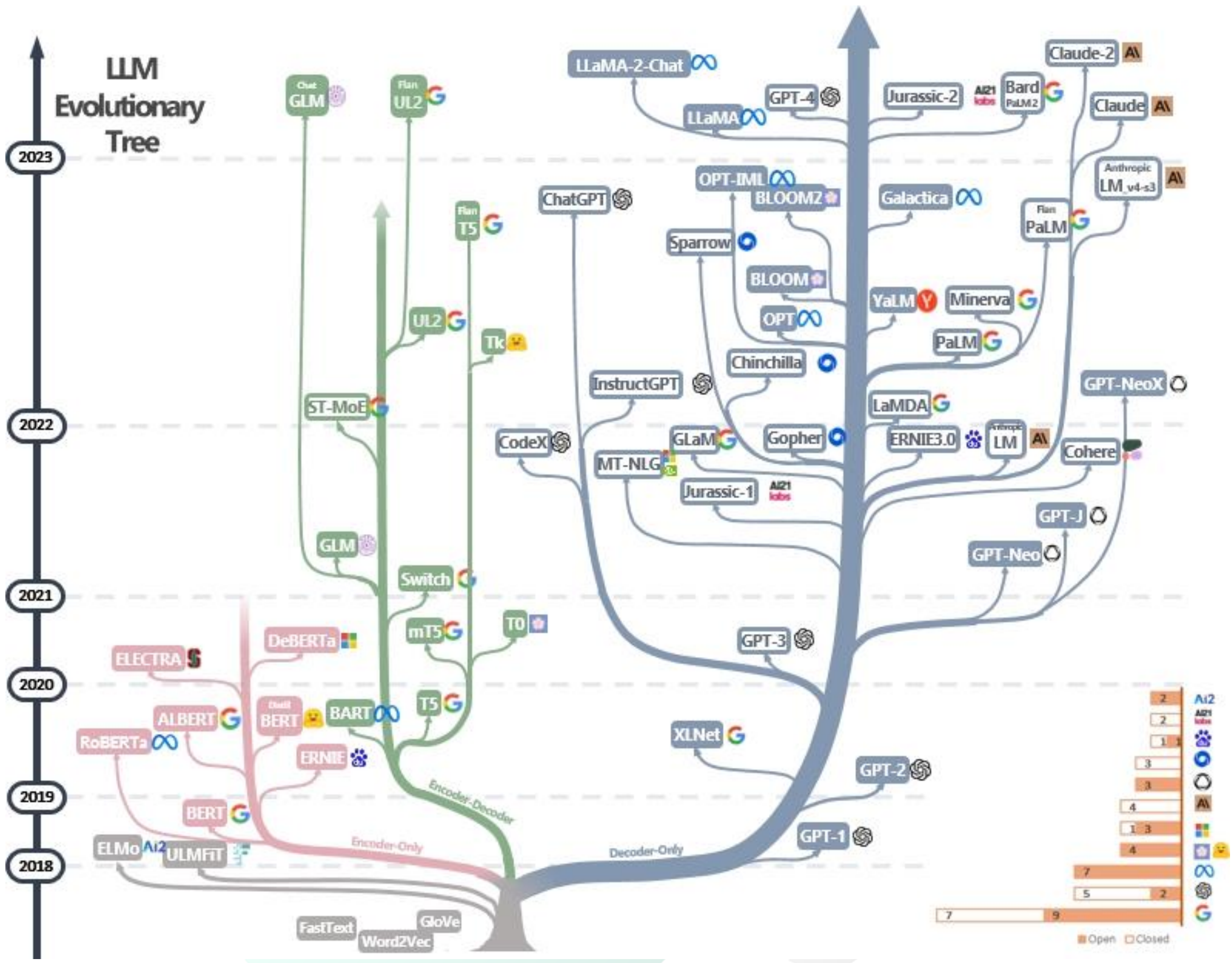

[**Evolution of Foundation LLM models**]

- Closed to open source, their size, their performance, scale, interesting characteristics, pros and cons
- From specialist models to general purpose assistants
- ChatGPT benchmarking paper
- Llama2 LLM

[Understanding Finetuning, RLHF and In-context Learning]

- Different types of finetuning mechanisms and their examples in LLMs
- What **In-Context Learning** "Learns" In-Context: Disentangling Task Recognition and Task Learning

- ICL reference
- In-context Learning and **Induction Heads**
- Finetuning with human preference
- RLHF (paper-pdf, InstructGPT), RLHF - base paper
- Prompt functions, Program aided LLMs
- Some more papers:
- Paper 1, Integrating human feedback in RL

[Walkthrough Prompting Techniques]
What is it? What are the different ways? Is there any best/general way that is better than others? Cover the most important ones in the permitted time.

- **Prompting**
- **RAG,** other paper
- **CoT** - Chain-of-Thought Prompting Elicits Reasoning in Large Language Models
    - CoT collection paper
- **Tree of Thoughts**: Deliberate Problem Solving with Large Language Models
- Multimodal CoT
- Self consistency
- Auto - prompting
- **Zero-shot** prompting
- ReACT
- **Active Prompting** with Chain-of-Thought for Large Language Models
- **GraphPrompt**: Unifying Pre-Training and Downstream Tasks for Graph Neural Networks

[**Alignment, Interpretability and Robustness in LLMs**]

- Alignment problem in LLMs
- Ethics, toxicity in LLMs and role of prompting / finetuning
- Automated interpretability
- Attention visualization: using dimensionality reduction to visualize the joint embedding space of key-query pairs
- Robustness & **adversarial** prompting
- Zero-Resource Hallucination Prevention for Large Language Models
- Certifying LLM Safety against Adversarial Prompting
- Improving Code Generation by Dynamic Temperature Sampling
- Limitations and challenges - blog
- Some more papers:
- Paper 1 on reasoning hallucinations,

[**Self-attention and improvements in terms of speed**]

- Multi-head self-attention: from self attention to its hardware level improvements:
- Flash attention, paged attention: based on reducing the IO in GPU's HBM and on-chip SRAM. Also improves the approximate block-sparse attention.

[**Distributed large scale training of LLMs and associated challenges**]

- Discrimination between models based on how they were trained: can take up models which differ in their training strategy and may discuss the differences. Related Survey
- Training spikes and divergences: use this as the starting point of your exploration of how large scale training converges
- ZeRO-fashion data parallel (distributed optimizer), Model parallel

[**Concept of vector database and LLM application dev. tools like Langchain**]

- Can discuss about performance, scalability, and flexibility in vector database
- Focus on pinecone
- Other database:, Chroma, Weaviate, Milvus, etc.
- Term vector database paper to get some idea on early vector databases
- Application development using LLMs:
- Langchain: framework for developing applications powered by LLMs. Agents ←use Tools. Memory ← to integrate and remember contexts.
- Other frameworks: FlowiseAI, AutoGPT, AgentGPT, BabyAGI, Langdock, GradientJ, LlamaIndex, MetaGPT

[**Parameter efficient finetuning and its application to LLMs**]

- Adapters - Parameter efficient fine tuning
- Prefix tunning
- Low Rank Adaptation of LLMs (LoRA)
- How does it compare with Prompting / In context learning? Is there any study or observation from literature?