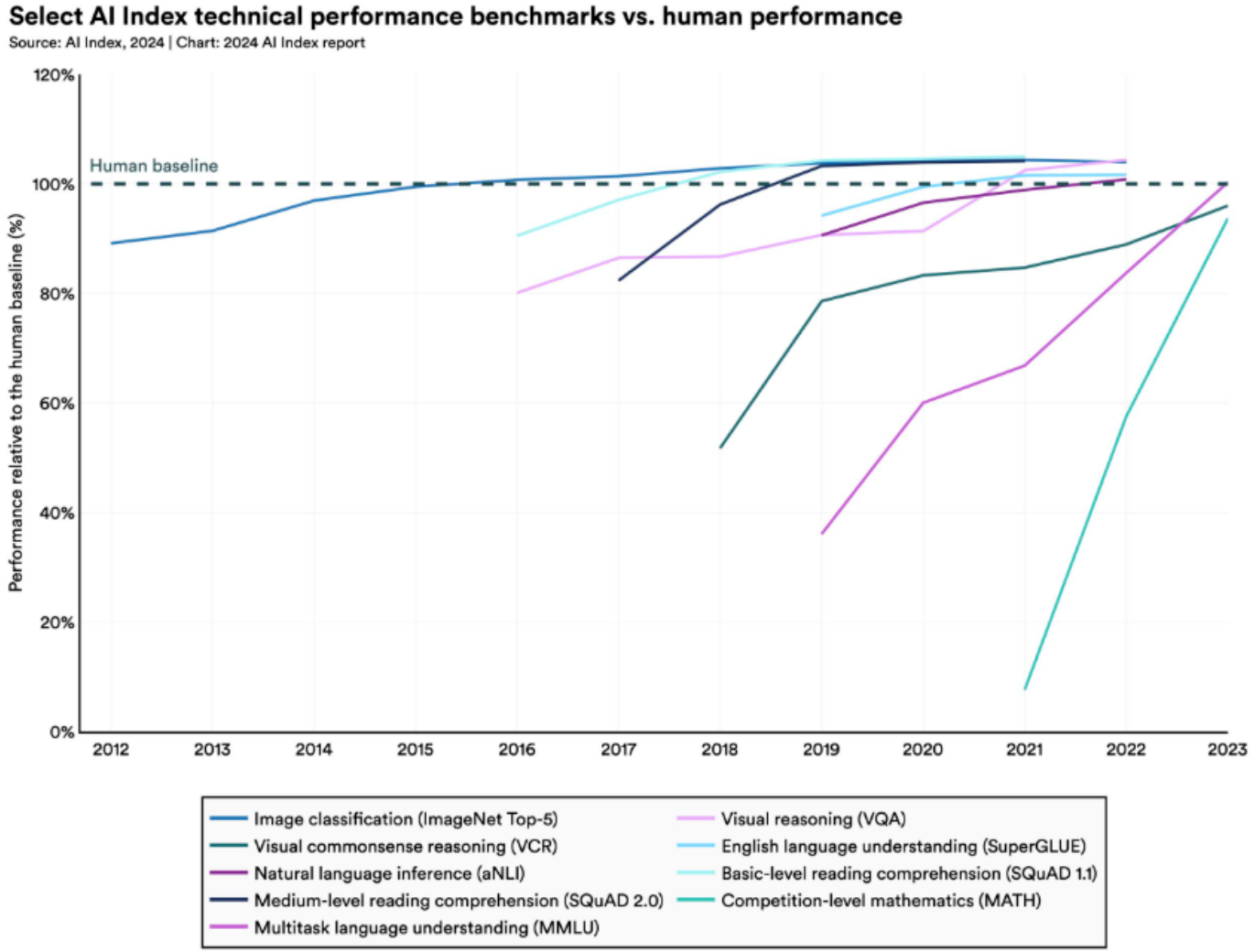
Terminologies



Terminologies



Human Performance as a Benchmark



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

Source: AI Index Report 2024 – Artificial Intelligence Index, Stanford AI index

Why Gen AI?

"Generative AI has the potential to change the world in ways that we can't even imagine. It has the power to create new ideas, products, and services that will make our lives easier, more productive, and more creative. It also has the potential to solve some of the world's biggest problems, such as climate change, poverty, and disease." – Bill Gates

Source: Forbes

2025 expectation

2025-2030 expectation

Source: Generative AI - Worldwide | Statista Market Forecast

Top 5 Challenges for IT Leaders in 2025

Unlock the Power of Data with New Applications

1

Data is the new energy of the economy. Whoever can first drill value out of the data field will gain advantage in business competition.

Build Intelligence Using ML and Generative AI

2

In 2025, ML and gen AI become the priority in enterprises worldwide. The power of ML may greatly affect the world and businesses this year through providing new insights.

Enforce Data Governance and Data Quality

3

Data governance not only ensures the compliance but also need to be combined with data quality control to ensure the information is accurate, on-time, and complete.

Ensure Data Resilience and Security

4

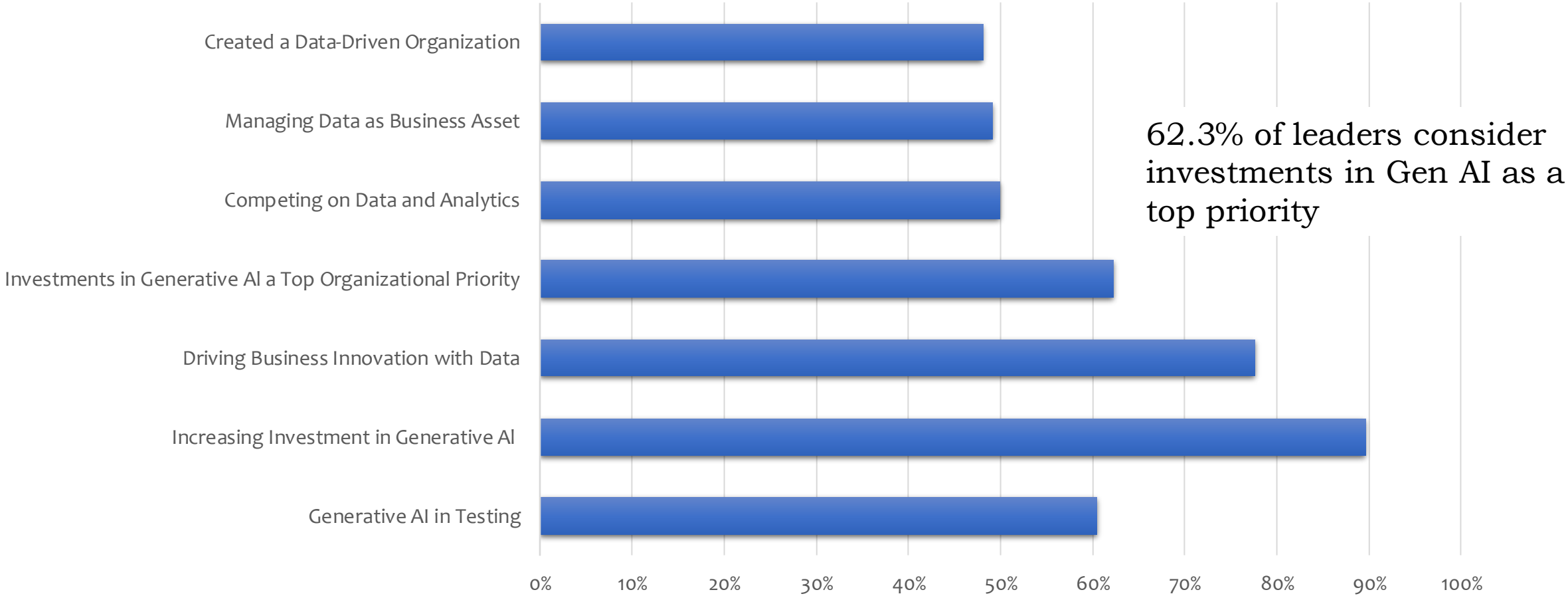
Data resilience and security will be facing new challenges in the new year. Ransomwares and disasters are in the trend of increasing.

Improve Cost Efficiency

5

The amount of data is growing in a rapid pace. We need innovations in infrastructure to control the budget.

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)

Model Arena Ranking (as of Jan 27, 2025)

https://lmarena.ai/?leaderboard

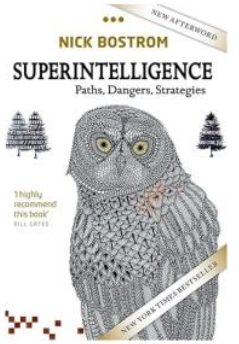
Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License
1	3	Gemini-2.0-Flash-Thinking-Exp-01-21	1382	+8/-6	6437	Google	Proprietary
1	1	Gemini-Exp-1206	1374	+5/-4	22116	Google	Proprietary
3	1	ChatGPT-4o-latest (2024-11-20)	1365	+4/-4	35328	OpenAI	Proprietary
3	1	DeepSeek-R1	1357	+12/-13	1883	DeepSeek	MIT
4	5	Gemini-2.0-Flash-Exp	1356	+4/-4	20939	Google	Proprietary
4	1	o1-2024-12-17	1352	+6/-6	9230	OpenAI	Proprietary
7	4	o1-preview	1335	+3/-3	33186	OpenAI	Proprietary
8	9	DeepSeek-V3	1317	+6/-5	13640	DeepSeek	DeepSeek
8	11	Step-2-16K-Exp	1305	+9/-7	4533	StepFun	Proprietary
9	12	o1-mini	1305	+2/-3	49952	OpenAI	Proprietary
9	9	Gemini-1.5-Pro-002	1302	+3/-4	46621	Google	Proprietary
12	14	Grok-2-08-13	1288	+3/-3	67150	xAI	Proprietary
12	17	Yi-Lightning	1287	+3/-4	28955	01 AI	Proprietary
12	10	GPT-4o-2024-05-13	1285	+2/-2	117745	OpenAI	Proprietary

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

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

???

Artificial General Intelligence (AGI)



Book by Nick Bostrom, 2014

The standards are still vague

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

AI Ethics

- **Fairness:**
  - Make AI treat everyone equally.
- **Transparency:**
  - Ensure people understand how AI works.
- **Responsibility:**
  - Hold someone accountable for AI decisions.
- **Privacy:**
  - Protect personal data and respect user rights.
- **Human alignment:**
  - Ensure AI follows human values.

Regulations are set by different government and organizations worldwide..

The word AI may mean different things in different places and context.

Regulations are used to protect us from current and future threats, including superintelligence.

Ethics are evolving with many debates.

AI Watch: Global regulatory tracker | White & Case LLP

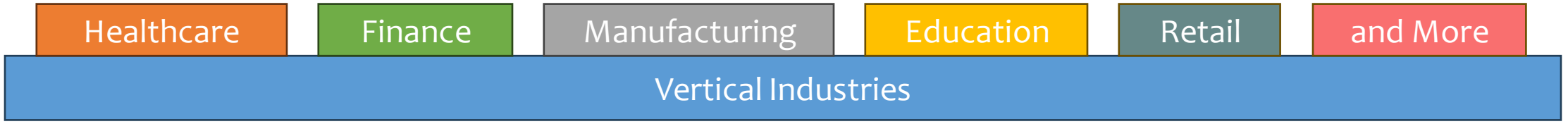
Task Examples

Traditional task examples:

- **Classification**: Assign labels (e.g., spam detection).
- **Regression**: Predict continuous values (e.g., price prediction).
- **Clustering**: Group similar data (e.g., customer segmentation).
- **Anomaly Detection**: Identify unusual patterns (e.g., fraud detection).
- **Recommendation Systems**: Suggest items (e.g., Music recommendations).
- **Time Series Forecasting**: Predict sequential data (e.g., sales forecasting).
- **Computer Vision**: Tasks like object detection or face recognition.

LLM-based task examples:

- **Text Generation**: Create coherent, context-aware text.
- **Question Answering (QA)**: Provide accurate answers to queries.
- **Conversational AI**: Enable human-like dialogue (e.g., chatbots) for customer service.
- **Code Generation**: Write or debug code.
- **Sentiment Analysis**: Extract sentiment from nuanced text.
- **Multimodal Tasks**: Process text, images, or audio together.
- **Creative Applications**: Assist with storytelling and content creation.
- **Personalization**: Adapt solutions to user preferences.



Training

Supervised Learning

- Trains on labeled data to map inputs to outputs.
- Used for classification (e.g., spam detection) and regression (e.g., price prediction).
- Minimizes prediction errors using a loss function.
- Common algorithms: linear regression, decision trees, neural networks.

Unsupervised Learning

- Works with unlabeled data to identify patterns or structures.
- Used for clustering, anomaly detection, and dimensionality reduction.
- Algorithms: k-means, hierarchical clustering, PCA.
- Evaluated using domain expertise or indirect metrics.

Semi-supervised Learning

- Combines small labeled data with large unlabeled data.
- Enhances generalization by leveraging both data types.
- Applied in NLP and fields where labeled data is costly.
- Methods: self-training and iterative labeling.

Reinforcement Learning

- Agents learn by interacting with an environment and receiving rewards/penalties.
- Focuses on optimizing cumulative rewards.
- Applied in games, robotics, and autonomous systems.
- Key algorithms: Q-learning, DQN, policy gradients.

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.

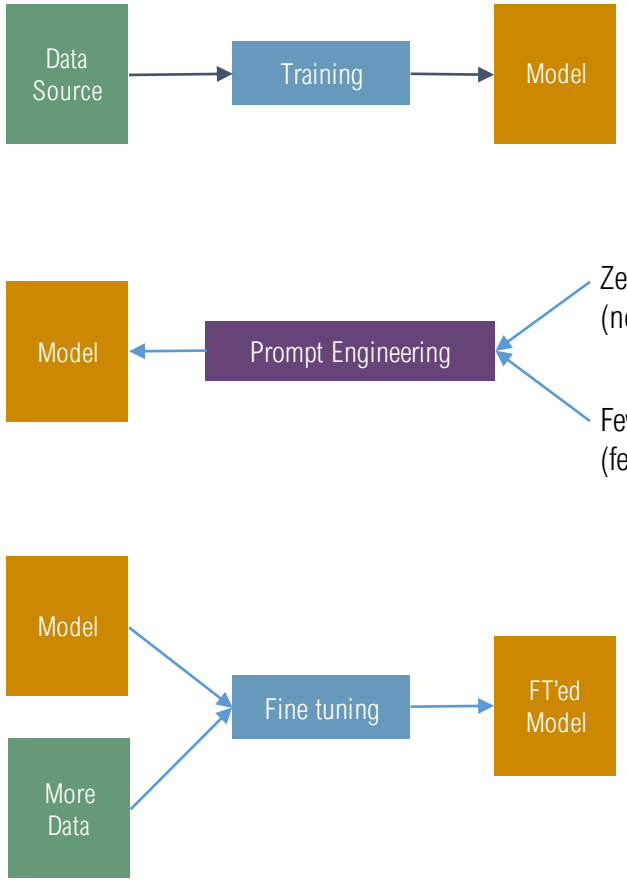
4. Evaluation: Test the model on validation data to measure performance and identify issues.

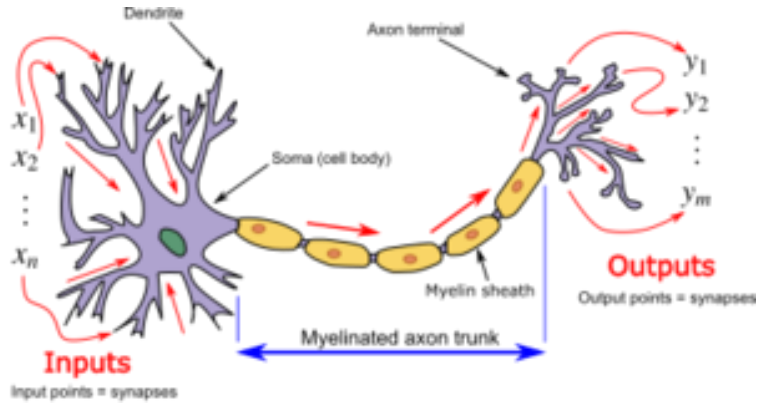
5. Deployment: Integrate the trained model into real-world applications or systems.

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

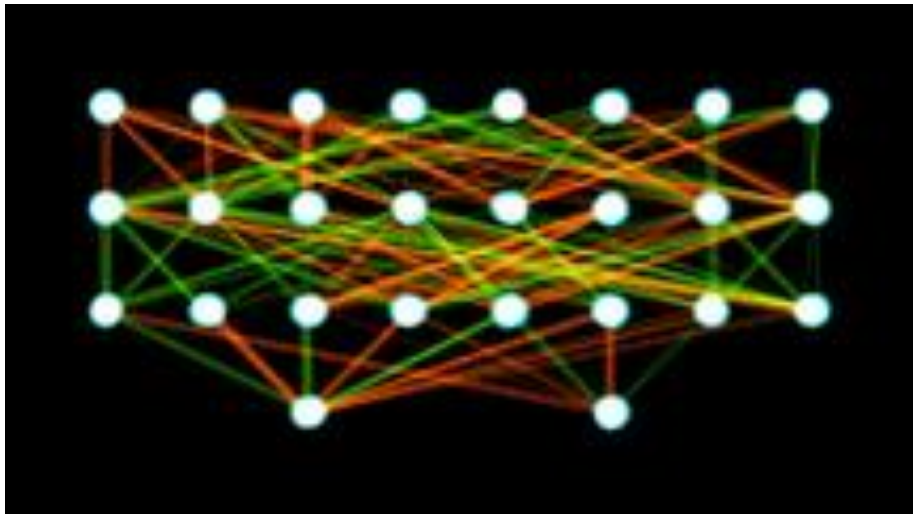
Transfer learning

- Utilizes knowledge from a pre-trained model to solve a new but related task.
- Reduces training time and improves performance with limited data.
- Common in computer vision (e.g., using ImageNet models) and NLP (e.g., BERT, GPT).
- Fine-tuning or feature extraction adapts the pre-trained model to the new task.
- Effective when tasks share similarities, such as domain or data structure.





a. Real neurons and neural network

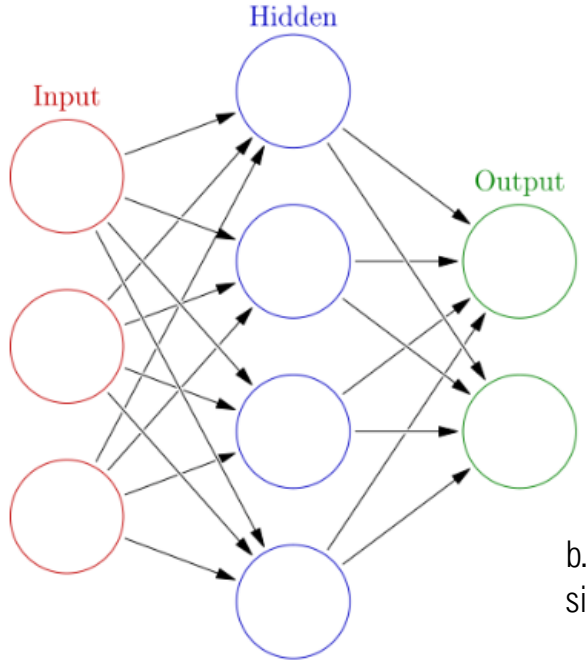


8 Neurons as input

2x8 Neurons (aka nodes) in the 2 hidden layers

2 Neurons as output

c. A 2-layer feedforward ANN



3 Neurons as input

2 Neurons as output

4 Neurons (aka nodes) in the hidden layer

b. Artificial Neural Network, a single layer ANN

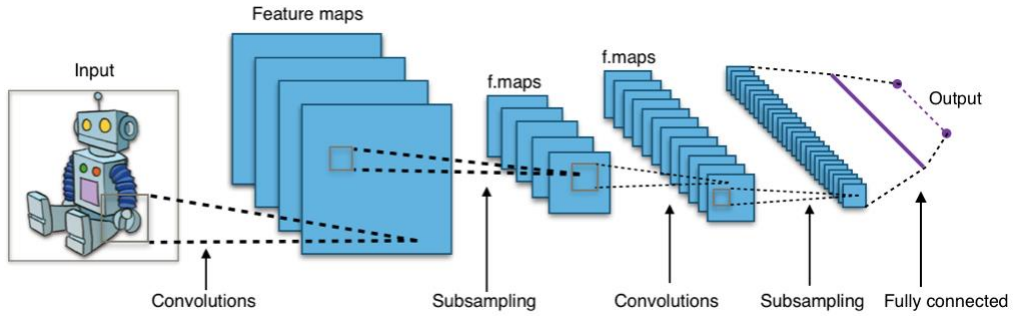
Source: [Neural network (machine learning) - Wikipedia]

Neural Network

Simulate real biological neural networks with math

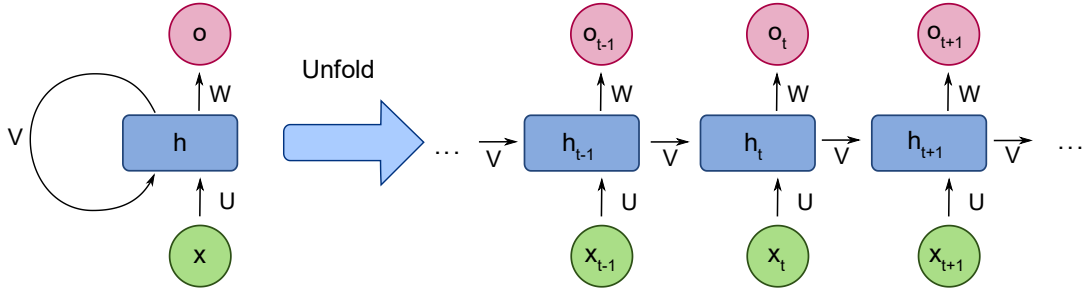
NN Architectures

- **Standard artificial neural network (ANN)**: Standard deep ANN are neural networks with multiple hidden layers with interconnected neurons, that communicate through synapses.
- **Convolutional neural network (CNN)**: CNNs are a type of neural network that are well-suited for image classification tasks. CNNs are composed of a series of convolutional layers, which extract features from images, and pooling layers, which down sample the feature maps.
- **Recurrent neural network (RNN)**: RNNs are deep neural networks that has the ability to store information from previous computations and passes it forward so as to work upon this data in a sequential manner. LSTM (Long Short-term Memory) is a sub-category of RNN
- **Transformer and beyond** (see next section)



CNN mimics how humans recognize patterns, like identifying shapes and objects in pictures, by breaking the image into small pieces and analyzing them. Convolutions are filters capturing features such as an edges, corners, or textures.

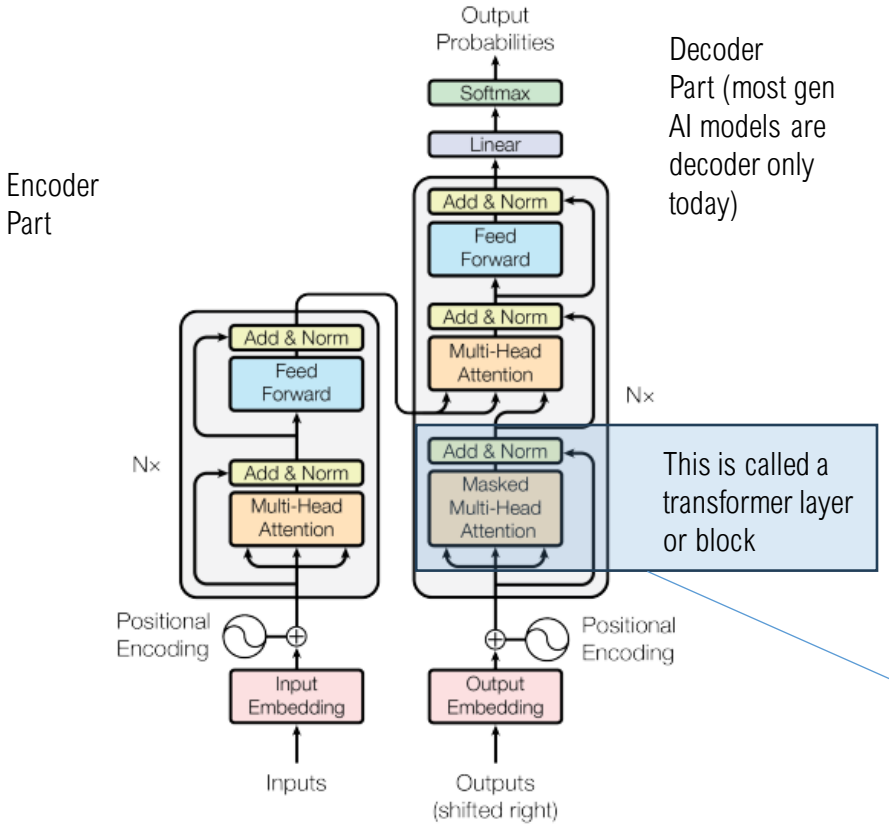
Source: Convolutional neural network - Wikipedia



RNN was designed to process a sequence of data. At each step, RNN combines the new input with the previous state, which is often referred to as "memory".

Source: Recurrent neural network - Wikipedia

Transformer



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.

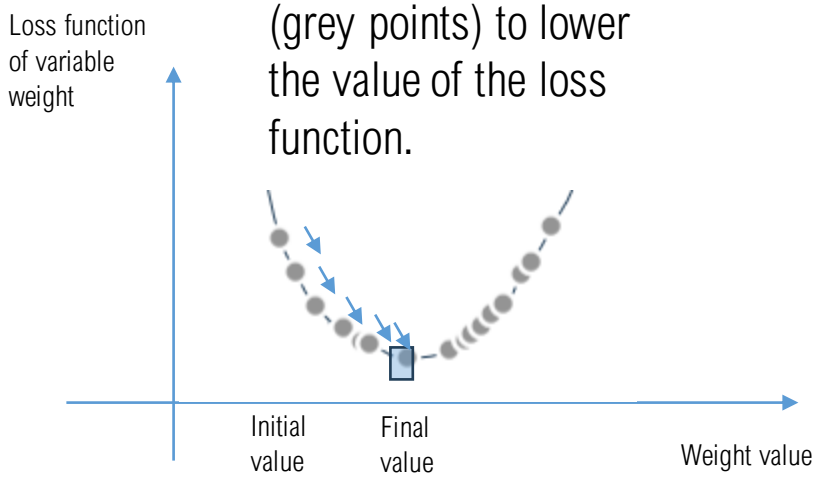
Learning Process

- **Loss Function**

A loss function measures the difference between predicted and actual values. The goal is to minimize this difference, improving model accuracy.

- **Gradient Descent**

Gradient Descent is an optimization algorithm used to minimize the loss function. It works by iteratively adjusting model parameters based on the gradients, with the step size controlled by the learning rate.



Gradient descent will try different values (grey points) to lower the value of the loss function.

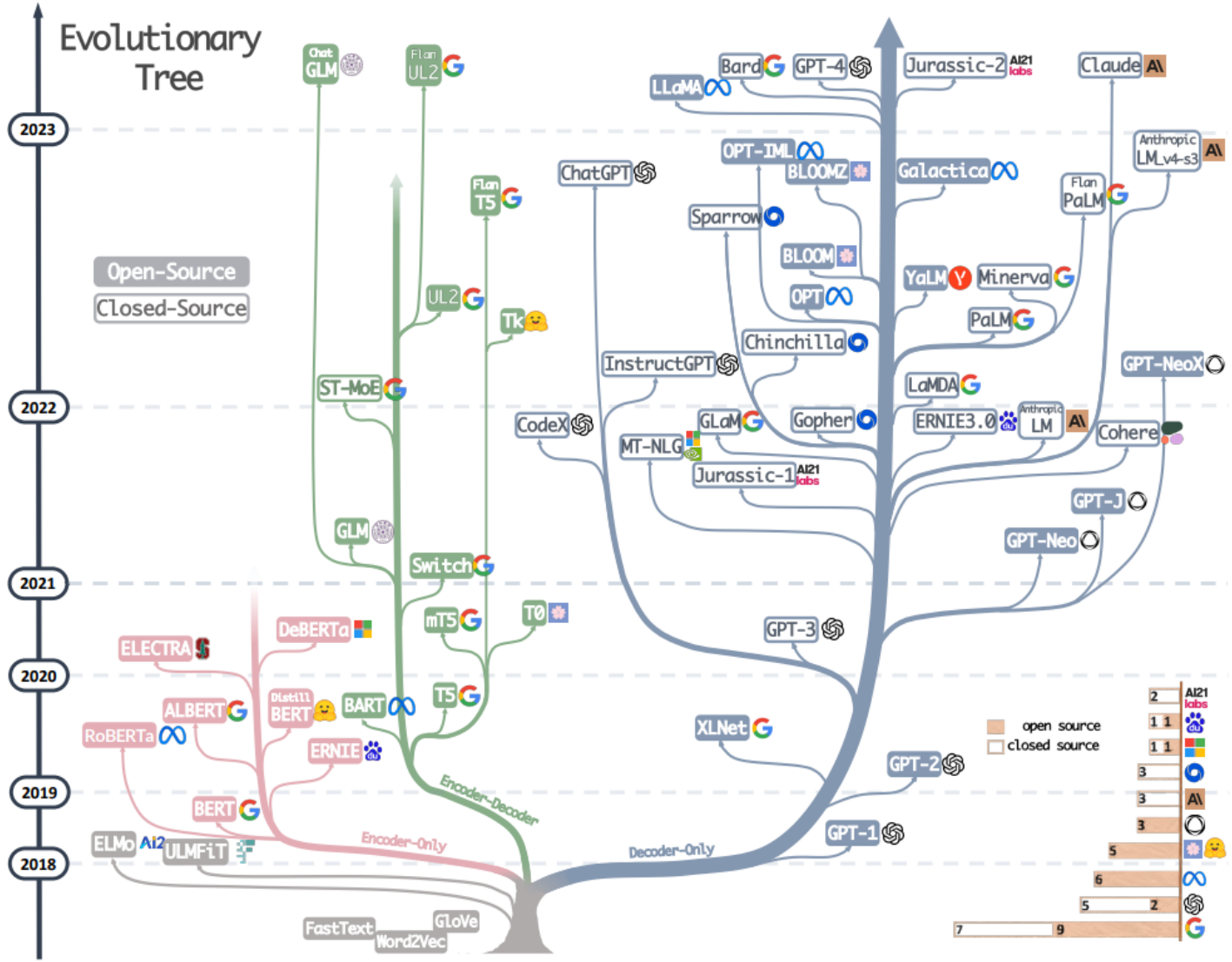
Goal: Find the weight point to minimize the loss function.

- **Backpropagation**

An algorithm that calculates the error between the network's predicted output and the actual target, then propagates this error backwards through the network, adjusting the weights and biases of each neuron proportionally to minimize the error.

Source: Backpropagation - Wikipedia

LLM Evolution History



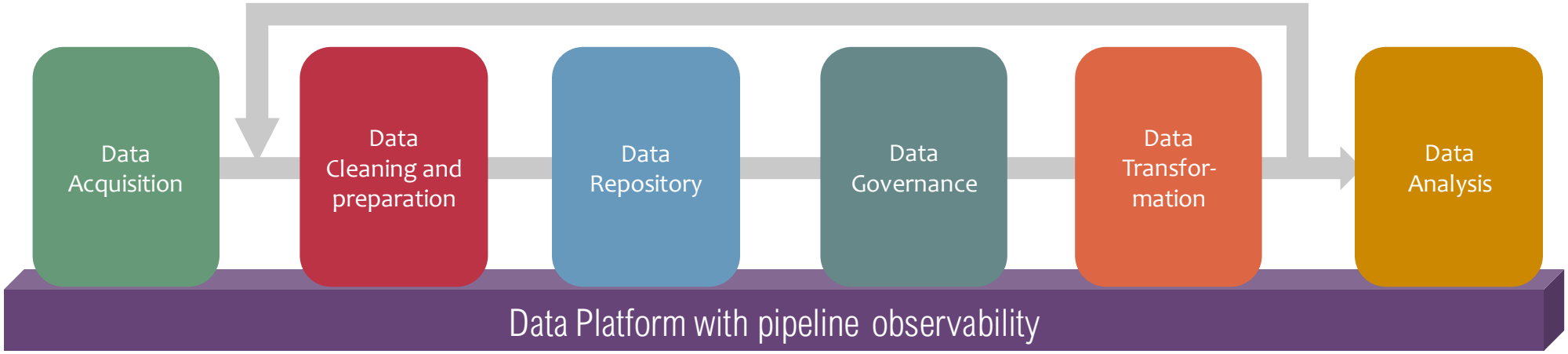
Source: Yang et al., 2023, [2304.13712] Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond

Data Needs Preparation

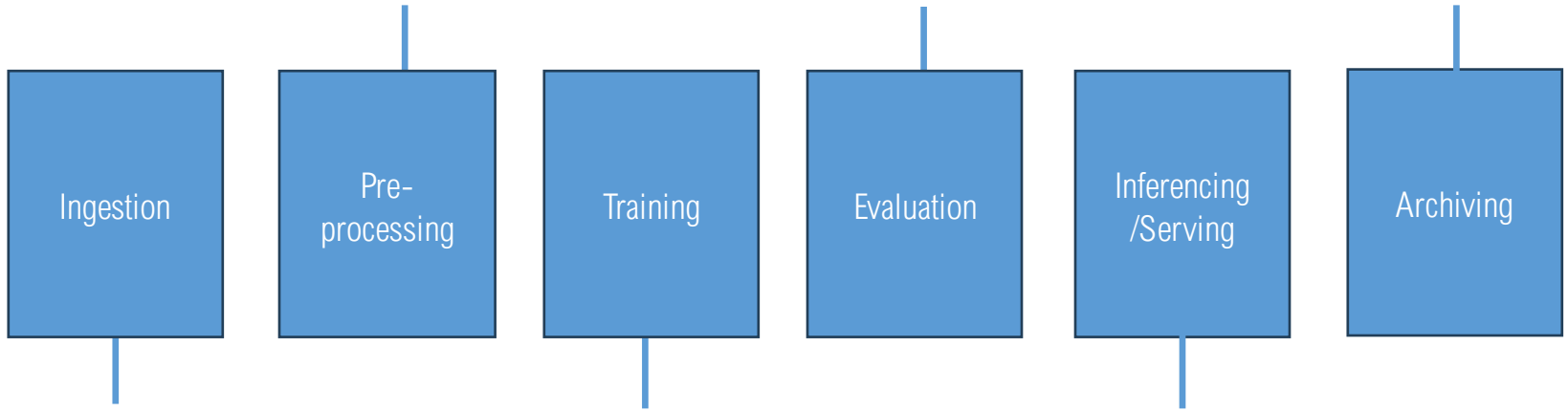
"Garbage in, garbage out"

Data needs preparation to be used.

- Cleaning and possibly labeling
- Reformatting
- Refreshing knowledge



Data Storage Needs



Data Pre-Processing

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

Model Evaluation

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

Data Archiving

- Data lifecycle management
- Low cost
- Frequent access

Data Ingestion

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

Data Training

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

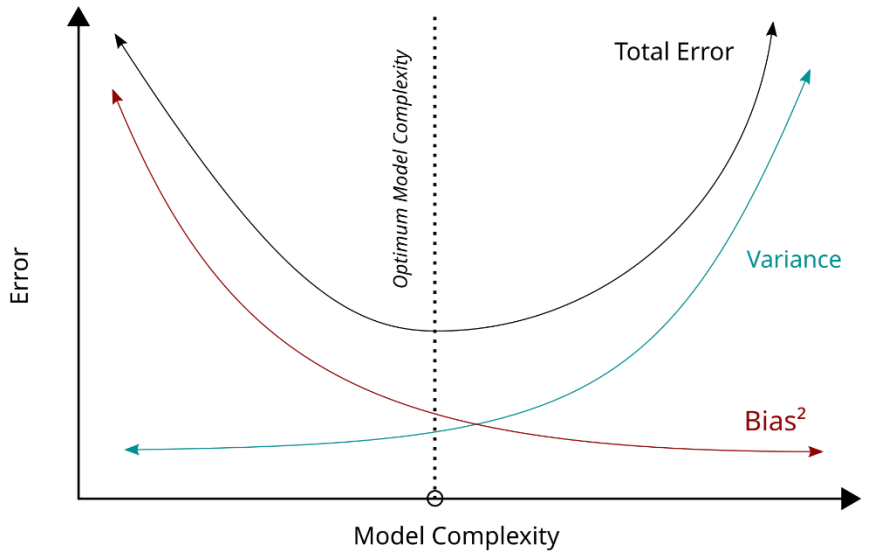
Model Inferencing

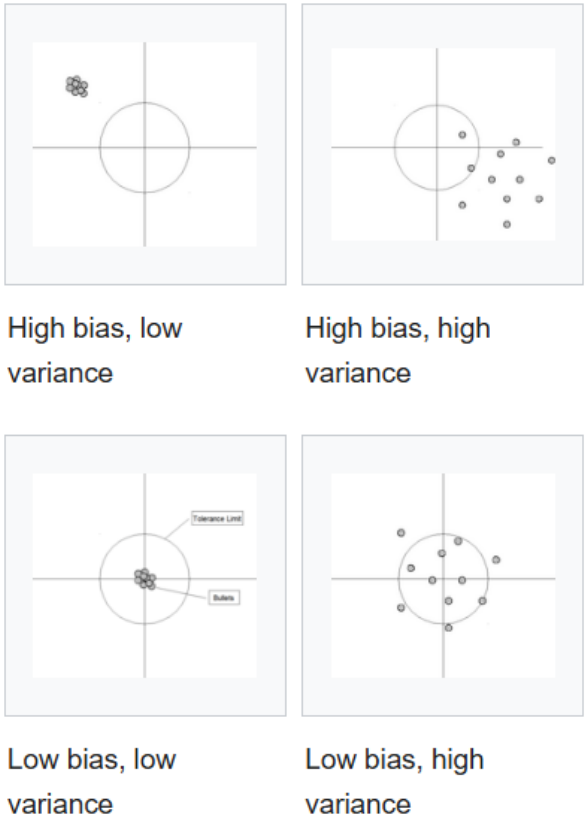
- High parallelization
- Low latency
- Model repository and deployment

Bias-Variance Tradeoff

- Train-test split, cross-validation
  - Don't allow test data leak into the training data set
- Bias: Overfitting vs underfitting
- "The bias–variance tradeoff is a central problem in supervised learning."

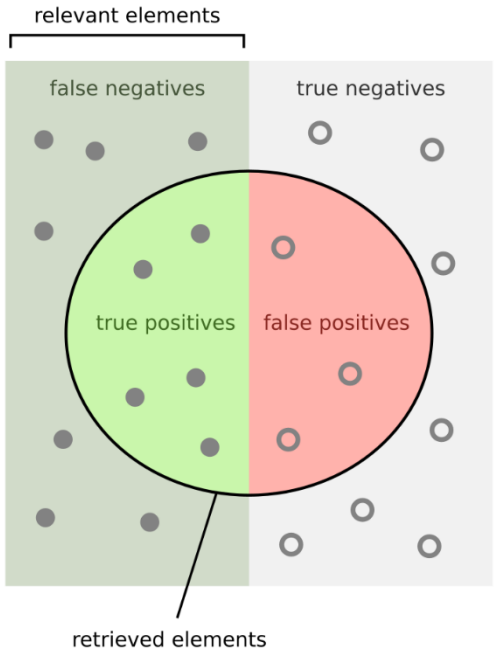
Source: Bias–variance tradeoff - Wikipedia





Performance Metrics

Performance metrics: quantify the quality of prediction made by ML models trained via data



Source: Precision and recall - Wikipedia

Regression (e.g., Price Prediction, Stock Forecasting)

- **MAE**: Average absolute error.
- **MSE**: Penalizes larger errors more.
- **RMSE**: Like MSE, but easier to interpret.
- **R^2**: How well the model explains the data.

Classification (e.g., Spam Detection, Image Recognition)

- **Accuracy**: How many predictions are correct.
- **Precision**: How many predicted positives are actually correct.
- **Recall**: How many actual positives were found.
- **F1-Score**: A balance between precision and recall.

NLP (e.g., Chatbots, Translation)

- **BLEU**: Checks how close machine translations are to human ones.
- **ROUGE**: Measures how well AI-generated summaries match real summaries.
- **Perplexity**: Evaluates how well a model predicts text.

Gen AI dataset as an example

Each model has its domain knowledge.
The percentage of mix is unique.

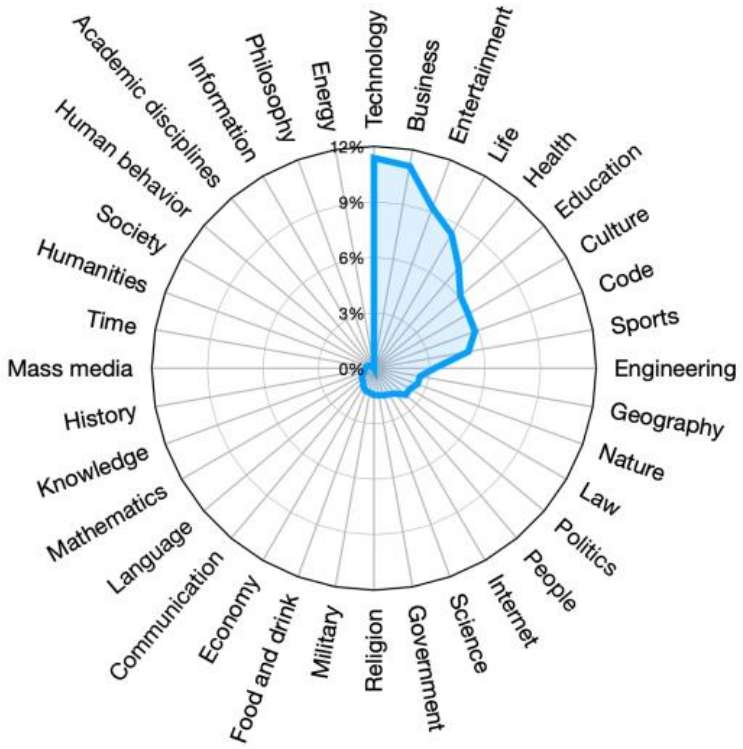


Figure 1: The distribution of different categories of Baichuan 2 training data.

Data needs to be pre-processed.
Only partial data is used for training

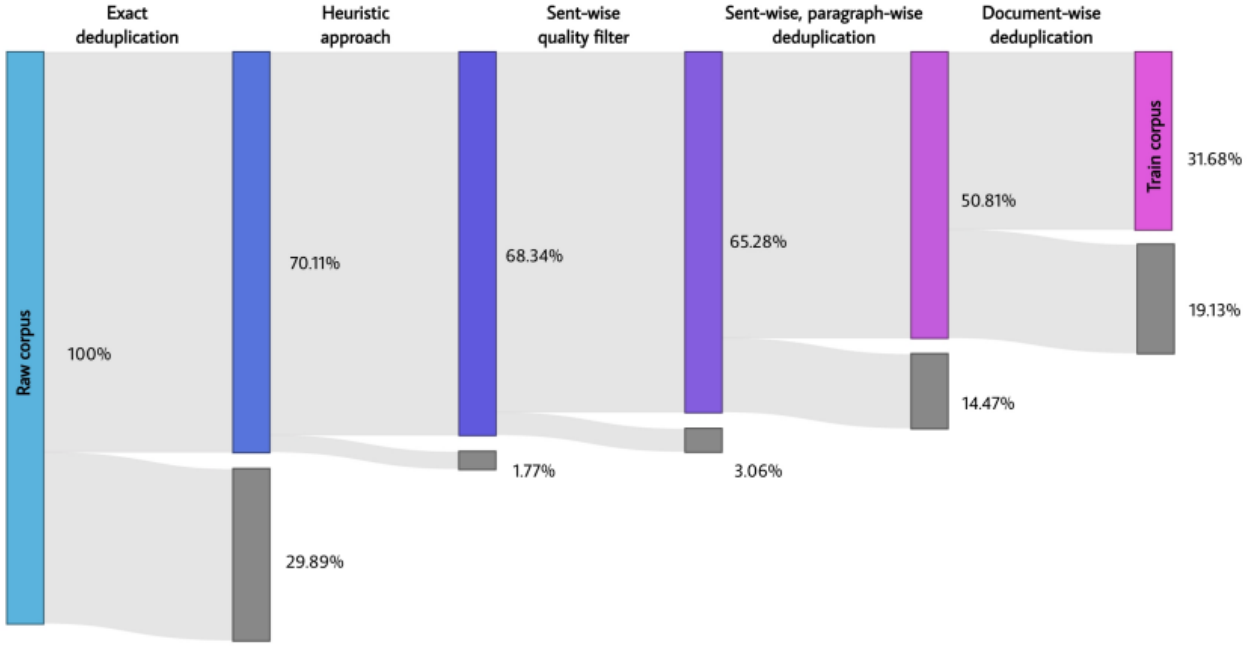


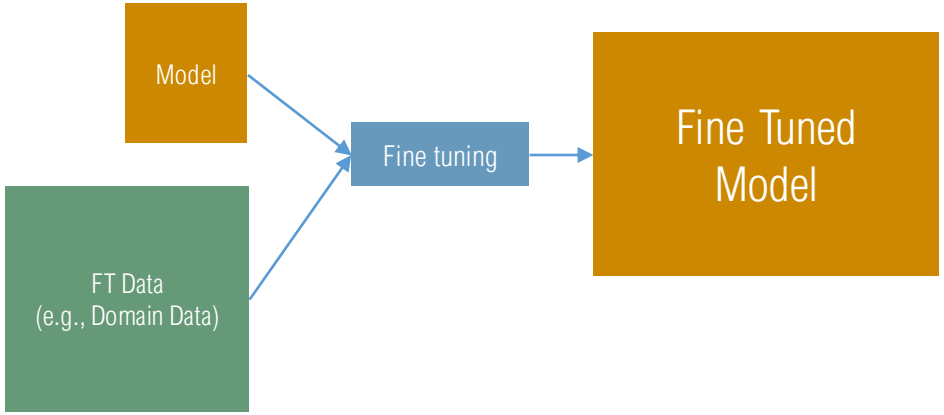
Figure 2: The data processing procedure of Baichuan 2's pre-training data.

Source: Yang et al., 2023, [2309.10305] Baichuan 2: Open Large-scale Language Models

Data for Fine Tuning

Why Fine-Tuning Data?

- Helps adapt pre-trained AI models to specific tasks or domains.
- Improves model accuracy and relevance for specialized applications.
- Supervised Fine Tuning (SFT) Reduces training time compared to training from scratch.



Characteristics of Good Fine-Tuning Data

- **High-Quality**: Clean, well-labeled, and free of errors.
- **Domain-Specific**: Matches the target application (e.g., medical, legal, finance).
- **Diverse**: Covers a range of variations within the domain.
- **Balanced**: Avoids bias and ensures fair representation.

Example Sources of Fine-Tuning Data

- **Existing Labeled Datasets**: Open-source datasets (e.g., ImageNet, Wikipedia).
- **Domain-Specific Data**: Medical records, legal documents, industry reports.
- **Synthetic Data**: AI-generated data for filling gaps in real data.
- **User Feedback**: Collecting real-world corrections and adjustments.

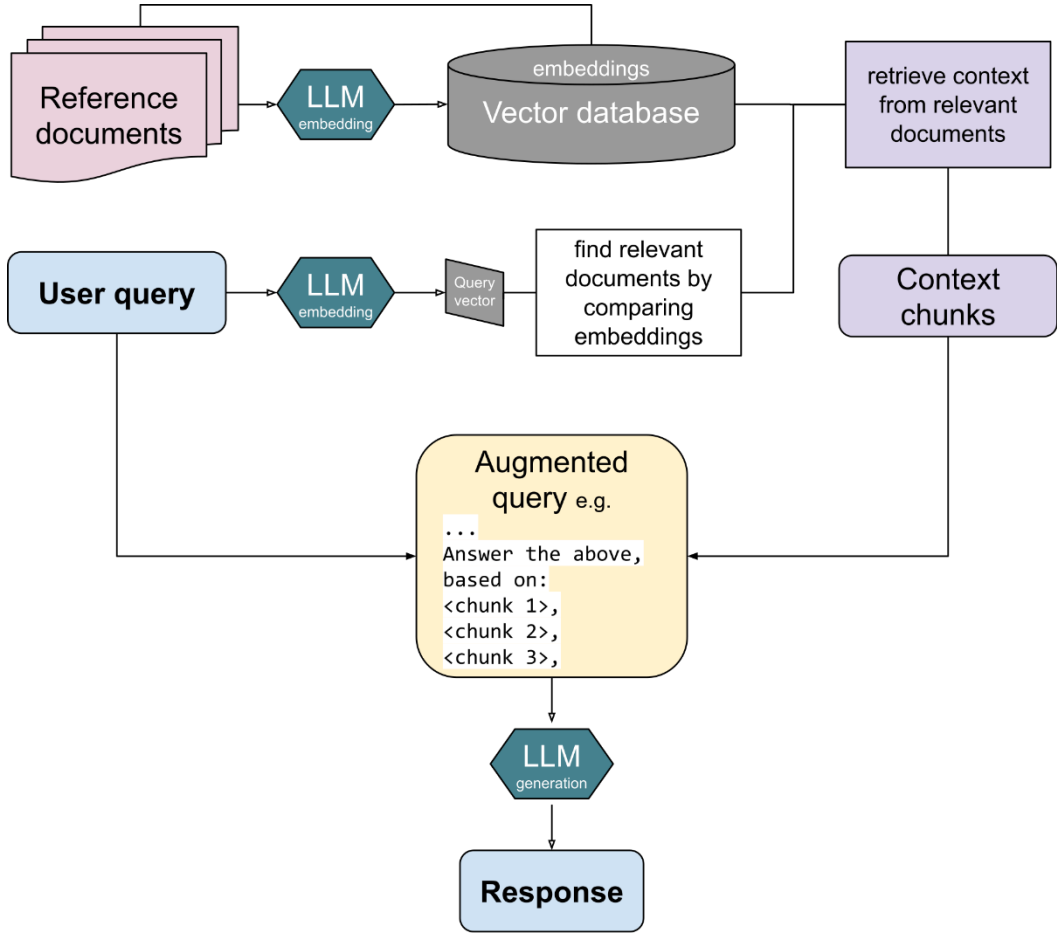
Retrieval Augmented Generation (RAG)

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

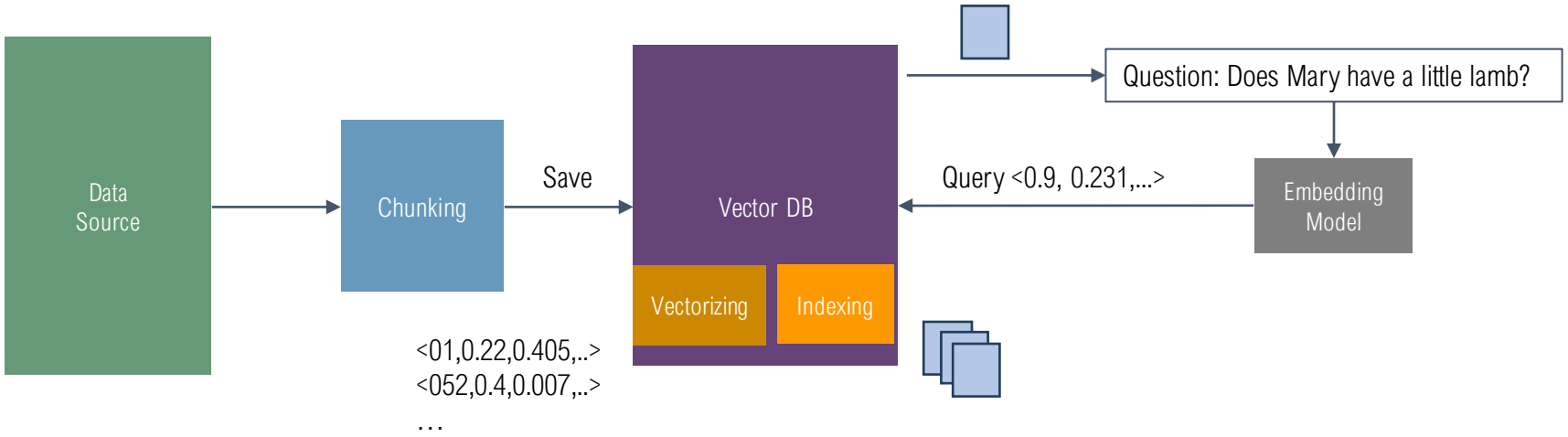
$3.04 Billion

2025 expectation

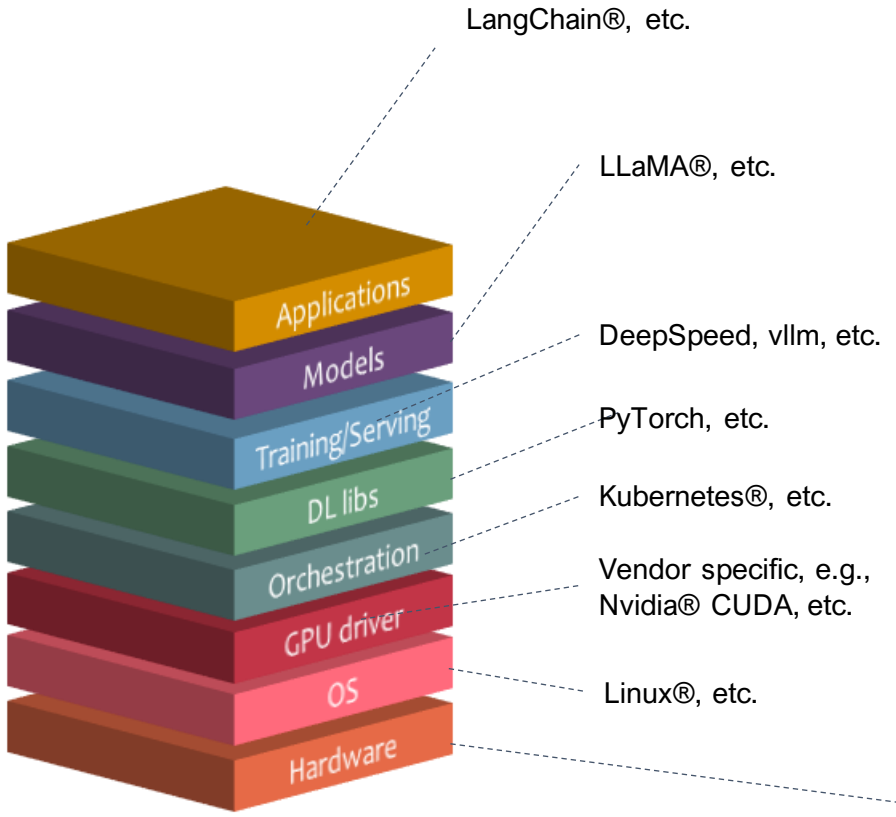
CAGR 23.7%

2024-2029 expectation

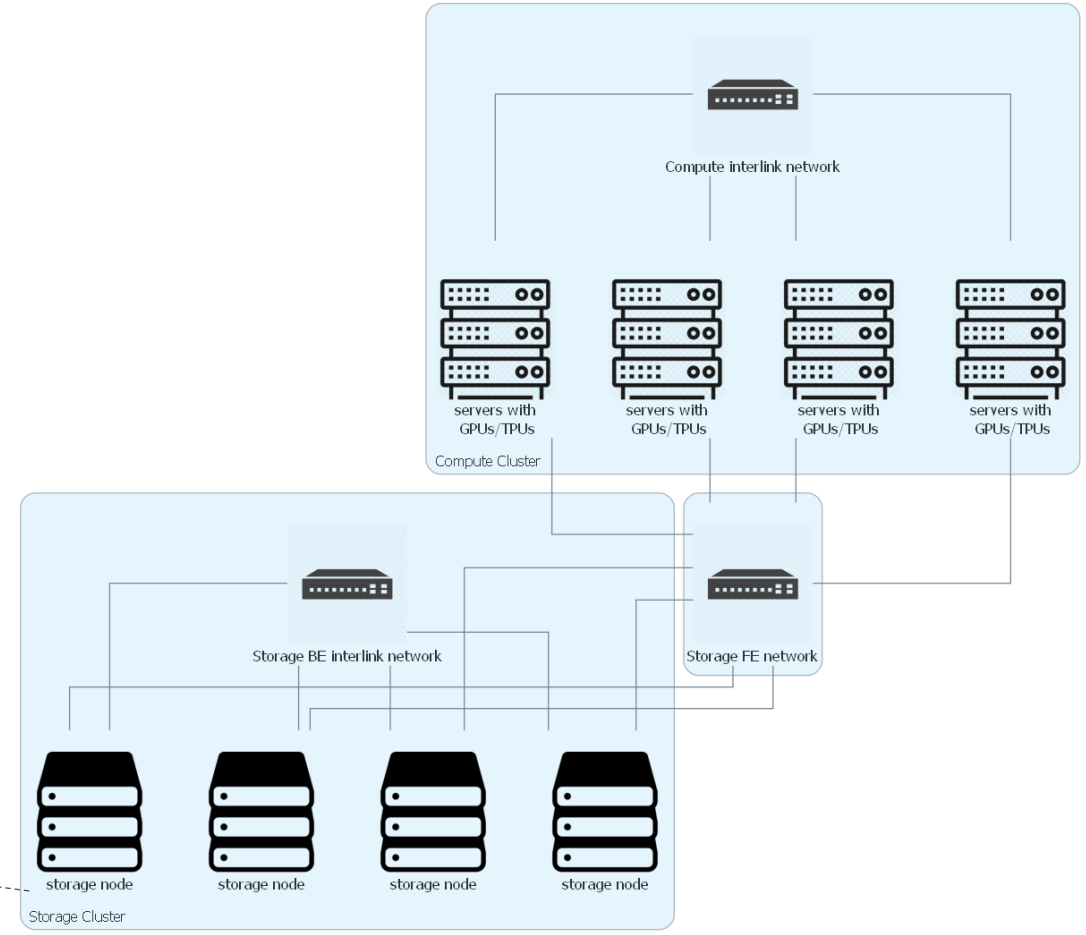
Source: The business research company, Vector Database Market Report 2025 - Vector Database Industry Analysis And Overview



Ecosystem Overview



Open Tech Stack of Gen AI



Common training and inference environment

GPUs

Why GPUs Are Important for AI?

- **Parallel Processing**: AI models perform billions of matrix multiplications, which GPUs handle efficiently.
- **Deep Learning Acceleration**: Neural networks involve heavy computations that GPUs execute much faster than CPUs.
- **Memory Bandwidth**: GPUs have **high-speed memory (VRAM)** to handle large AI datasets.
- **Support for AI Frameworks**: Optimized for many ML libraries.

Feature	CPU	GPU
Cores	Few (4-64)	Thousands
Processing Type	Sequential	Parallel
AI Performance	Slow	Fast
Best for	General computing	Deep learning, AI, graphics

For large scale matrix operations, GPUs can be thousands of times faster than CPUs due to the large number of cores.

Scaling laws and more

- The **Scaling Law** (Kaplan et al., 2020, [2001.08361] Scaling Laws for Neural Language Models) describes how the performance of AI models improves as we **increase** the amount of **compute, model size, and training data**.
- It shows that **larger models, when trained on more data with more compute, follow predictable improvements** in loss and capability.
- Chinchilla law (Hoffmann et al. 2022, [2203.15556] Training Compute-Optimal Large Language Models) emphasizes that for fixed budget, smaller model with more data can be more effective than bigger model with less data.

Computing is measured in FLOPs (Floating Point Operations per Second)
Or total GPU hours with a certain model

✅ **Larger Models Keep Getting Better** (So far still is).
✅ **More Data is as Important as Model Size** (Chinchilla: 4x data = 2x model size).
✅ **Training Efficiency Matters** (Scaling laws help optimize compute costs).
✅ **Transfer to Multimodal AI** (Text, images, video, and code models scale similarly).

Training/Inference Tool Examples

Tool #1
Deepspeed®: It's an open-source tool for training and inferencing. It was widely used worldwide.

Follow the link for more information:
Latest News - DeepSpeed

Tool #2
vLLM: It's an open-source tool for inferencing. It evolved from a UC Berkeley project to a full-fledged open-source project.

Follow the link for more information:
Welcome to vLLM — vLLM

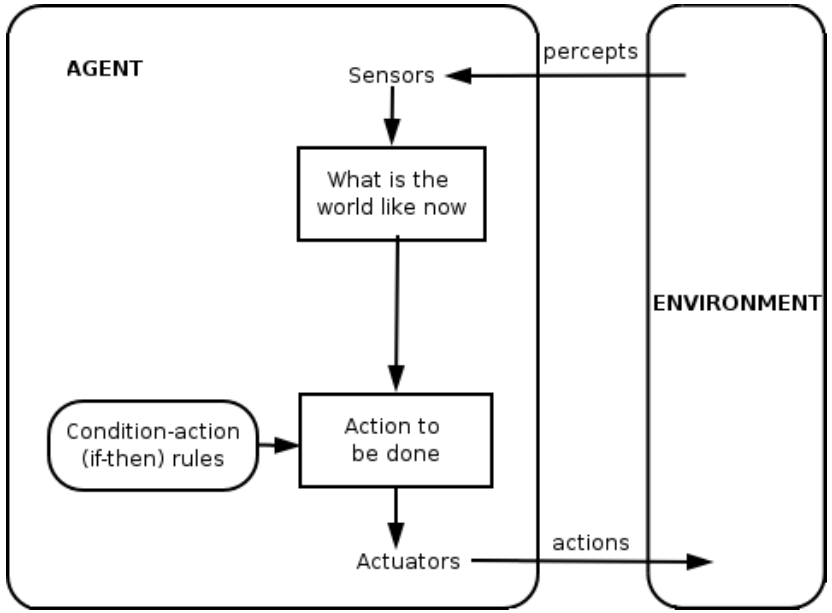
Agentic Workflow

"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

Tools and Framework Examples

Frame #1
LangChain® provides not only the packages for using LLMs for different tasks, but also the packages called LangGraph® to build agents.

Follow the link for more information:
Introduction | 🦜🔗 LangChain

Frame #2
LlamaIndex® provides similar functionalities as LangChain® but focuses on efficient data retrieval for RAG.

Follow the link for more information:
LlamaIndex - Build Knowledge Assistants over your Enterprise Data

Frame #3
Huggingface® is where you can find all kinds of information about LLM, including models, datasets, and different examples.

Follow the link for more information:
Hugging Face – The AI community building the future.

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