Week2.1 More LLMs

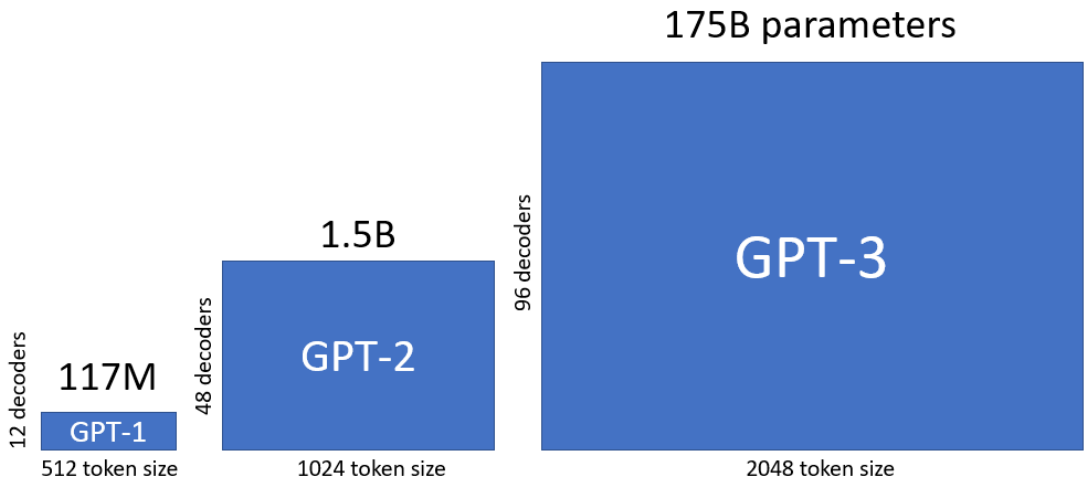
2025 Spring GenAI

Dr. Yanjun Qi

20250121

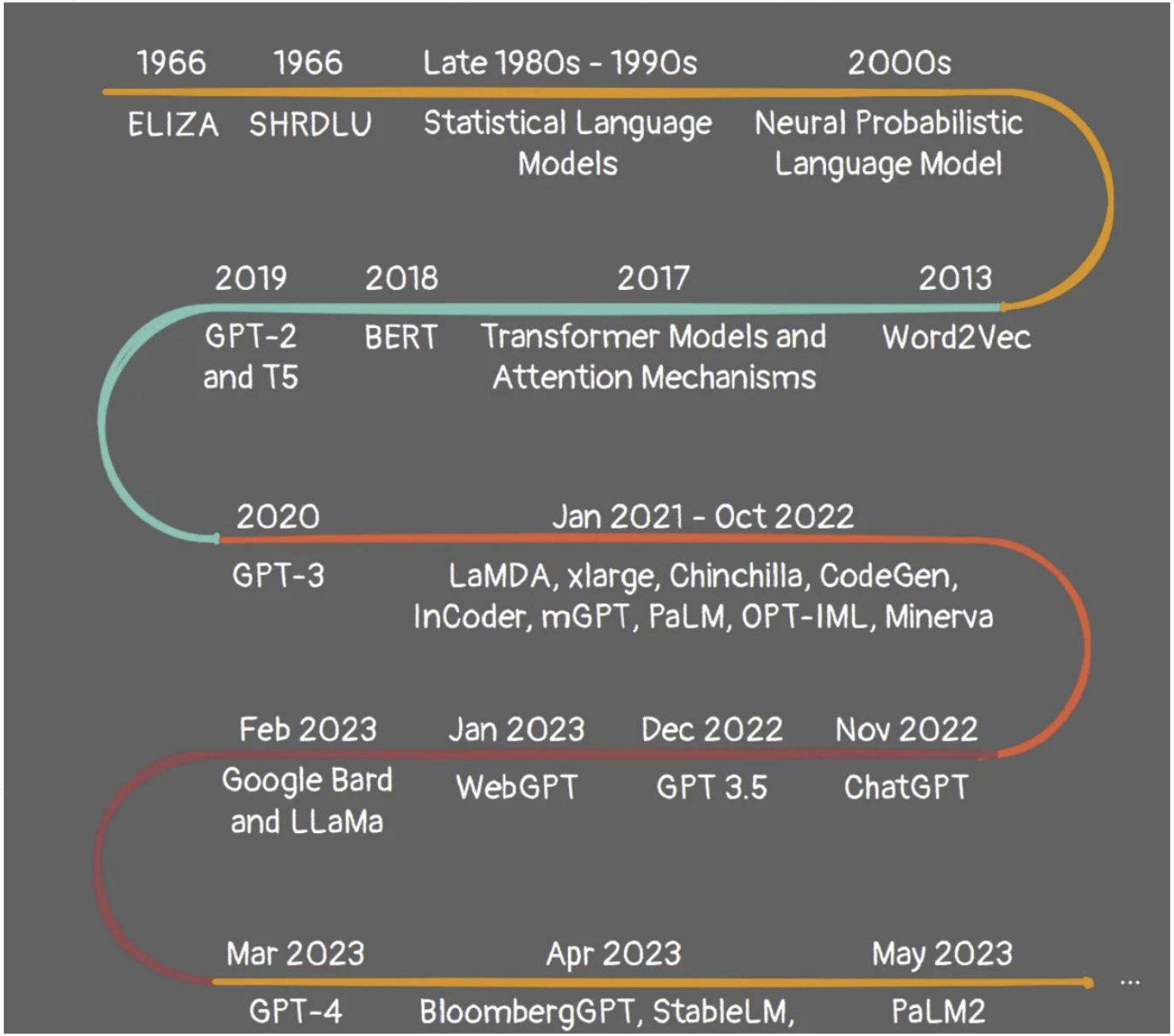
Last Class:

- GPT1 / 2/ 3
- Emergent Abilities of Large Language Models
- Scaling Instruction-Finetuned Language Models
- On the Opportunities and Risks of Foundation Models



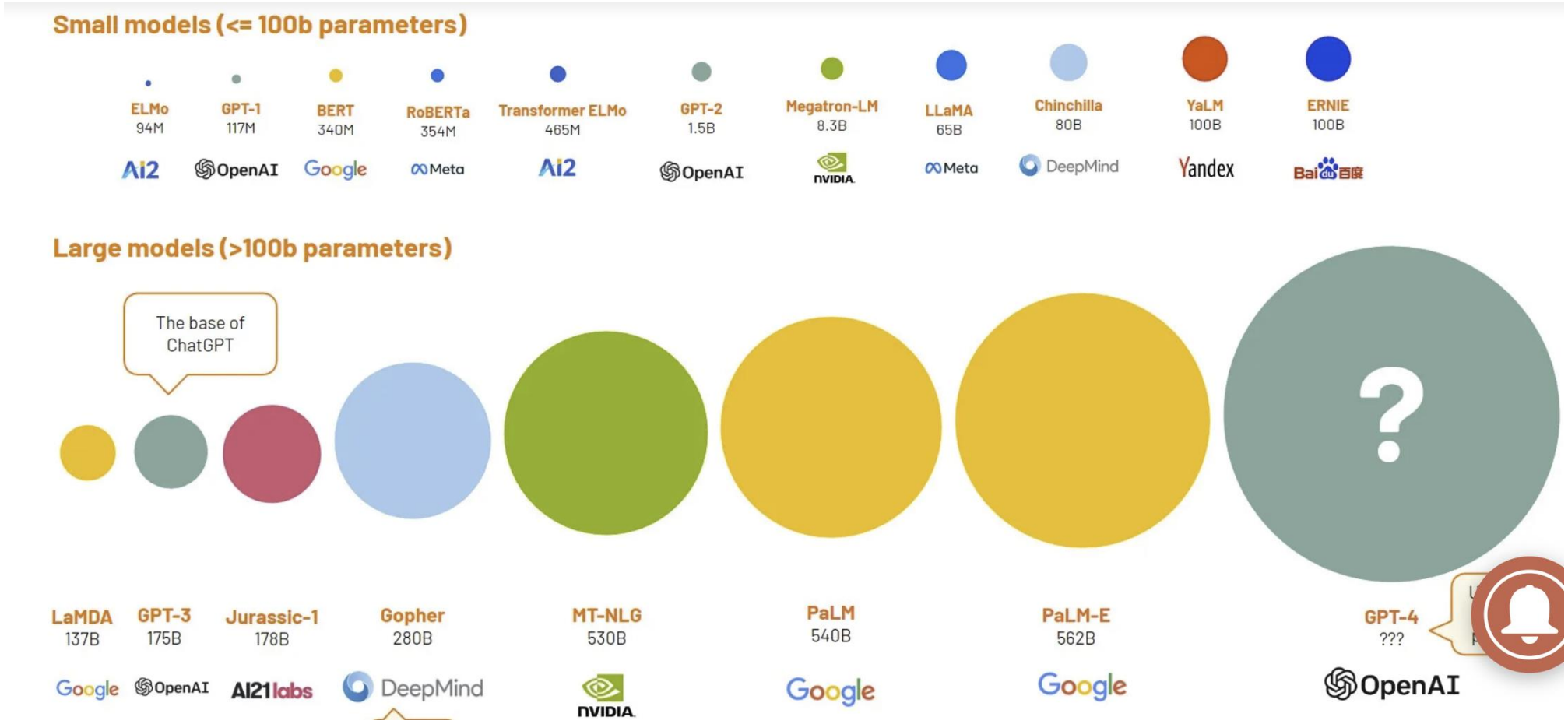
https://medium.com/@YanAIx/step-by-step-into-gpt-70bc4a5d8714

Many new LLMs in 2022-2023 ->2024



https://levelup.gitconnected.com/the-brief-history-of-large-language-models-a-journey-from-eliza-to-gpt-4-and-google-bard-167c614af5af

LLMs Size changes



https://levelup.gitconnected.com/the-brief-history-of-large-language-models-a-journey-from-eliza-to-gpt-4-and-google-bard-167c614af5af

🏆 LMSYS Chatbot Arena Leaderboard

| Vote | Blog | GitHub | Paper | Dataset | Twitter | Discord |

LMSYS Chatbot Arena is a crowdsourced open platform for LLM evals. We've collected over **200,000** human preference votes to rank LLMs with the Elo ranking system.

Arena Elo | Full Leaderboard

Total #models: **55**. Total #votes: **230875**. Last updated: Jan 18, 2024.

Contribute your vote 🗳️ at chat.lmsys.org! Find more analysis in the notebook.

Rank	🤖 Model	⭐ Arena Elo	📊 95% CI	🗳️ Votes	Organization	License
1	GPT-4-Turbo	1249	+14/-13	27399	OpenAI	Proprietary
2	GPT-4-0314	1191	+15/-14	17372	OpenAI	Proprietary
3	GPT-4-0613	1160	+13/-13	24888	OpenAI	Proprietary
4	Claude-1	1150	+14/-13	17333	Anthropic	Proprietary
5	Mistral Medium	1148	+14/-13	9370	Mistral	Proprietary
6	Claude-2.0	1131	+14/-13	11475	Anthropic	Proprietary
7	Mixtral-8x7b-Instruct-v0.1	1124	+15/-13	13485	Mistral	Apache 2.0
8	Gemini Pro (Dev)	1121	+15/-15	5304	Google	Proprietary

🏆 Chatbot Arena LLM Leaderboard: Community-driven Evaluation for Best LLM and AI chatbots

小红书 | Twitter | Discord | Blog | GitHub | Paper | Dataset | Kaggle Competition

Chatbot Arena is an open platform for crowdsourced AI benchmarking, developed by researchers at UC Berkeley SkyLab and LMArena. With over 1,000,000 user votes, the platform ranks best LLM and AI chatbots using the Bradley-Terry model to generate live leaderboards. For technical details, check out our paper.

Chatbot Arena thrives on community engagement — cast your vote to help improve AI evaluation!

New Launch! WebDev Arena: web.lmarena.ai - AI Battle to build the best website!

💬 Language | 🎛️ Overview | 👀 Vision | 🎨 Text-to-Image | 📟 Copilot Arena | 🌐 WebDev Arena | Arena-Hard-Auto

Total #models: **194**. Total #votes: **2,557,144**. Last updated: 2025-01-20.

Code to recreate leaderboard tables and plots in this notebook. You can contribute your vote at lmarena.ai!

Category: Overall

Apply filter: Style Control | Show Deprecated

Overall Questions

#models: 194 (100%) **#votes: 2,557,144 (100%)**

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License
1	3	Gemini-2.0-Flash-Thinking-Exp-01-21	1380	+8/-9	5572	Google	Proprietary
1	1	Gemini-Exp-1206	1374	+5/-5	21004	Google	Proprietary
3	1	ChatGPT-4o-latest (2024-11-20)	1365	+4/-3	34209	OpenAI	Proprietary
4	4	Gemini-2.0-Flash-Exp	1356	+4/-6	19823	Google	Proprietary
4	1	o1-2024-12-17	1351	+8/-5	8124	OpenAI	Proprietary
6	4	o1-preview	1335	+4/-4	33202	OpenAI	Proprietary
7	7	DeepSeek-V3	1320	+5/-5	11893	DeepSeek	DeepSeek

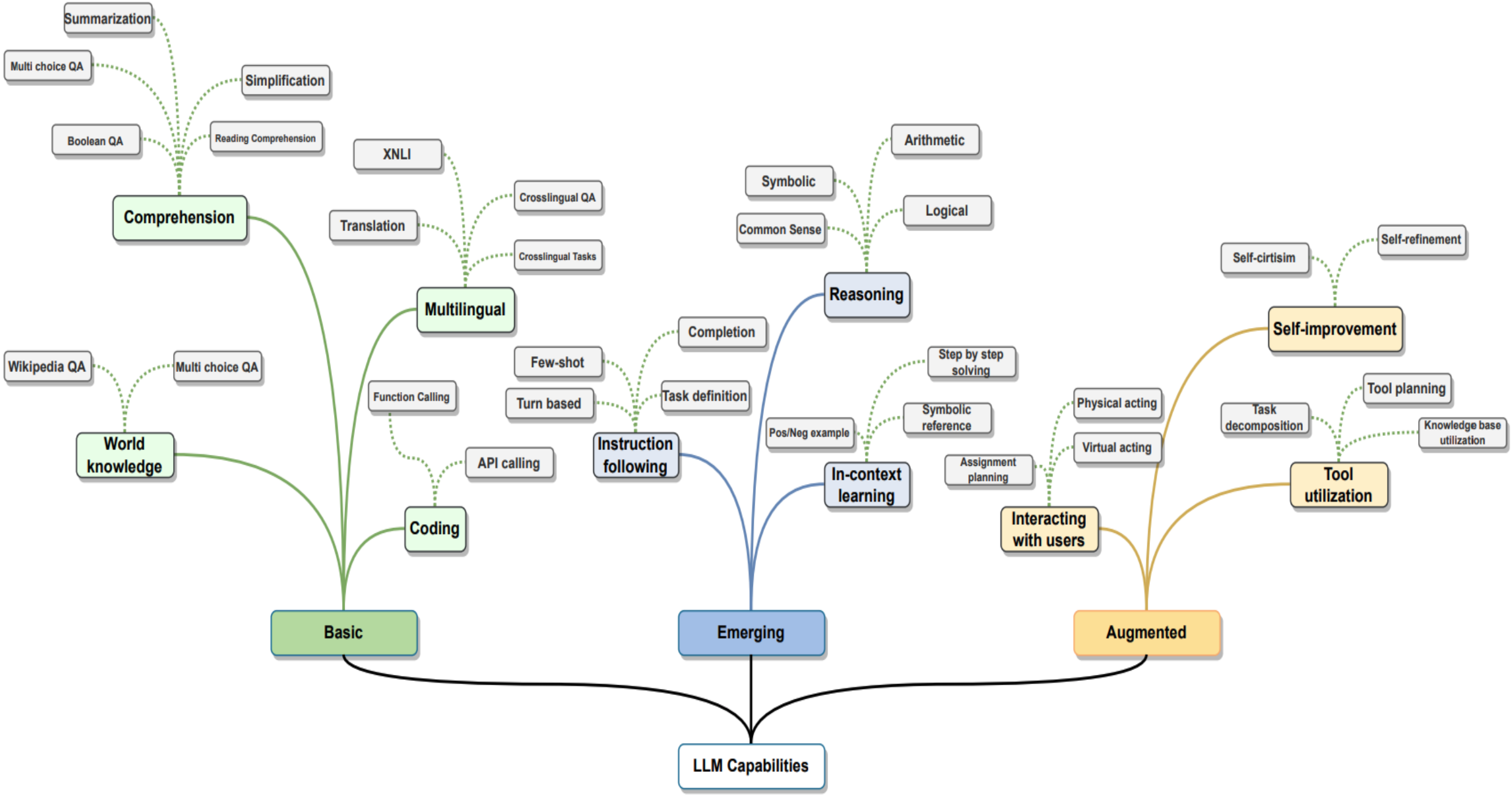
Large Language Models: A Survey

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu
Richard Socher, Xavier Amatriain, Jianfeng Gao

***Abstract*—Large Language Models (LLMs) have drawn a lot of attention due to their strong performance on a wide range of natural language tasks, since the release of ChatGPT in November 2022. LLMs' ability of general-purpose language understanding and generation is acquired by training billions of model's parameters on massive amounts of text data, as predicted by scaling laws [1], [2]. The research area of LLMs, while very recent, is evolving rapidly in many different ways. In this paper, we review some of the most prominent LLMs, including three popular LLM families (GPT, LLaMA, PaLM), and discuss their characteristics, contributions and limitations. We also give an overview of techniques developed to build, and augment LLMs. We then survey popular datasets prepared for LLM training, fine-tuning, and evaluation, review widely used LLM evaluation metrics, and compare the performance of several popular LLMs on a set of representative benchmarks. Finally, we conclude the paper by discussing open challenges and future research directions.**

that have different starting points and velocity: statistical language models, neural language models, pre-trained language models and LLMs.

Statistical language models (SLMs) view text as a sequence of words, and estimate the probability of text as the product of their word probabilities. The dominating form of SLMs are Markov chain models known as the n-gram models, which compute the probability of a word conditioned on its immediate proceeding $n-1$ words. Since word probabilities are estimated using word and n-gram counts collected from text corpora, the model needs to deal with data sparsity (i.e., assigning zero probabilities to unseen words or n-grams) by using *smoothing*, where some probability mass of the model is reserved for unseen n-grams [12]. N-gram models are widely used in many NLP systems. However, these models are incomplete in that they cannot fully capture the diversity and variability of natural language due to data sparsity.


LLM Capabilities
Basic
Emerging
Augmented
Comprehension
Summarization
Multi choice QA
Simplification
Boolean QA
Reading Comprehension
Multilingual
XNLI
Translation
Crosslingual QA
Crosslingual Tasks
World knowledge
Wikipedia QA
Multi choice QA
Coding
Function Calling
API calling
Instruction following
Few-shot
Completion
Turn based
Task definition
Reasoning
Symbolic
Arithmetic
Common Sense
Logical
In-context learning
Pos/Neg example
Step by step solving
Symbolic reference
Interacting with users
Assignment planning
Physical acting
Virtual acting
Tool utilization
Task decomposition
Tool planning
Knowledge base utilization
Self-improvement
Self-cirtisim
Self-refinement

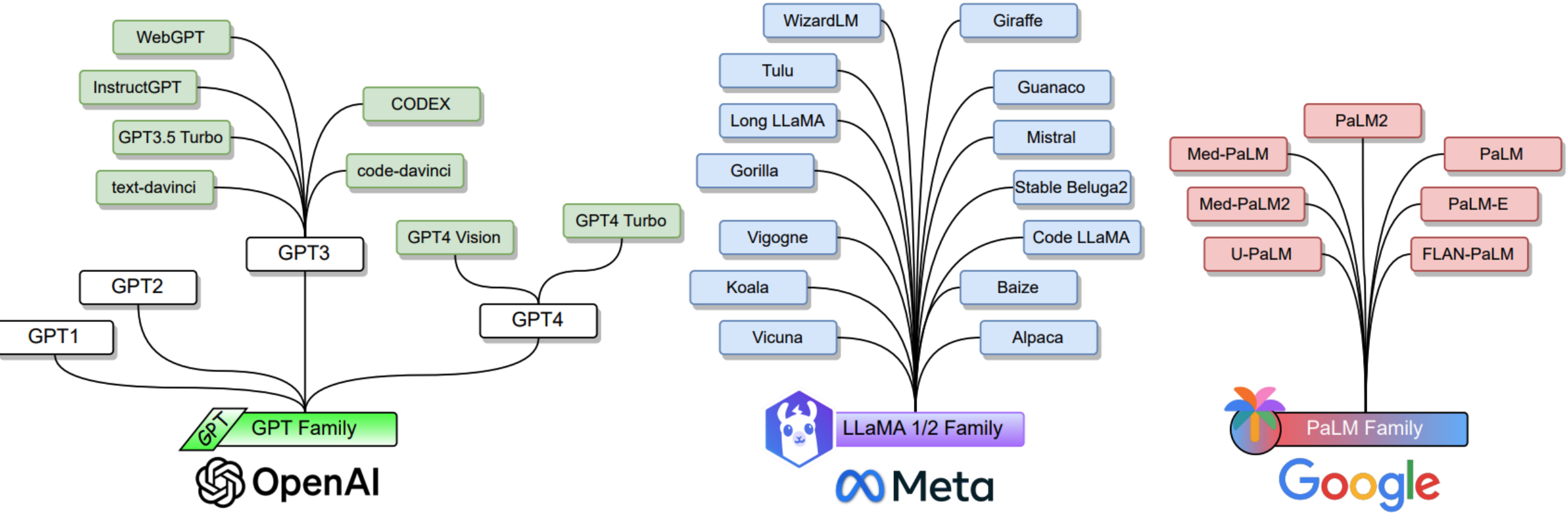


Fig. 8: Popular LLM Families.

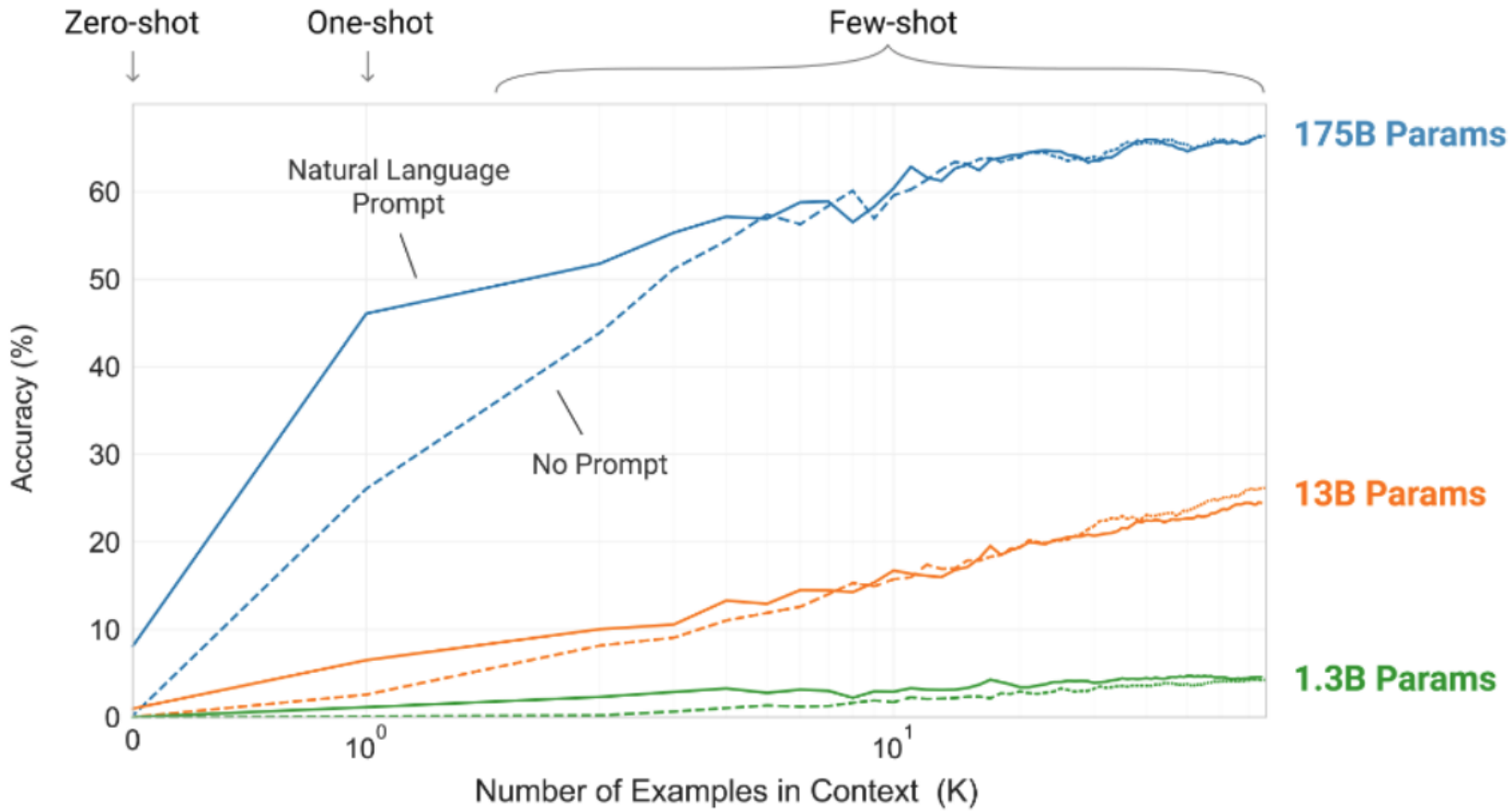


Fig. 9: GPT-3 shows that larger models make increasingly efficient use of in-context information. It shows in-context

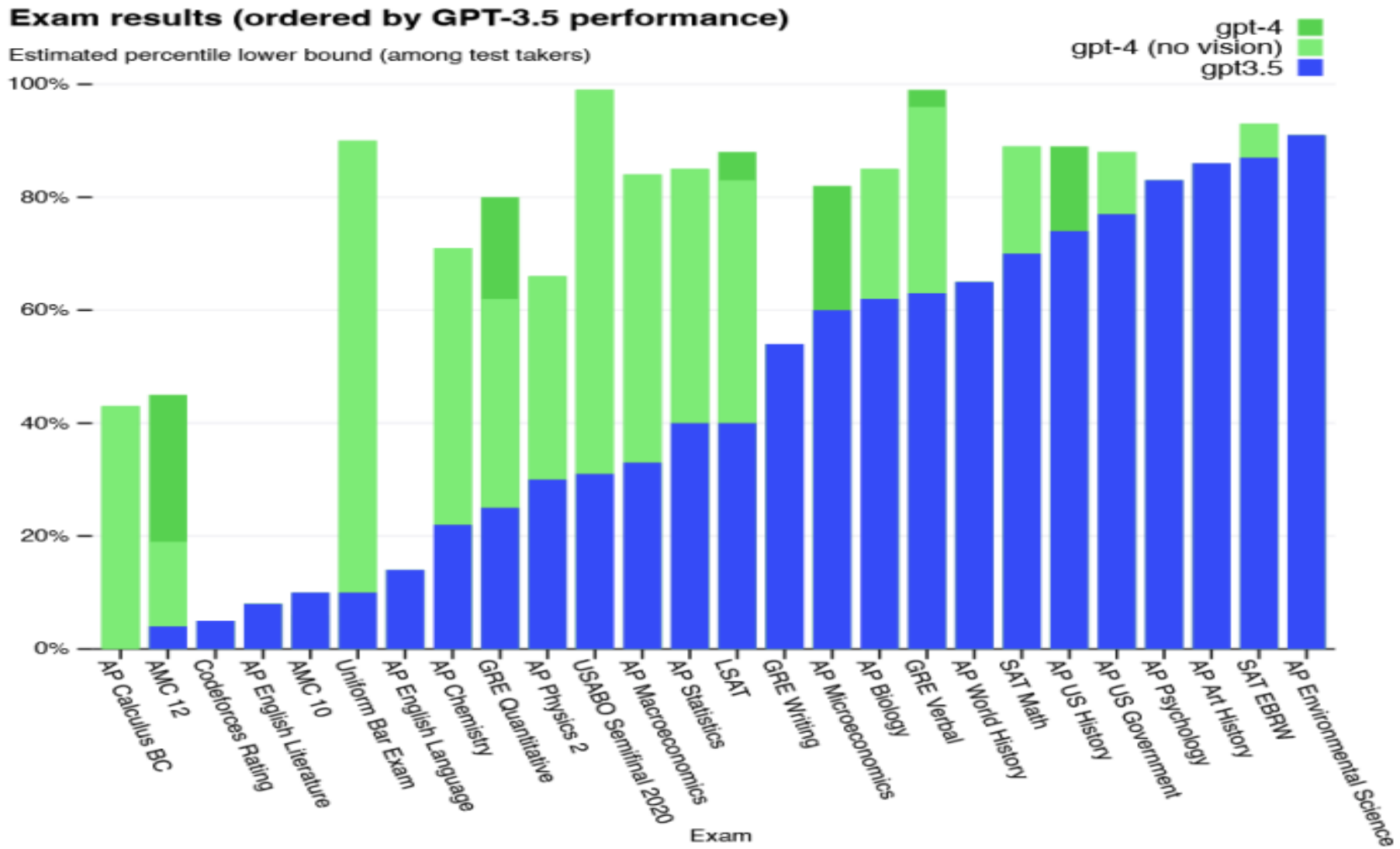
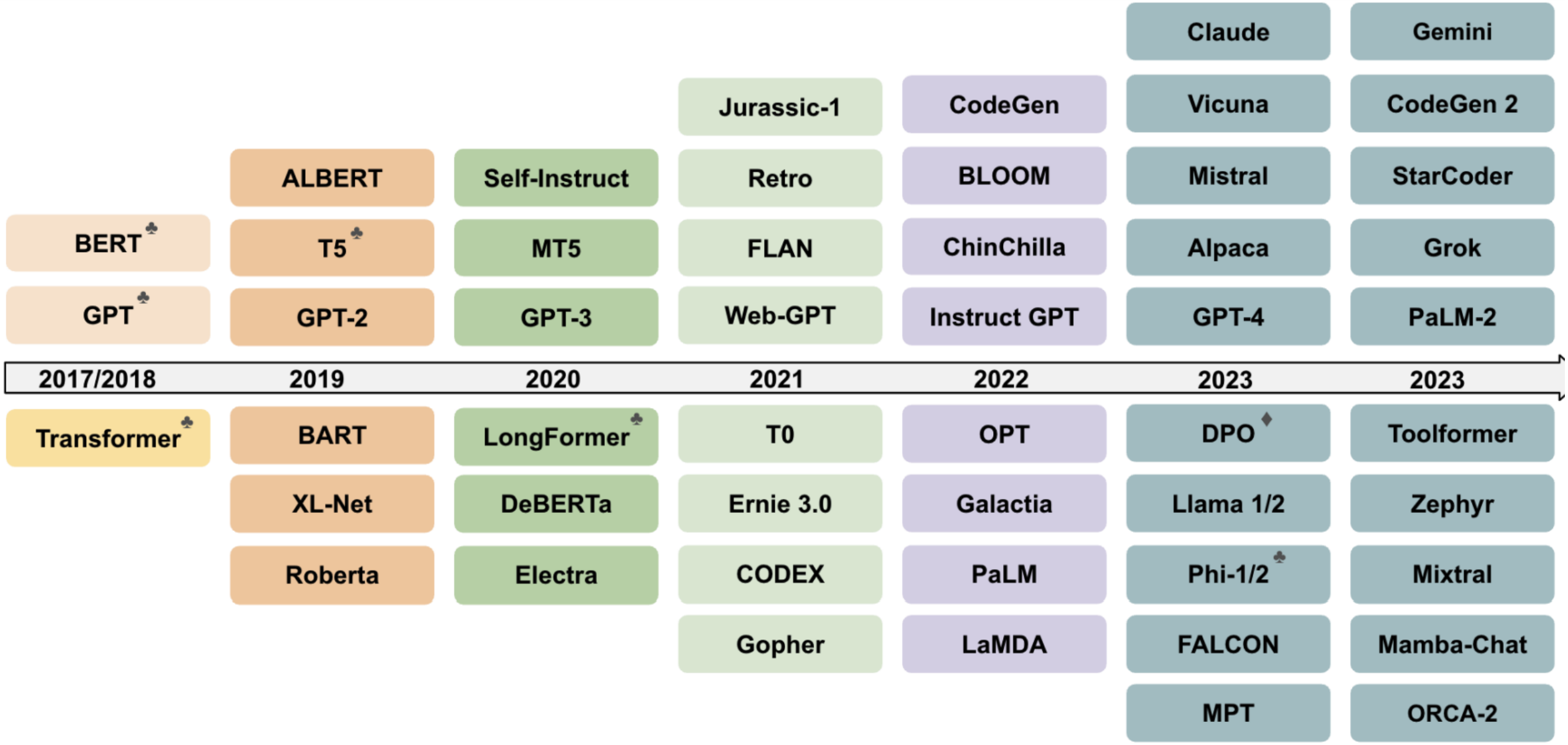


Fig. 11: GPT-4 performance on academic and professional exams, compared with GPT 3.5. Courtesy of [33].

Timeline of some of the most representative LLM frameworks



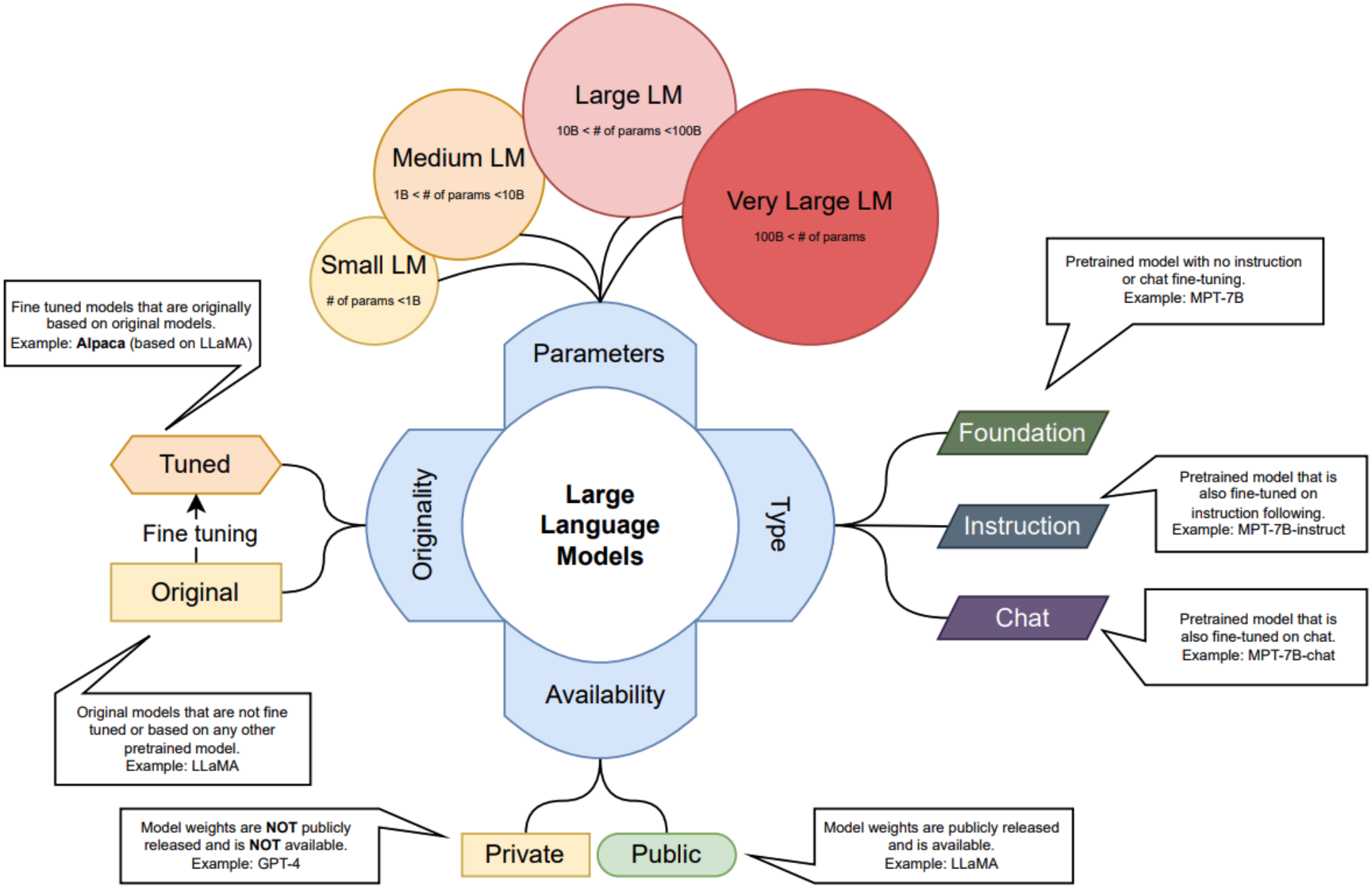
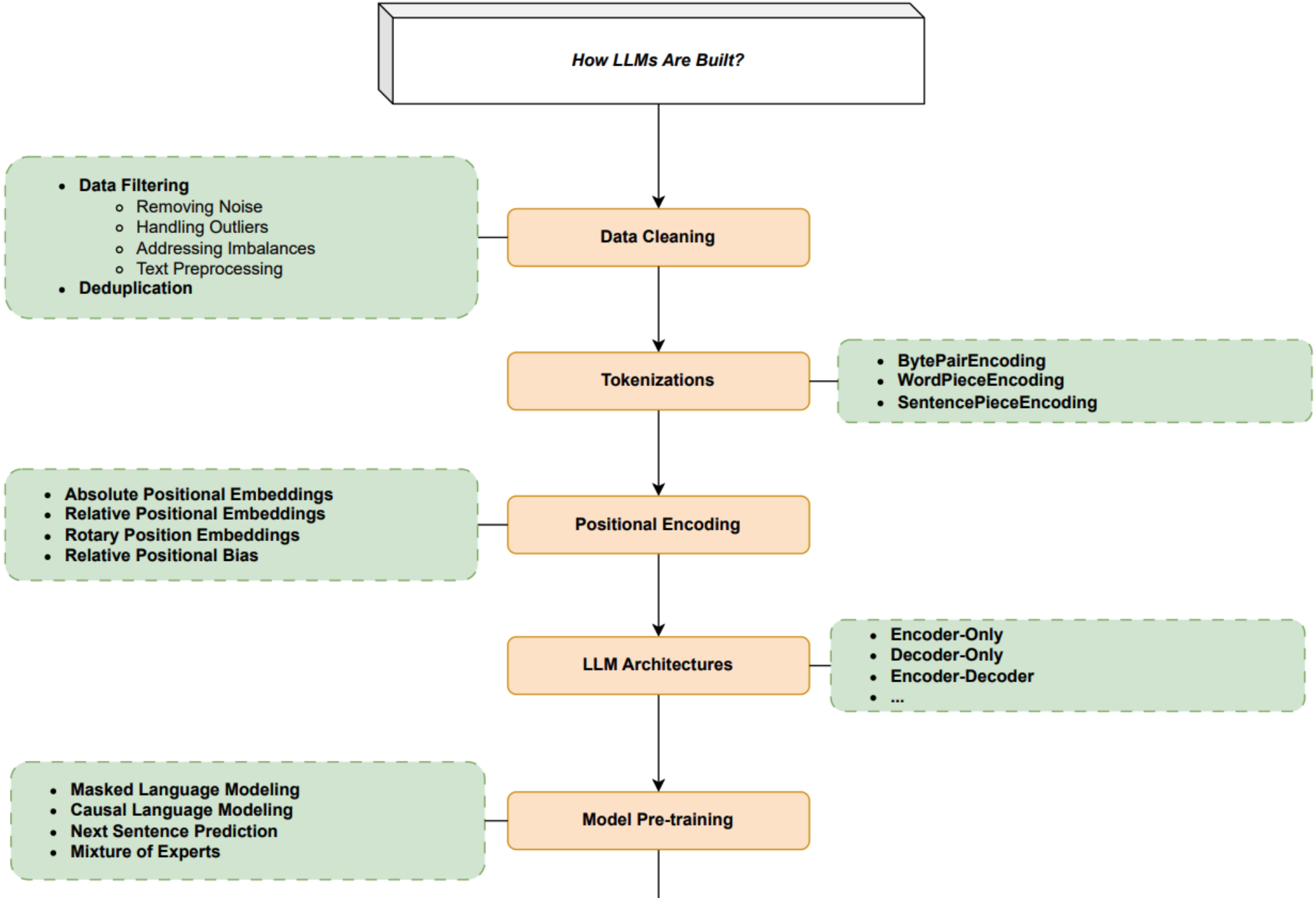


Fig. 43: LLM categorizations.


How LLMs Are Built?
Data Filtering
Removing Noise
Handling Outliers
Addressing Imbalances
Text Preprocessing
Deduplication
Data Cleaning
Tokenizations
BytePairEncoding
WordPieceEncoding
SentencePieceEncoding
Absolute Positional Embeddings
Relative Positional Embeddings
Rotary Position Embeddings
Relative Positional Bias
Positional Encoding
LLM Architectures
Encoder-Only
Decoder-Only
Encoder-Decoder
...
Masked Language Modeling
Causal Language Modeling
Next Sentence Prediction
Mixture of Experts
Model Pre-training

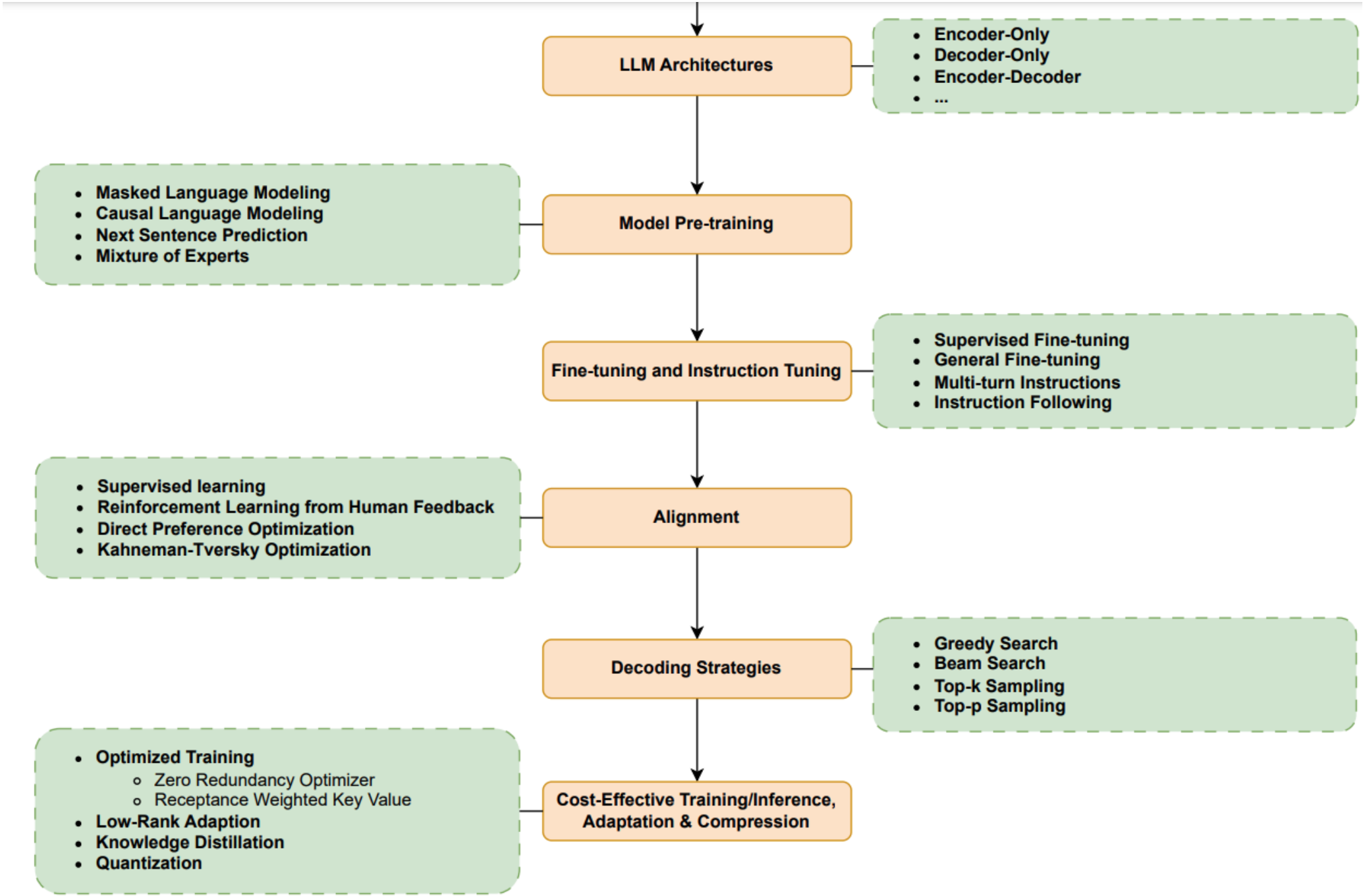


Fig. 25: This figure shows different components of LLMs.

How LLMs Are Used and Augmented

A) LLM limitations

- **Hallucination**
- **Hallucination Quantification**
  - Automated metrics
    - Statistical Metrics
    - Model-Based Metrics
      - IE-Based Metrics
      - QA-Based Metrics
      - NLI-Based Metrics
  - Human judgment
    - Scoring
    - Comparative Analysis

PROMPT ENGINEERING

B) Using LLMs

Prompt Design and Engineering

1) Chain of Thought
- Zero-Shot CoT
- Manual CoT

2) Tree of Thought

3) Self-Consistency

4) Reflection

5) Expert Prompting

6) Chains

7) Rails
- Topical Rails
- Fact-Checking Rails
- Jailbreaking Rails

8) Automatic Prompt Engineering
- Prompt Generation
- Prompt Scoring
- Refinement and Iteration

B) Augmenting LLMs through external knowledge - RAG

Components of a RAG
- Retrieval
- Generation
- Augmentation

RAG Tools
- LangChain
- LlamaIndex
- HayStack
- Meltano
- Cohere Coral
- Flowise AI

a) RAG-aware prompting techniques

C) Using External Tools

a) Tool-aware prompting techniques

D) LLM Agents

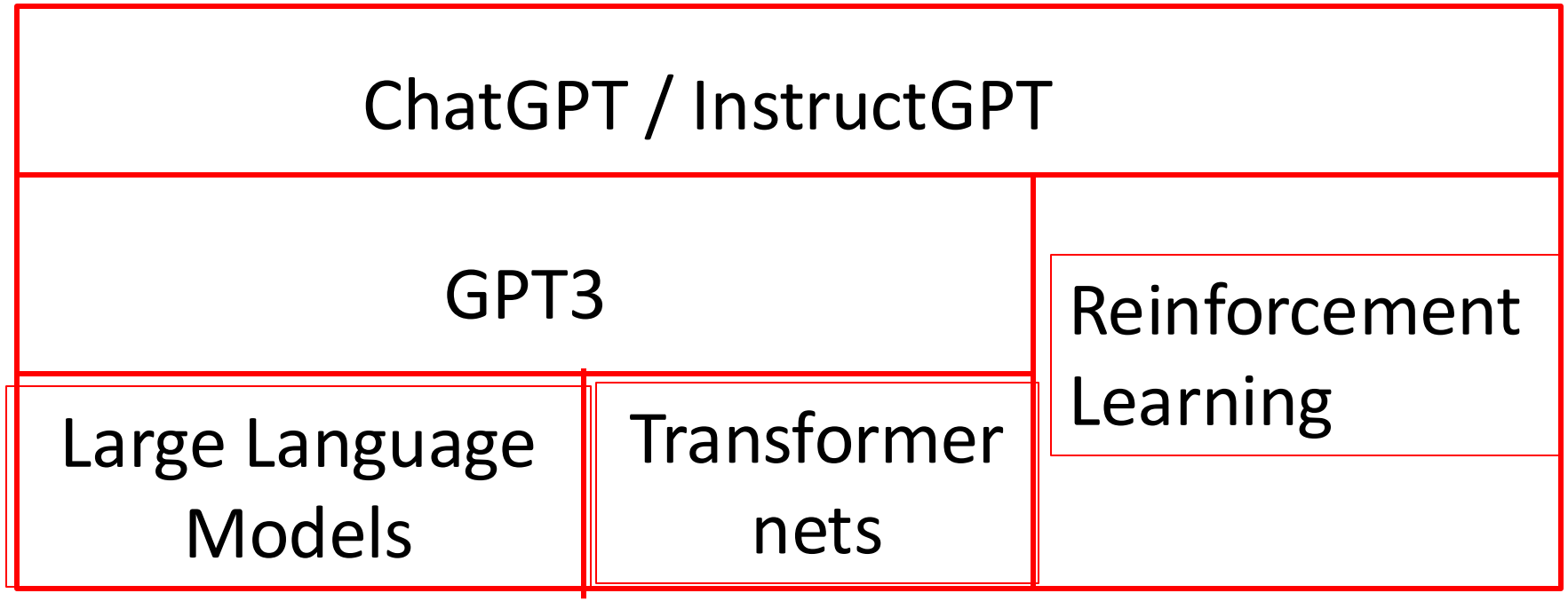
Functionality of an LLM-based agent
- Tool Access and Utilization
- Decision Making

Prompt engineering techniques for agents
- Reasoning without Observation
- Reason and Act
- Dialog-Enabled Resolving Agents

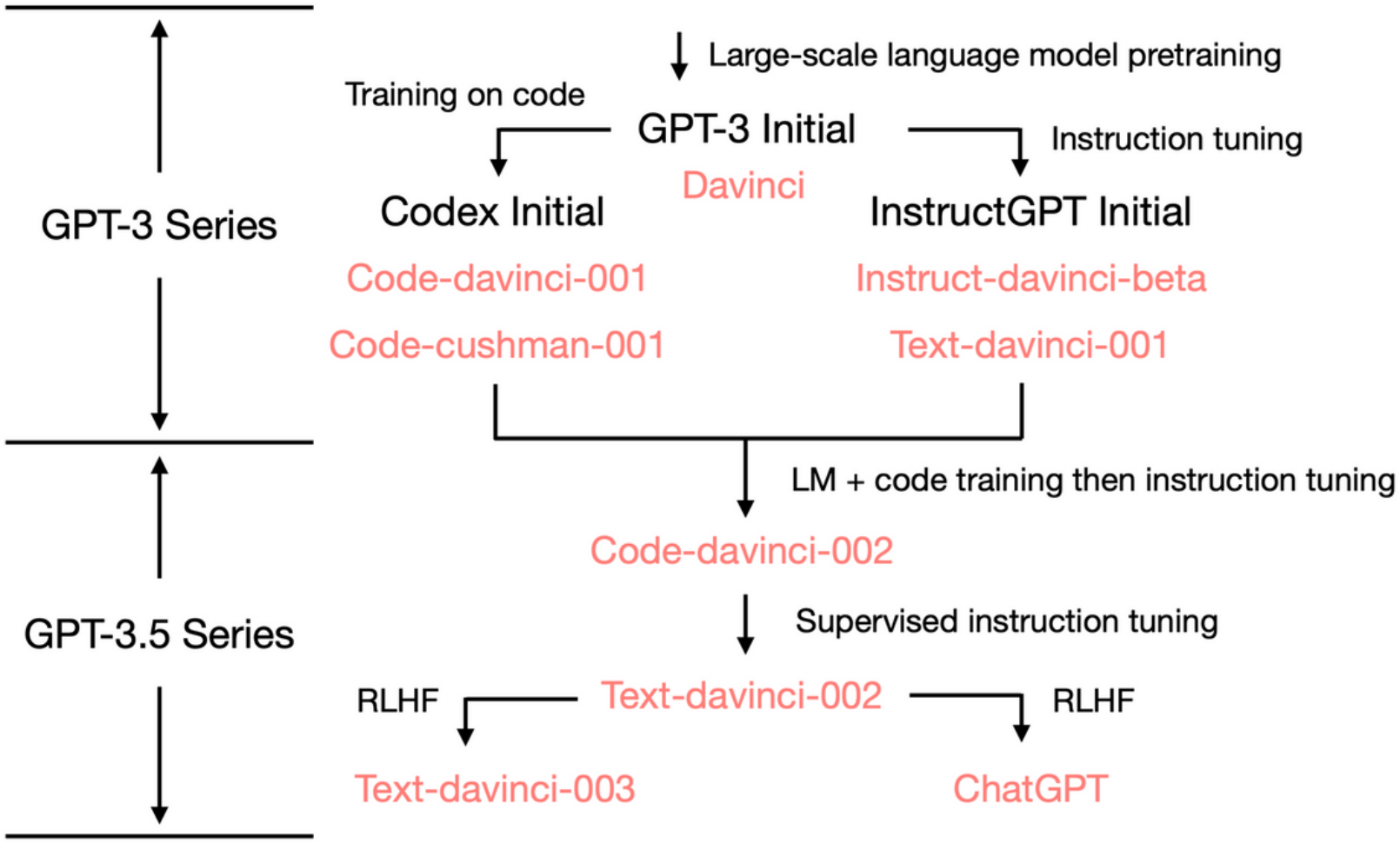
- No paper / Just a blog / Released Nov 30 2022
- It took 5 days to reach 1M users

CHATGPT: OPTIMIZING LANGUAGE MODELS FOR DIALOGUE” BY OPENAI

Concepts that ChatGPT builds on



Family of GPT-3.5



Blog: [How does GPT Obtain its Ability? Tracing Emergent Abilities of Language Models to their Sources]

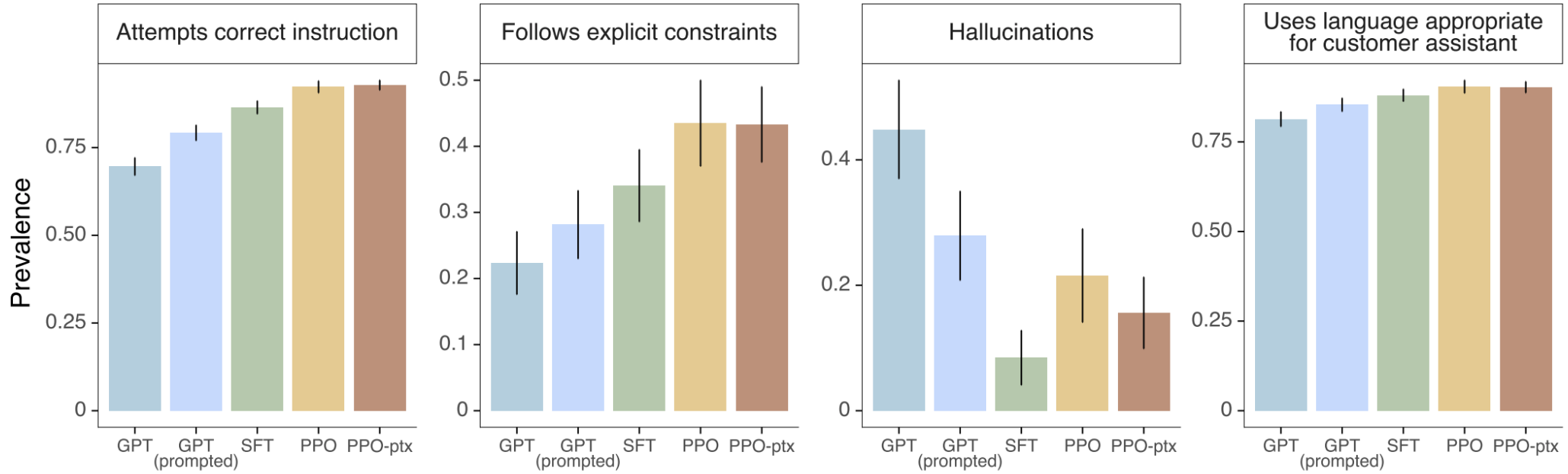


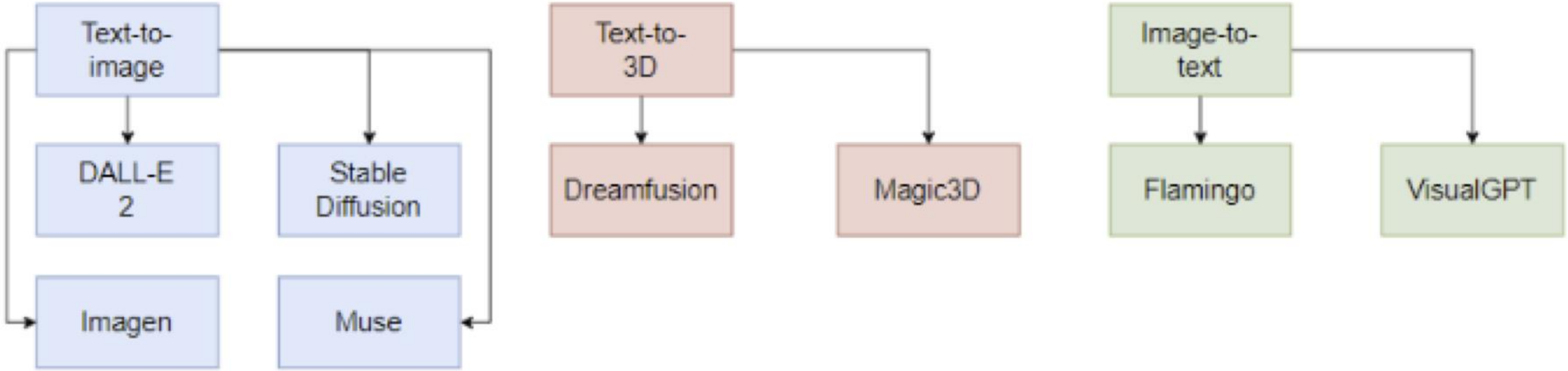
Figure 4: Metadata results on the API distribution. Note that, due to dataset sizes, these results are collapsed across model sizes. See Appendix E.2 for analysis that includes model size. Compared to GPT-3, the PPO models are more appropriate in the context of a customer assistant, are better at following explicit constraints in the instruction and attempting the correct instruction, and less likely

Paper: Training language models to follow instructions with human feedback

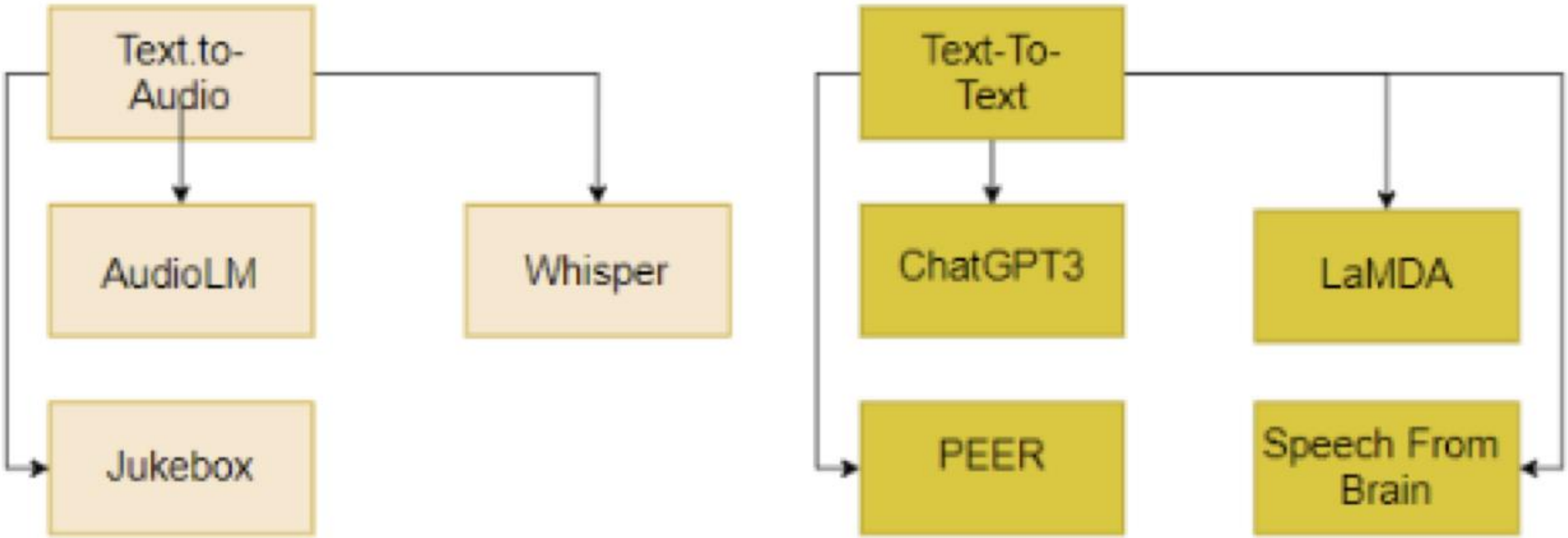
[Submitted on 11 Jan 2023] --- OLD for GenAI

CHATGPT IS NOT ALL YOU NEED. A STATE OF THE ART REVIEW OF LARGE GENERATIVE AI MODELS

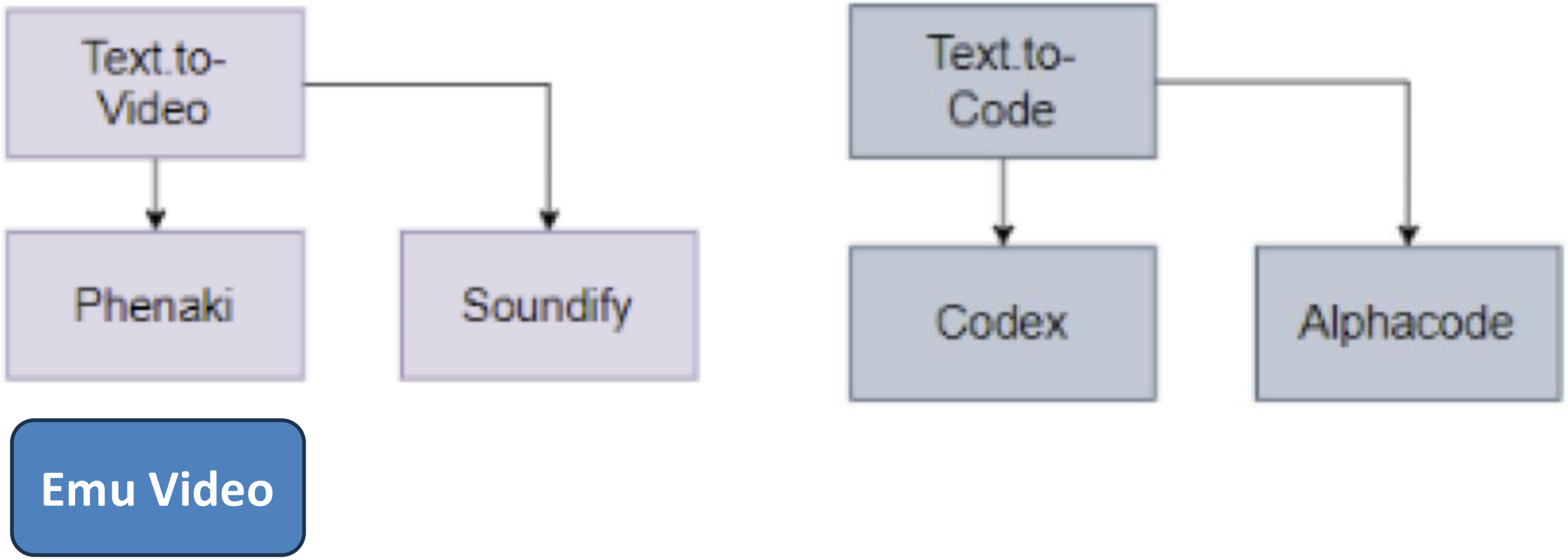
A taxonomy of the most popular generative AI models that have recently appeared classified according to their input and generated formats.



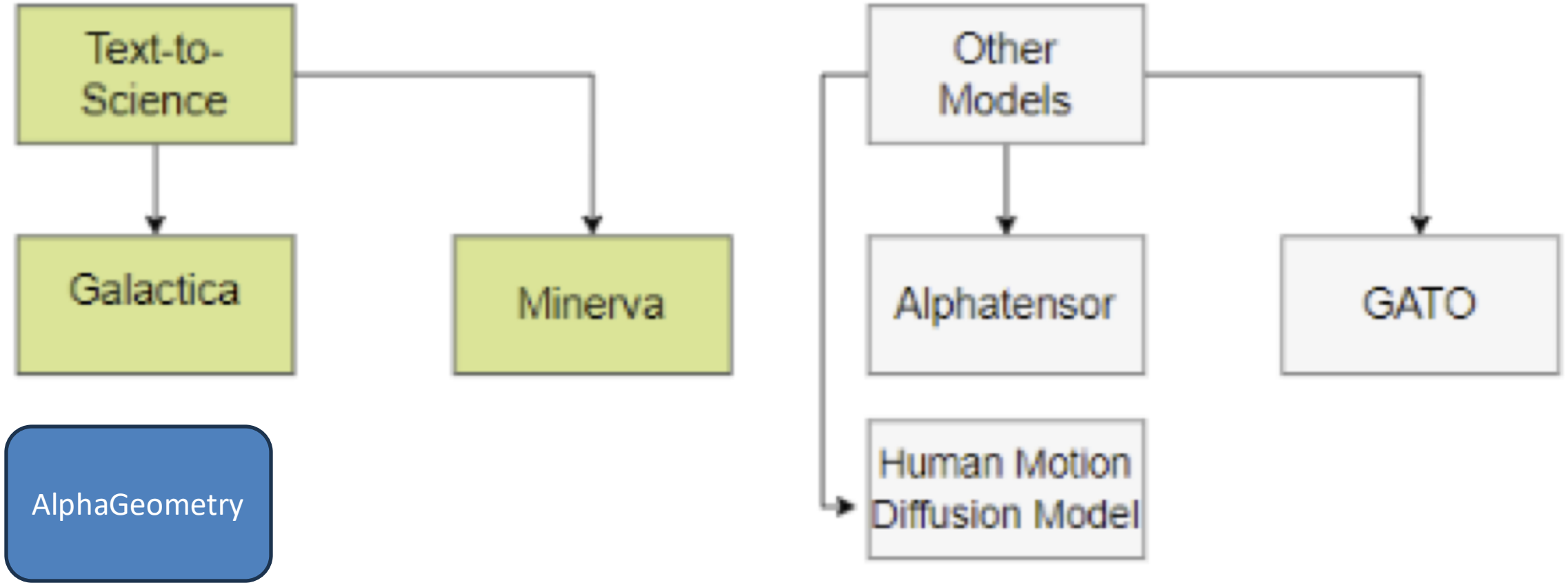
Many new ones out in each type in 2023



Many new ones out in each type in 2023



Many new ones out in each type in 2023

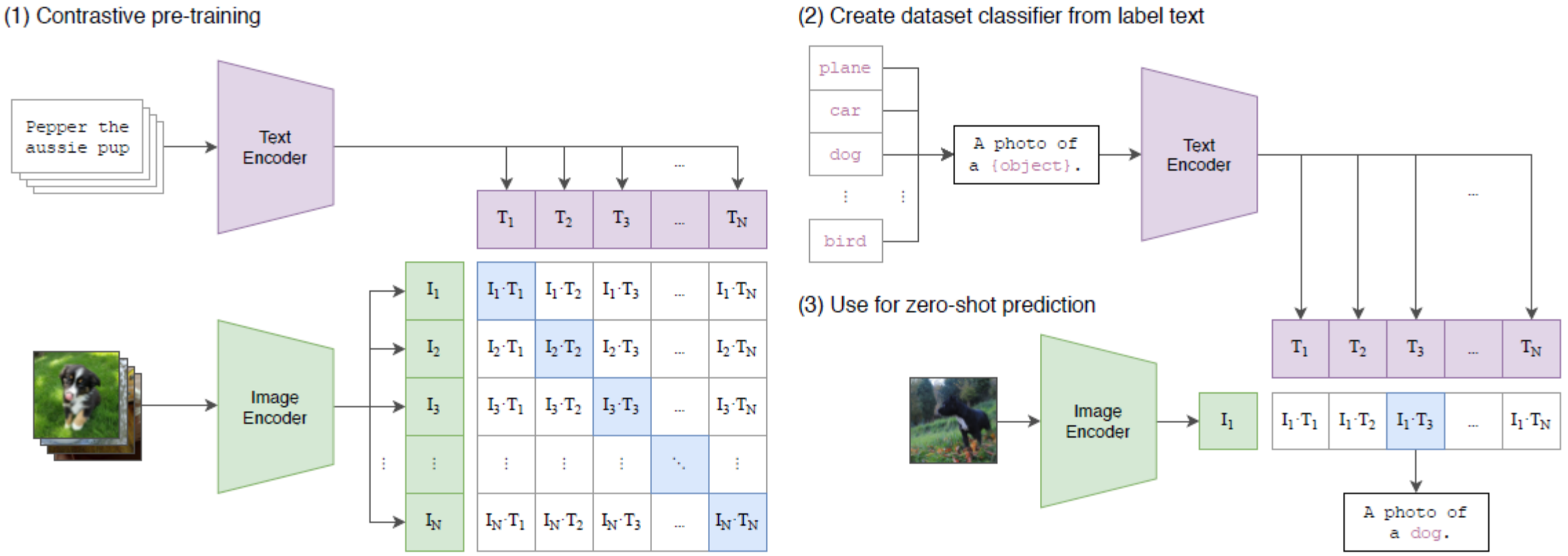


Many new ones out in each type in 2023

BACKUP OLD RELATED

CLIP: CONTRASTIVE LANGUAGE-IMAGE PRETRAINING FOR VISION

CLIP: Learning Transferrable Visual Models from Natural Language Supervision (Radford et al. 2021)

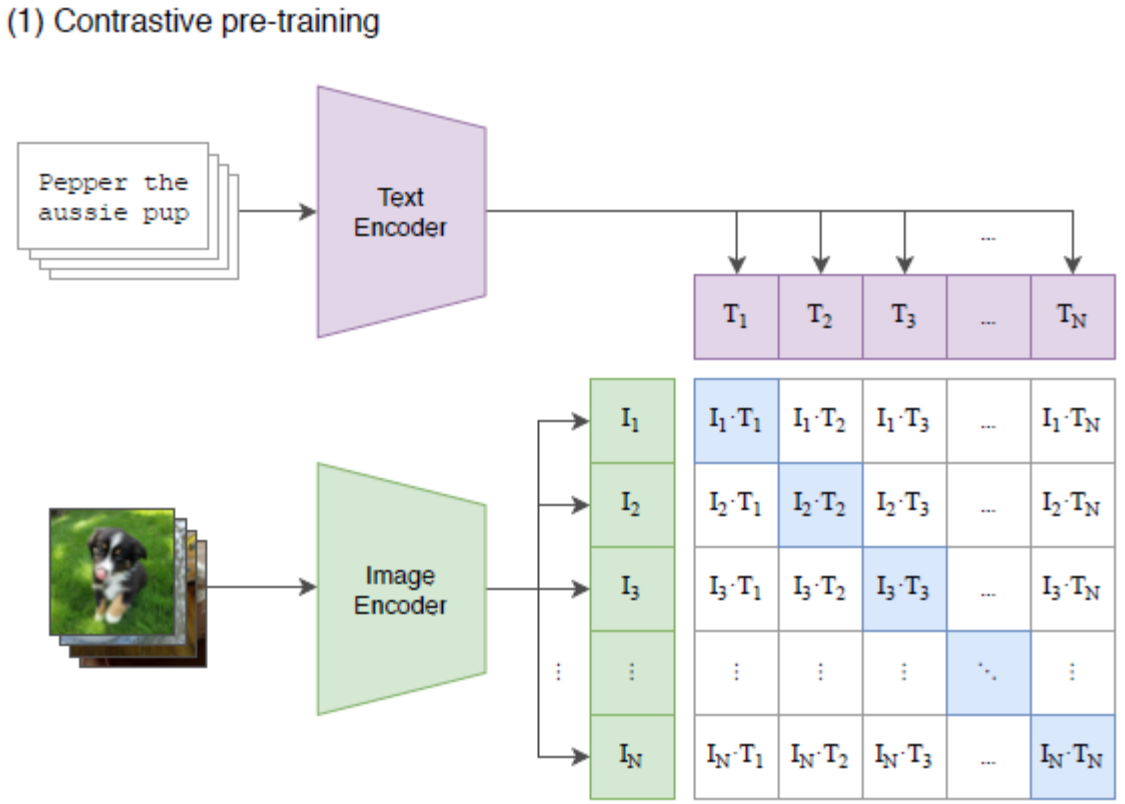


use a text encoder as a classifier

Second key idea(s): **contrastively match text to image**

- Use small transformer language model (76M parameters for base)

"The largest ResNet model RN50x64, took 18 days to train on 592 V100 GPUs, while the largest Vision Transformer took 12 days on 256 V100 GPUs"

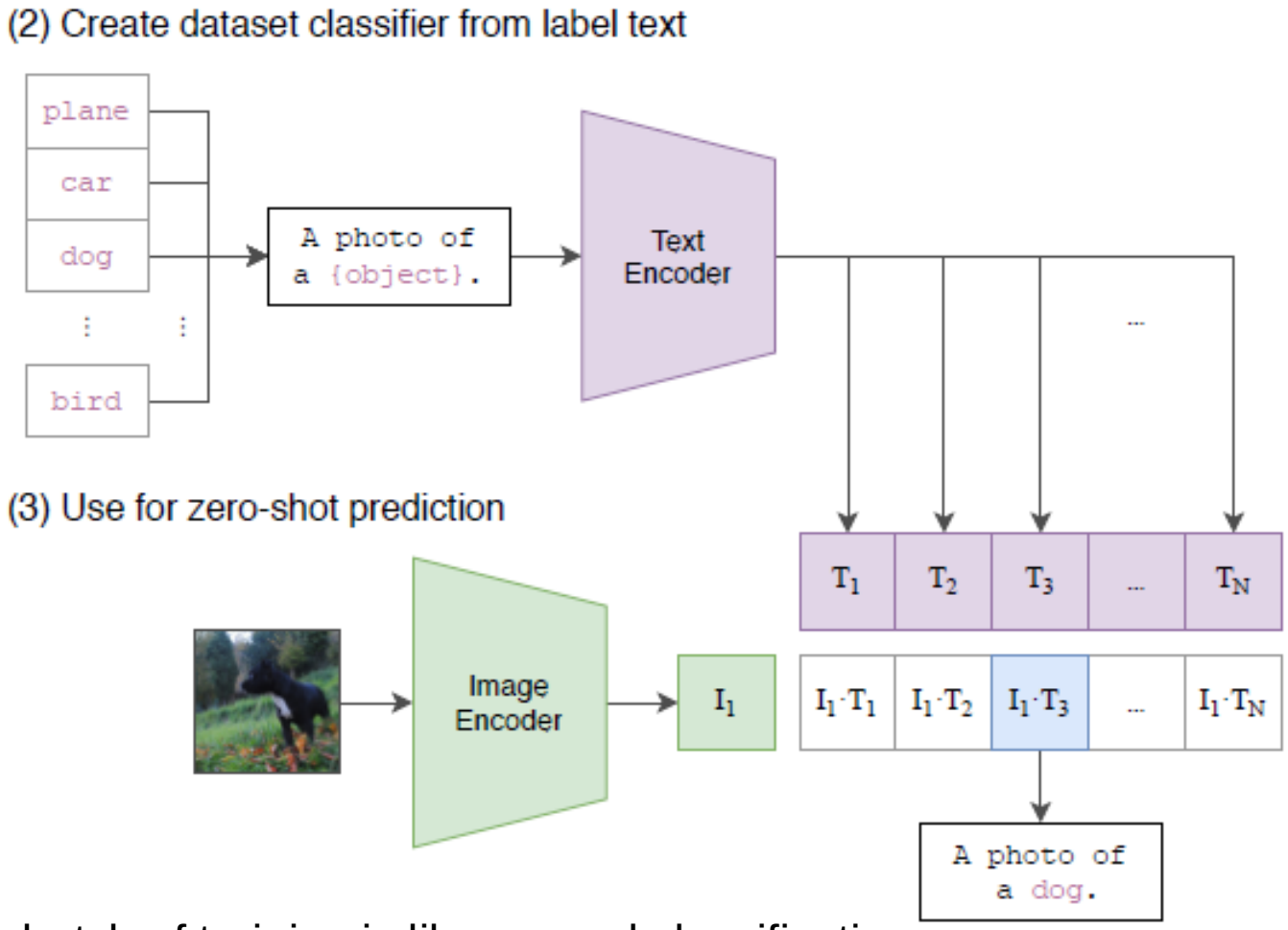


Contrastive formulations is a good general pretrain way to learn when exact target is unpredictable

Key idea 3: **zero-shot classification**

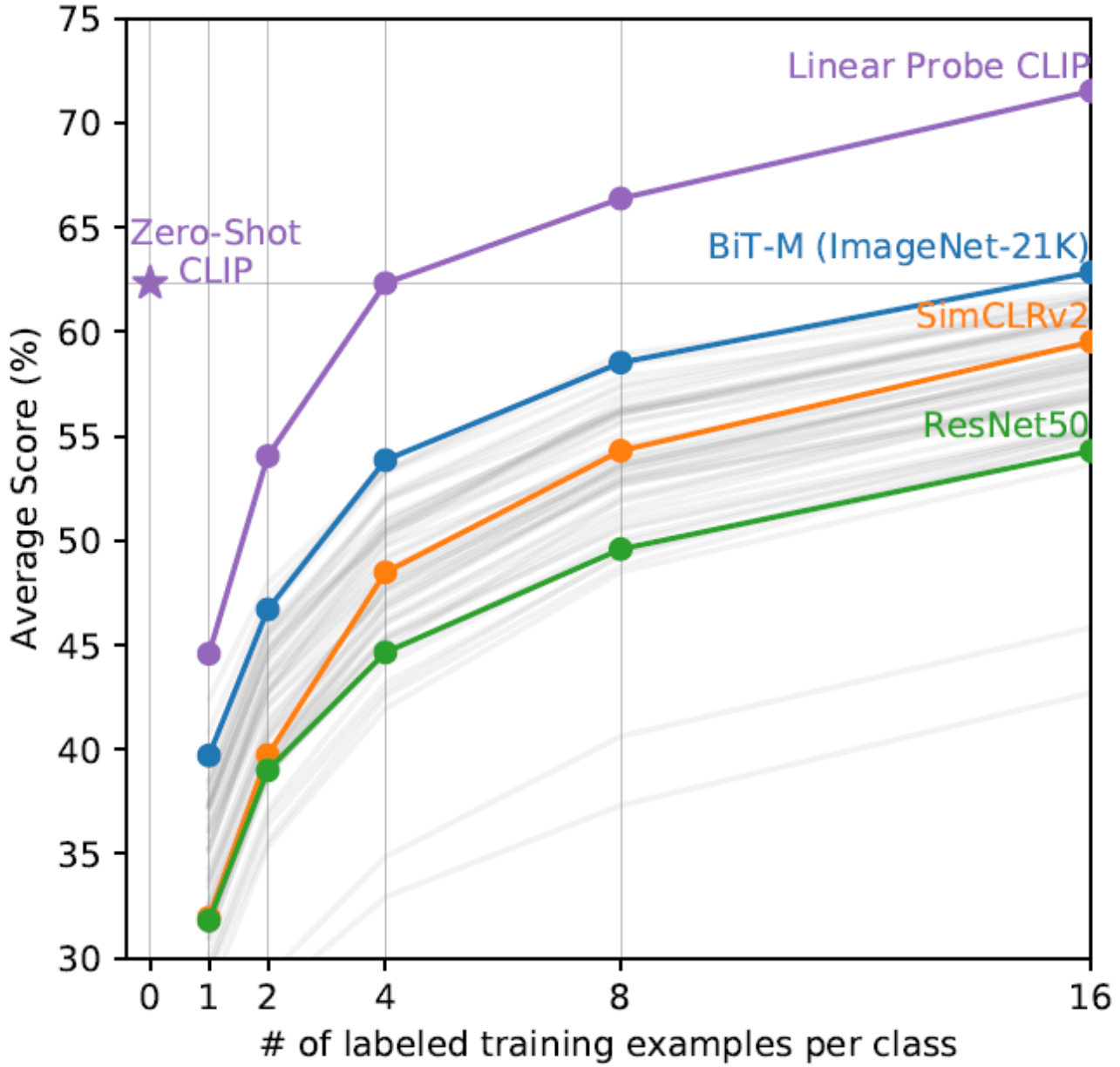
To create a new classification task:

1. Convert class labels into captions and encode the text
2. Encode the image
3. Assign the image to the label whose caption matches best



Every batch of training is like a novel classification task, matching 32K classes to 32K images

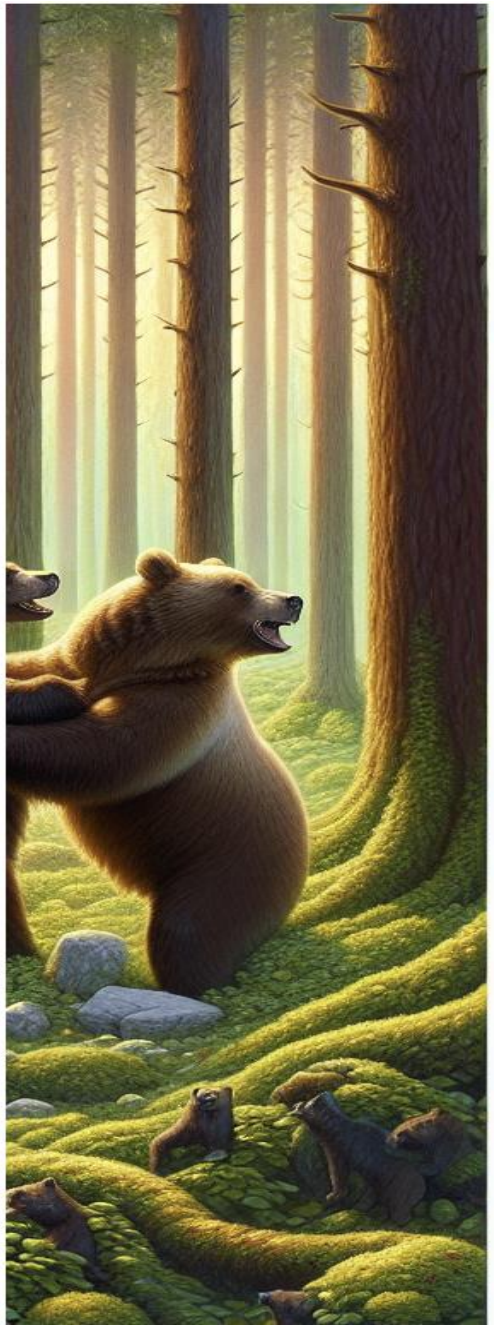
Pretrain learning that match images to text produces a good zero-shot classifier and an excellent image encoder

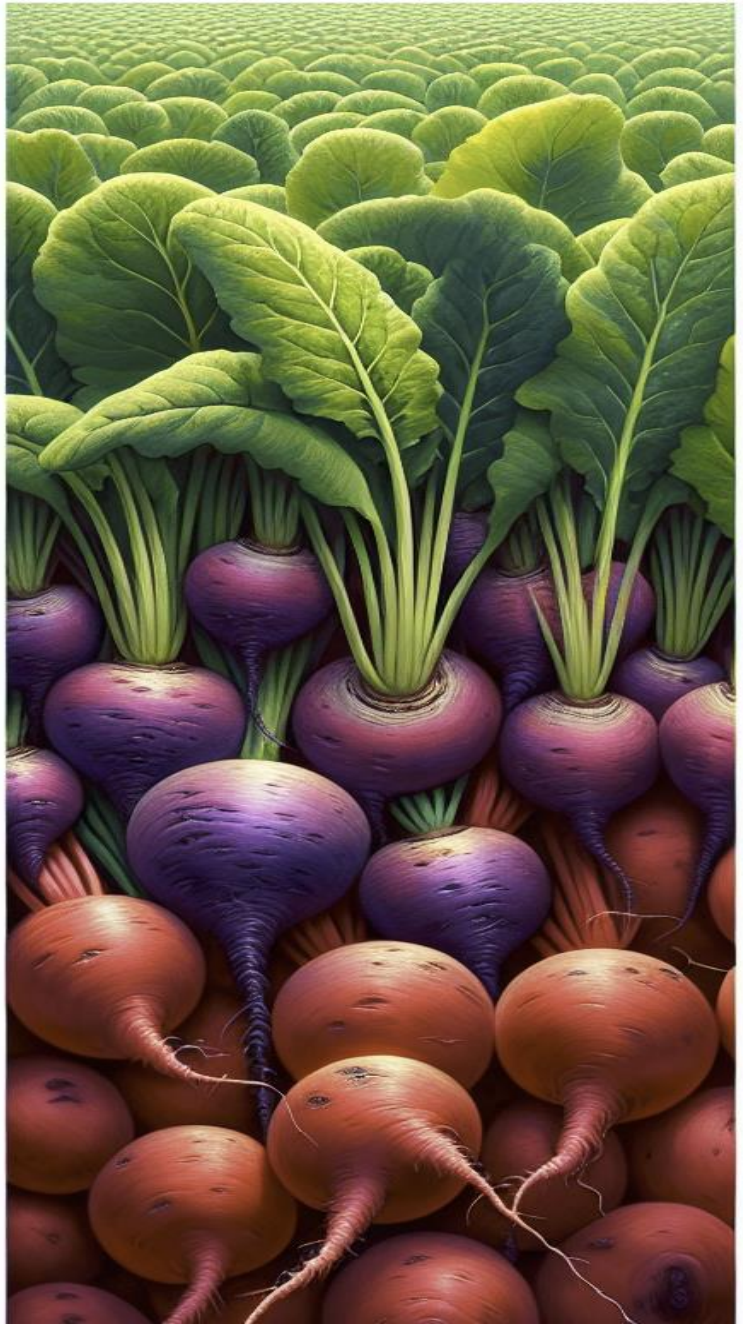


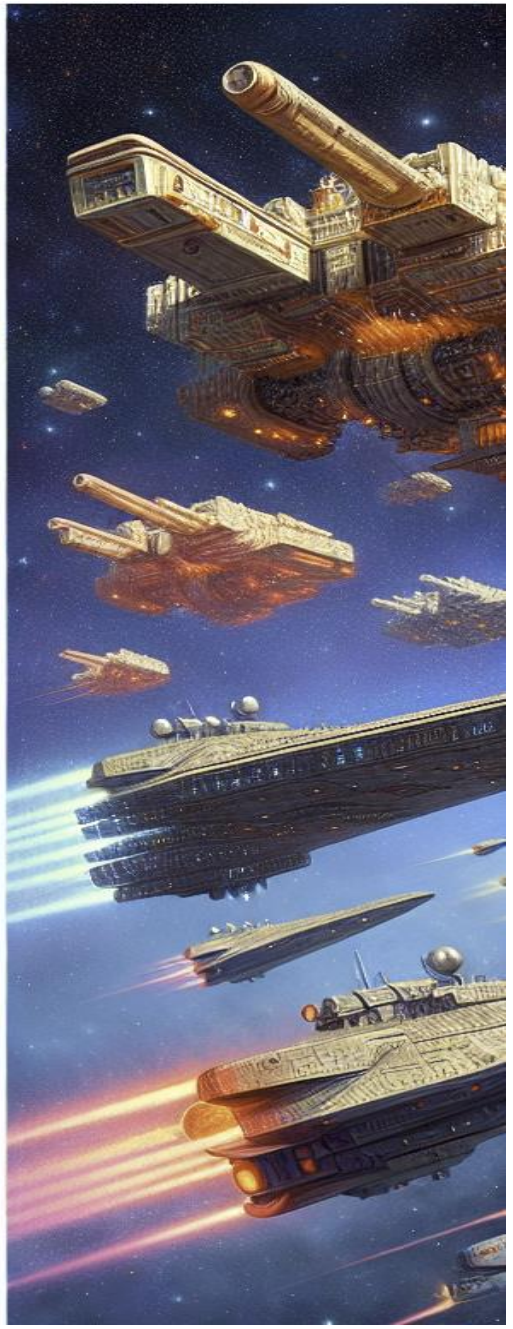
Zero shot to few shot image prediction

openAI DALLE-3 with prompt = "Bears, Beets or Battlestar Galactica”

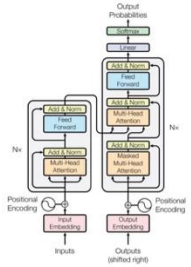
TEXT 2 IMAGE GENERATION

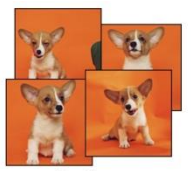






History of Image Generation





GAN

Style-GAN

DALL-E 2

Imagen

DALL-E3

2012 2014 2015 2017 2018 2020 2021 2021 2022 2022 2022 2022 2023

ImageNet w. CNN

DCGAN

Transformer

ViT (VQGAN)

DALL-E

GLIDE

Stable Diffusion

DreamBooth

Input images

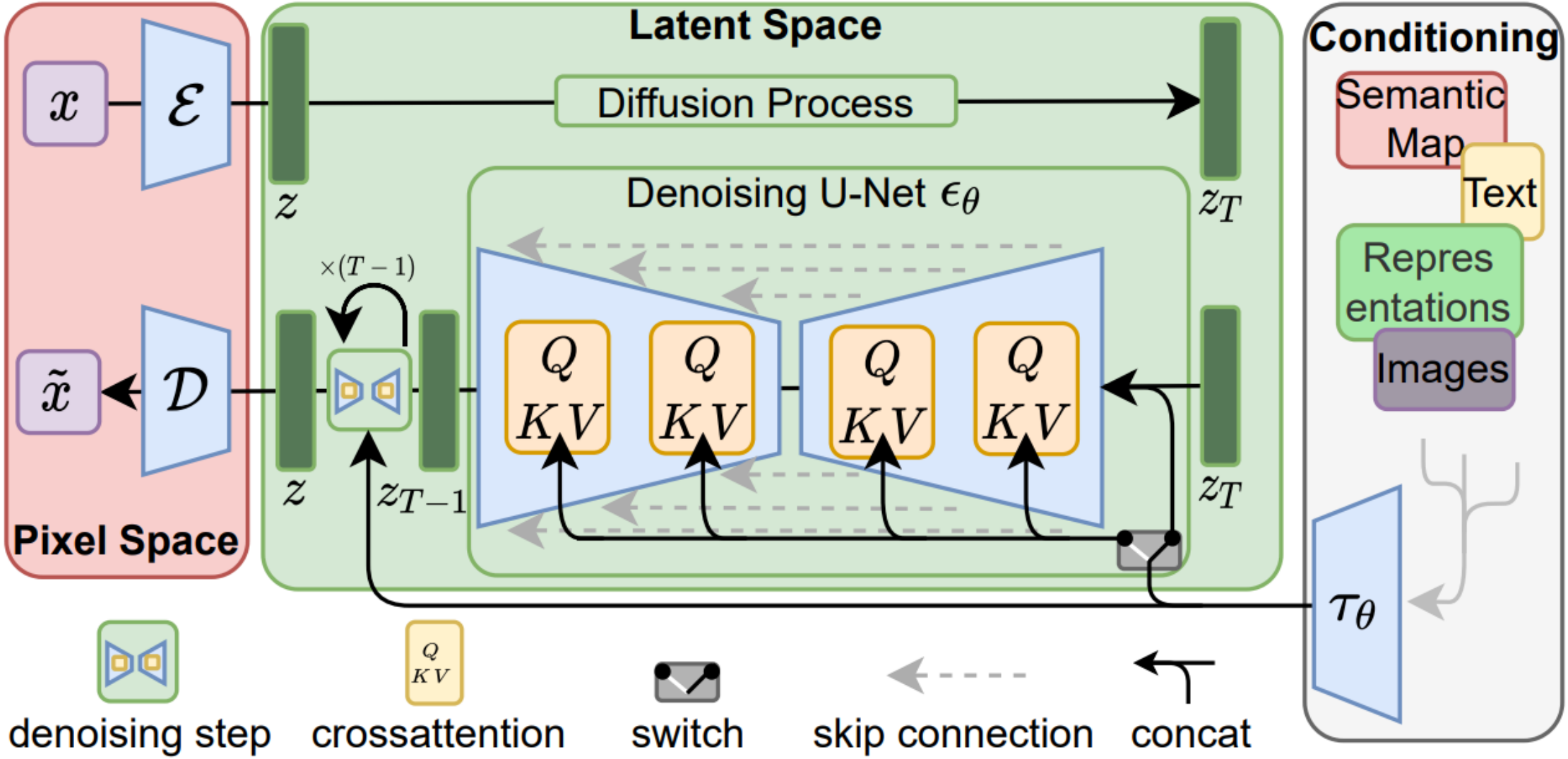
in the Acropolis

Blog: [Ten Years of Image Synthesis]

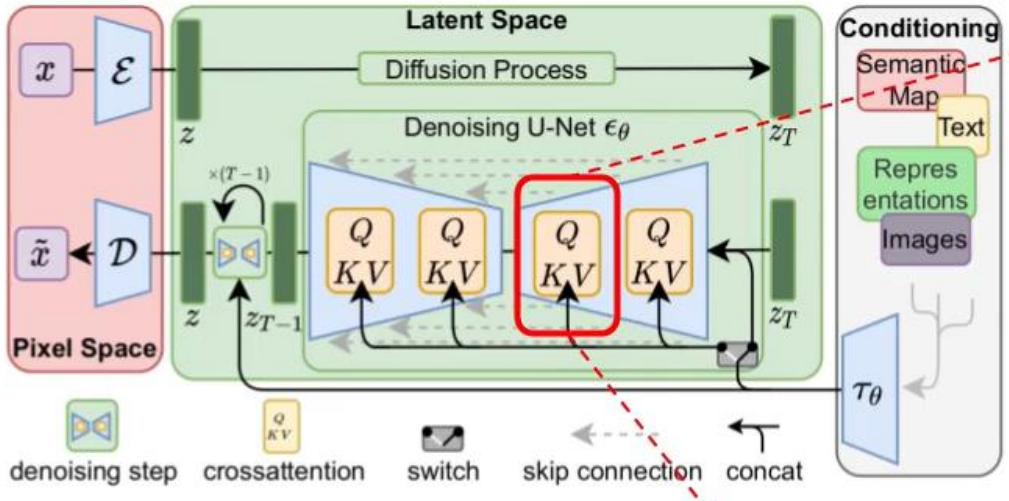
Stable diffusion (2022-08)

- High-Resolution Image Synthesis with Latent Diffusion Models
  CVPR'22 Rombach et al.
- Code and models released Open Source
- The CreativeML OpenRAIL M license
- by Stability AI and Runway

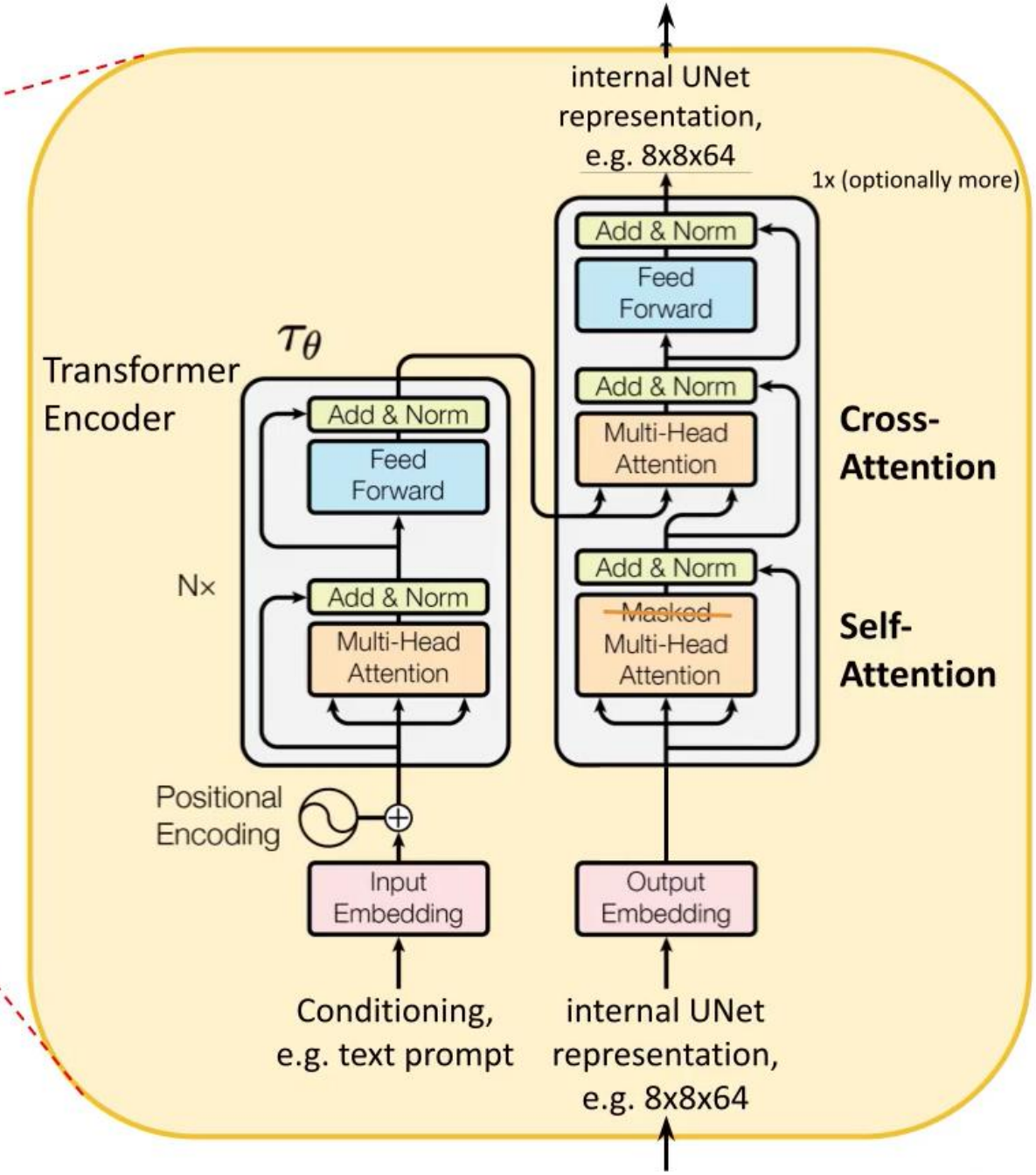
Latent Diffusion Model (LDM)



Paper: [High-Resolution Image Synthesis with Latent Diffusion Models](#) CVPR'22 Rombach et al.



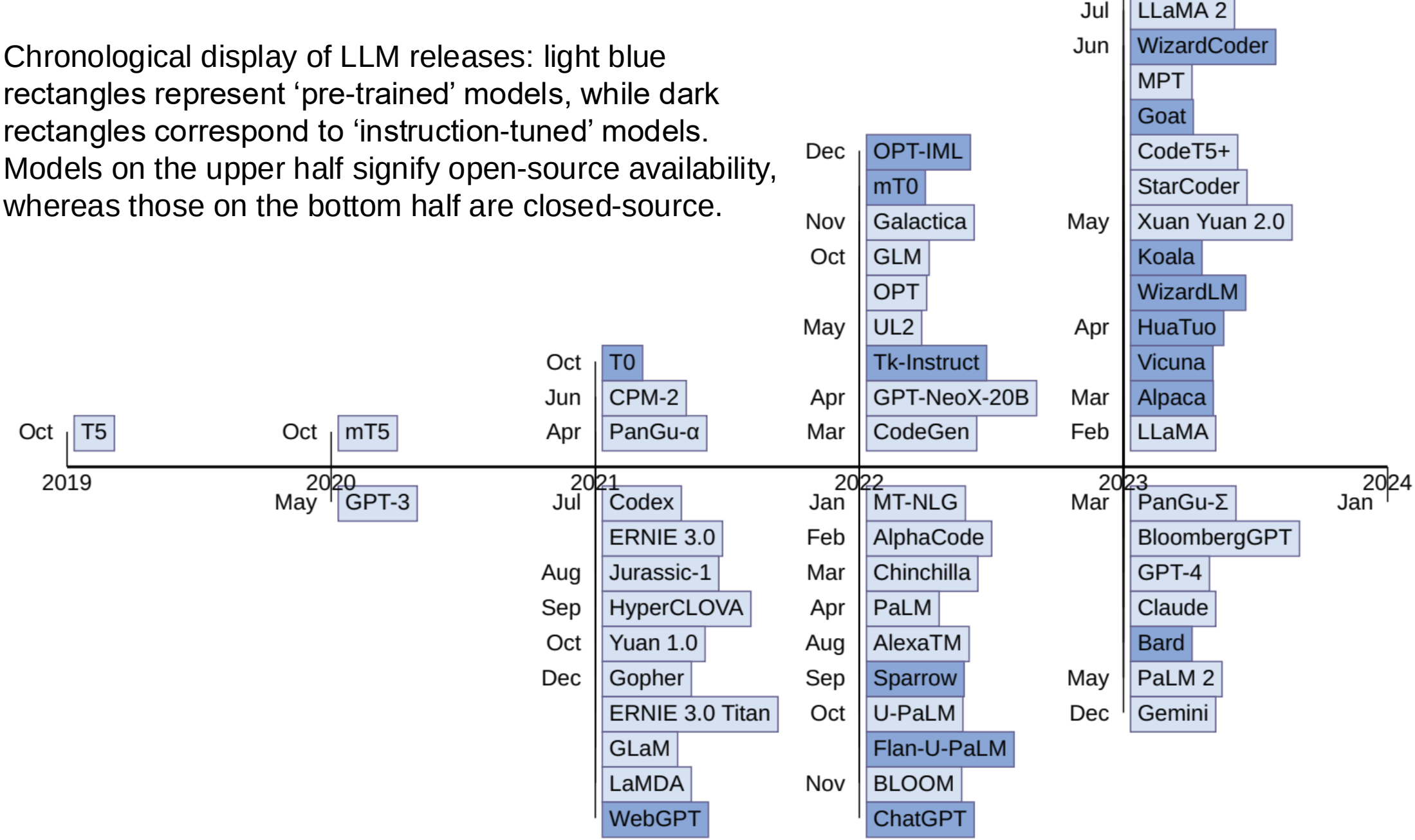
Conditioning on text

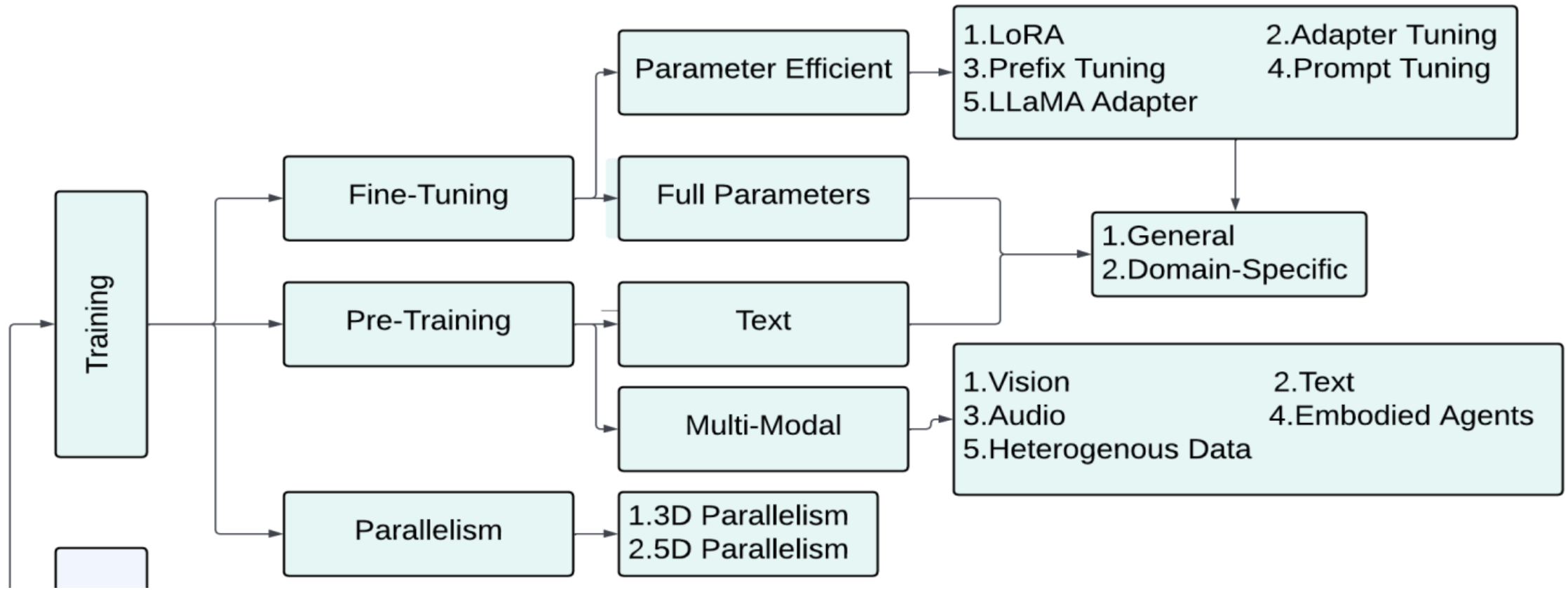


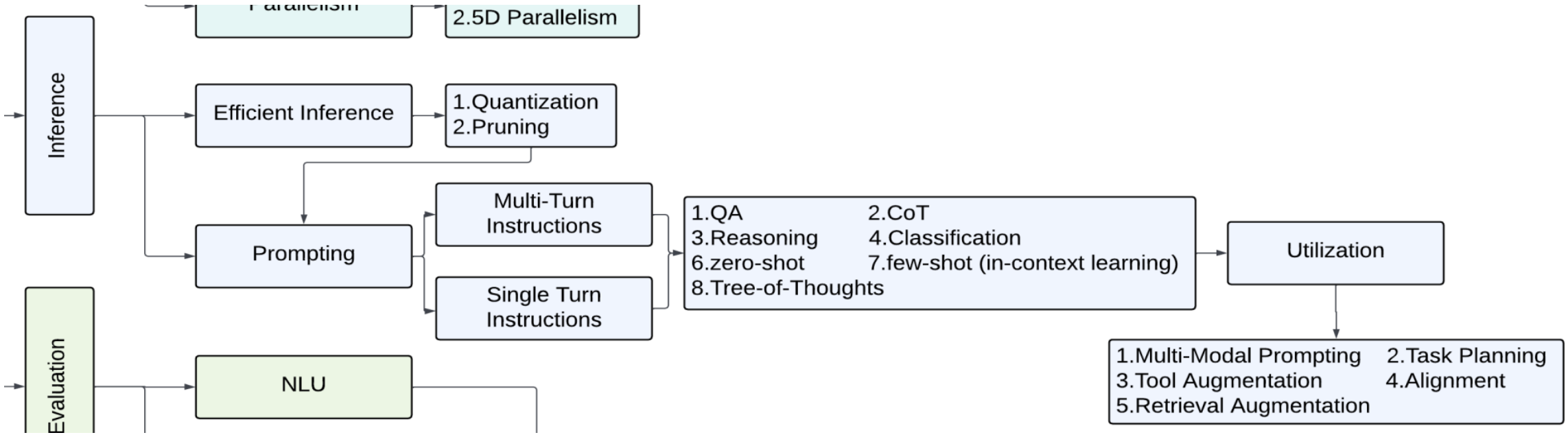
A Comprehensive Overview of Large Language Models

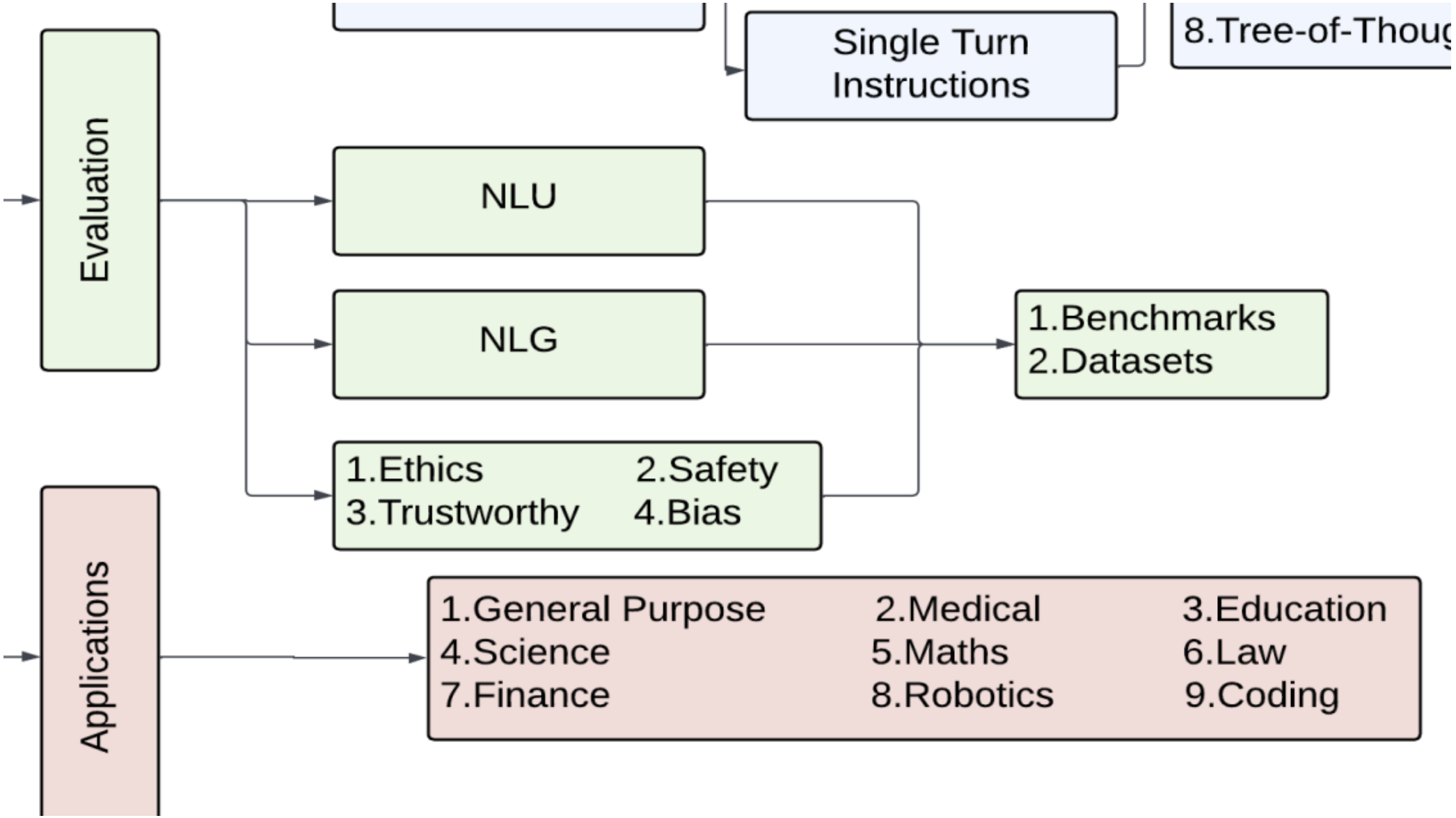
https://arxiv.org/pdf/2307.06435.pdf

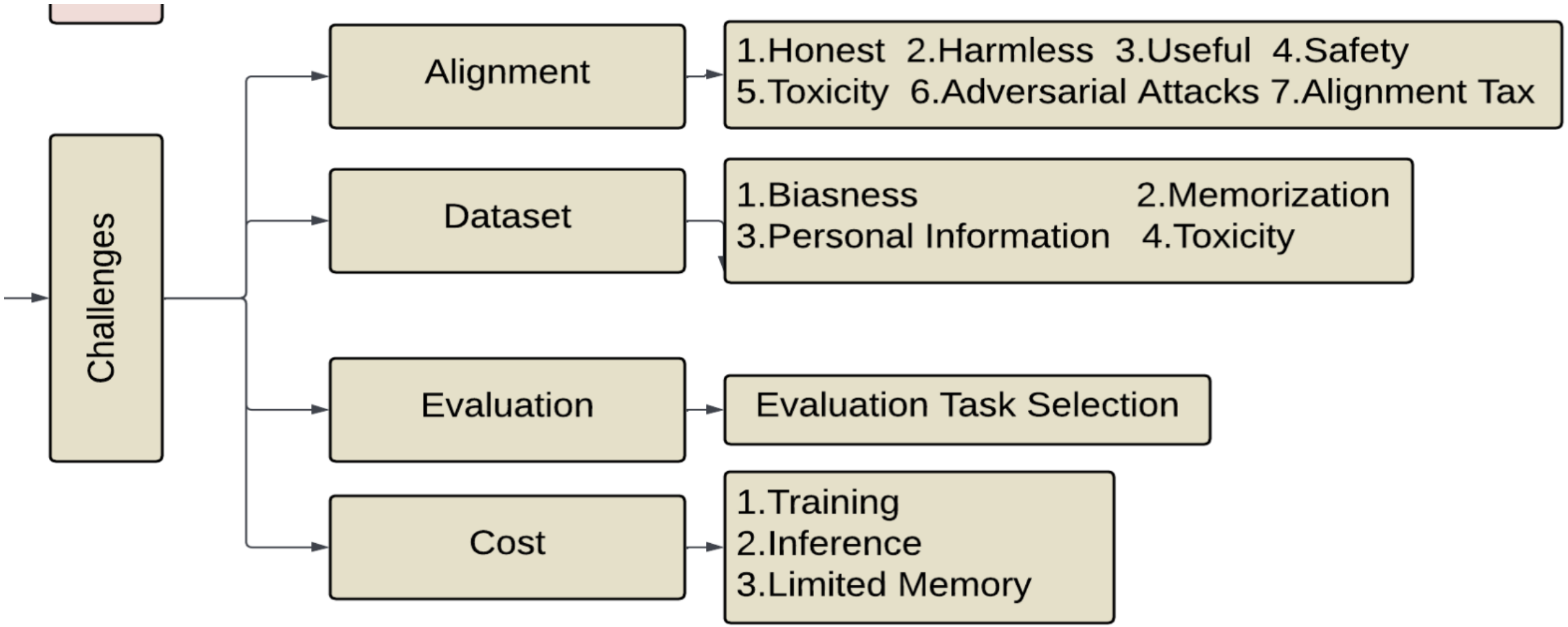
Chronological display of LLM releases: light blue rectangles represent ‘pre-trained’ models, while dark rectangles correspond to ‘instruction-tuned’ models. Models on the upper half signify open-source availability, whereas those on the bottom half are closed-source.




Training
Fine-Tuning
Pre-Training
Parallelism
Parameter Efficient
Full Parameters
Text
Multi-Modal
1.LoRA
2.Adapter Tuning
3.Prefix Tuning
4.Prompt Tuning
5.LLaMA Adapter
1.General
2.Domain-Specific
1.Vision
2.Text
3.Audio
4.Embodied Agents
5.Heterogenous Data
1.3D Parallelism
2.5D Parallelism


2.5D Parallelism
Inference
Efficient Inference
1.Quantization
2.Pruning
Multi-Turn Instructions
Prompting
Single Turn Instructions
1.QA
2.CoT
3.Reasoning
4.Classification
6.zero-shot
7.few-shot (in-context learning)
8.Tree-of-Thoughts
Utilization
1.Multi-Modal Prompting
2.Task Planning
3.Tool Augmentation
4.Alignment
5.Retrieval Augmentation
Evaluation
NLU


Single Turn Instructions
8.Tree-of-Thoug
Evaluation
NLU
NLG
1.Benchmarks
2.Datasets
1.Ethics
2.Safety
3.Trustworthy
4.Bias
Applications
1.General Purpose
2.Medical
3.Education
4.Science
5.Maths
6.Law
7.Finance
8.Robotics
9.Coding


Challenges
Alignment
1.Honest 2.Harmless 3.Useful 4.Safety
5.Toxicity 6.Adversarial Attacks 7.Alignment Tax
Dataset
1.Biasness 2.Memorization
3.Personal Information 4.Toxicity
Evaluation
Evaluation Task Selection
Cost
1.Training
2.Inference
3.Limited Memory