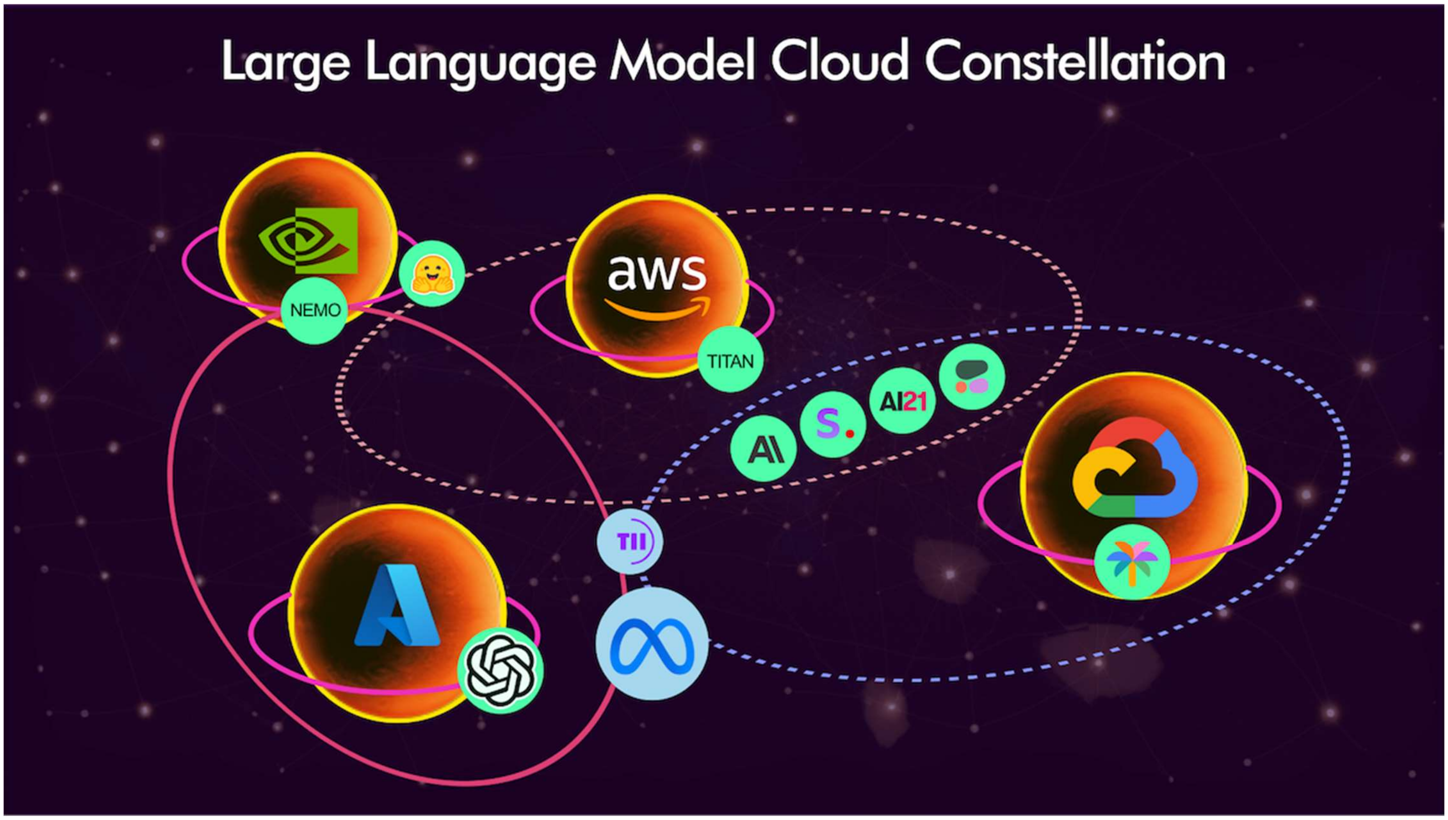
How the Cloud Hyperscalers are Shaping the Generative AI Market

LLMs are the actors on the stage, but the cloud providers are the theaters of AI



Large language models (LLM) or foundation models, if you prefer, receive a lot of attention today. On the surface, it seems like a pitched battle between LLM providers to capture new users and use cases. That is true, to a degree. However, another way to look at the market is that the LLM wars exist in the larger context of the cloud computing wars.

Cloud Hyperscaler Generative AI Strategies

The three cloud-computing LLM ecosystems are depicted in the diagram above. There is little argument that OpenAI is the leader in LLM use and mindshare today. A recent survey by Voicebot of developers in the conversational AI industry showed that over 90% of those who had used an LLM API had at least some experience with OpenAI. The next most tried LLM was identified by less than 50% of the developers.

Azure is the only leading cloud provider that offers OpenAI today, and it is also the cloud that has the fewest third-party options today. AWS started with a strategy of featuring multiple LLMs from third-parties alongside its Tritan LLM in the Bedrock offering. It is a strategy designed to be a counterbalance to Azure's OpenAI approach by providing more user choice.

Google began by focusing on its own in-house LLM PaLM and other foundation models. However, it has quickly added third-party offerings to match AWS's selection. Azure is likely to add more third-party LLM offerings over time, but it is clear the first priority today is to serve the significant demand for OpenAI Services.

Meta's Middle Road

A prominent third-party option that Azure does provide is Meta's Llama 2. Meta has quickly become a popular consideration for an open-source foundation model. To the surprise of many, Google Cloud also recently announced the availability of Llama 2. Every cloud provider knows they need to offer open-source options and Meta is going to become a standard. The company has announced it will also be available on AWS.

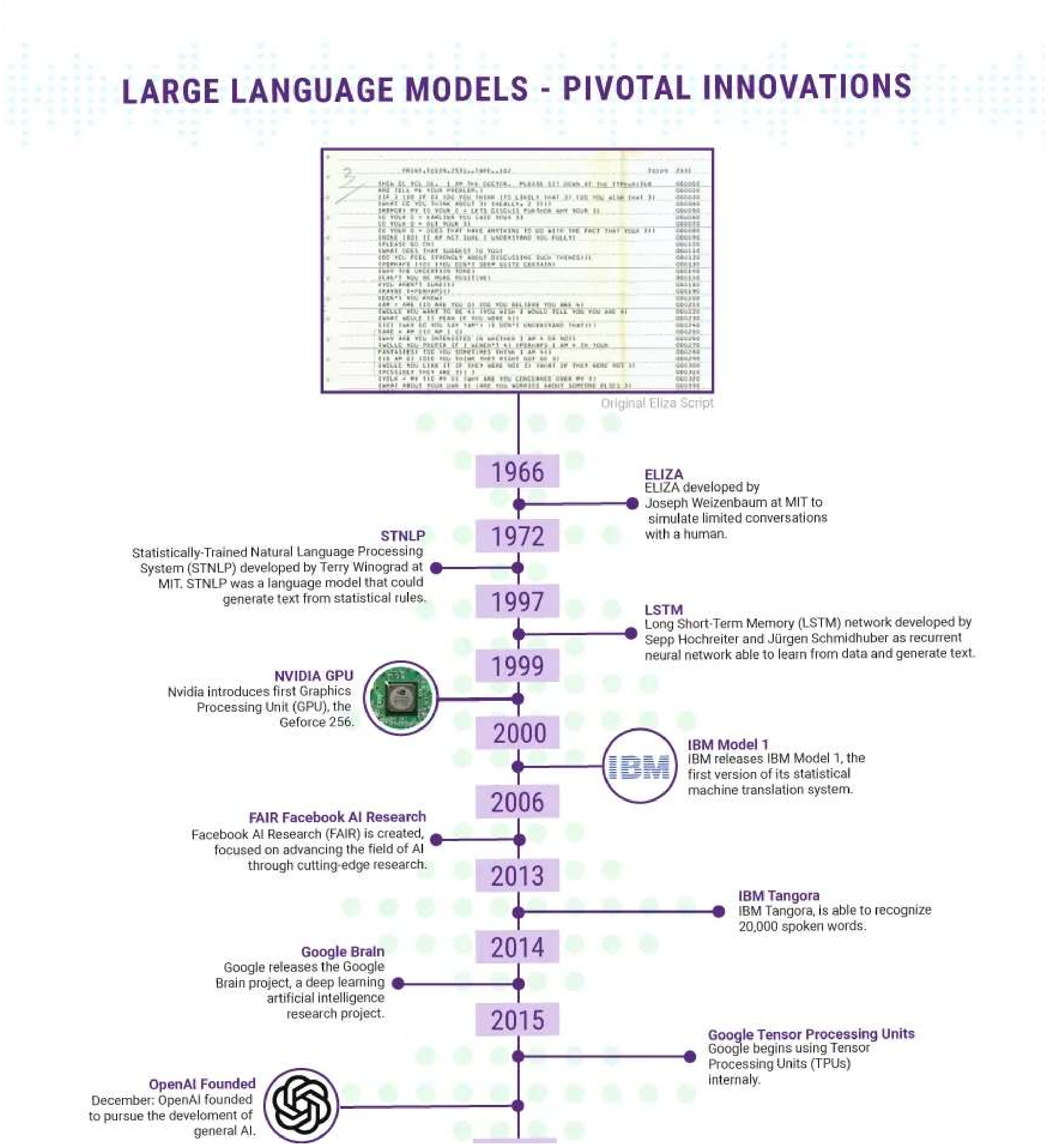
While other tech giants offer proprietary models, Meta has chosen to provide Llama 2 as open-source and is not looking to generate revenue from the foundation model. Meta would instead like to see a large developer community organize around Llama 2 and ultimately provide greater interoperability and compatibility of LLM-backed features. However, Meta understands that the primary route for enterprises to access LLMs will be through their preferred cloud providers. Meta may eventually become the leading LLM choice overall as the second most popular choice in each Cloud.

LLM Distribution

Cloud providers have emerged as the key distribution channels for LLMs. What they provide determines the selection set for enterprises. LLMs that cannot secure distribution through the leading cloud hyperscalers will have difficultly driving adoption. It is not to say they cannot be successful. The path will just include more significant barriers.

The key takeaway here is that you can understand a lot about how the LLM market is developing by observing the actions of the leading cloud providers. They are deciding which LLMs will be the easiest to adopt. They are offering discounted foundation model training for partners. The cloud providers know that LLMs drive a lot of compute processing both for training and inference. They each want to maximize how much of that flows through their servers. LLMs are just the latest technology venue for them to play out cloud competition.

A Timeline of Large Language Model Innovation



LARGE LANGUAGE MODELS - PIVOTAL INNOVATIONS

Original Eliza Script

1966

ELIZA
ELIZA developed by Joseph Weizenbaum at MIT to simulate limited conversations with a human.

1972

STNLP
Statistically-Trained Natural Language Processing System (STNLP) developed by Terry Winograd at MIT. STNLP was a language model that could generate text from statistical rules.

1997

LSTM
Long Short-Term Memory (LSTM) network developed by Sepp Hochreiter and Jürgen Schmidhuber as recurrent neural network able to learn from data and generate text.

1999

NVIDIA GPU
Nvidia introduces first Graphics Processing Unit (GPU), the Geforce 256.

2000

IBM Model 1
IBM releases IBM Model 1, the first version of its statistical machine translation system.

2006

FAIR Facebook AI Research
Facebook AI Research (FAIR) is created, focused on advancing the field of AI through cutting-edge research.

2013

IBM Tangora
IBM Tangora, Is able to recognize 20,000 spoken words.

2014

Google Brain
Google releases the Google Brain project, a deep learning artificial intelligence research project.

2015

Google Tensor Processing Units
Google begins using Tensor Processing Units (TPUs) internaly.

OpenAI Founded
December: OpenAI founded to pursue the develoment of general AI.

2016

Stanford SQuAD
Stanford University's NLP Group releases the Stanford Question Answering Dataset (SQuAD), a dataset for NLP research.

2017

Introduction of Transformer Models
Transformer Models are introduced through papers like Googles Transformer: A Novel Neural Network Architecture for Language Understanding and Attention Is All You Need, Vaswani et al., 2017.

Elmo
October: Allen Institute for AI researchers develop Embeddings from Language Models (Elmo) trained on 5.5B words and 1B parameters.

Tensor Flow Framework
November: Google releases its TensorFlow framework, which allows developers to build and train their own AI models.

Universal Sentence Encoder
December: The Universal Sentence Encoder, a pre-trained model for encoding sentence pairs, is introduced by Google.

2018

Google Tensor Processing Units
Google Makes TPUs availabe to third parties.

Open AI Generative Pre-Trained Transformer
June: OpenAI publishes paper on Generative Pre-Trained Transformer (GPT) model for natural language processing with 110 million parameters, setting the foundation for future generative AI models

Google Introduces BERT
October: BERT-Large, a pre-trained transformer model for natural language processing, is introduced by Google.

2019

OpenAI GPT-2
February: OpenAI releases GPT-2, Trained on 1.6 Billion parameters GPT-2 demonstrates the ability to generate coherent and human-like text.

Baidu ERNIE
March: Baidu develops Enhanced Representation through kNowledge IntEgration (ERNIE) as a large-scale pre-trained language model, with a capacity of 4.5 billion parameters.

XLNet
June: XLNet released, beating BERT model with pre-training on a dataset of over 1 billion words with 570 million parameters.

Microsoft
Microsoft invests $1 billion into OpenAI.

Microsoft LXMERT
August: Microsoft Research announces its new language model, LXMERT, which uses pre-training on large-scale tasks and fine-tuning on target tasks.

Facebook RoBERTa
August: RoBERTa, a robustly optimized BERT model, is introduced by Facebook AI.

Amazon DeepComposer
December: Amazon Web Services introduces the DeepComposer, a language model that generates music based on text input.

2020

OpenAI GPT-3
May: OpenAI releases GPT-3, the largest language model to date with 175 billion parameters.

AI21 Wordtune
AI21 Wordtune generative text AI tool released.

Nvidia Megatron
October: Nvidia/Microsoft Megatron language model launches.

2021

Google T5-xxl
January: Google trains T5-xxl with over 1.6 trillion parameters.

OpenAI DALL-E
January: OpenAI releases DALL-E a 12-billion parameter version of GPT-3 trained to generate images from text descriptions.

Google Transformer-X
February: Google releases its next-generation language model Transformer-X 7.5x faster than base T5.

EleutherAI GPT-Neo
March: EleutherAI releases its first open-source GPT-Neo LLMs.

Facebook XLM-R
April: The XLM-R language model, introduced by Facebook AI, is released.

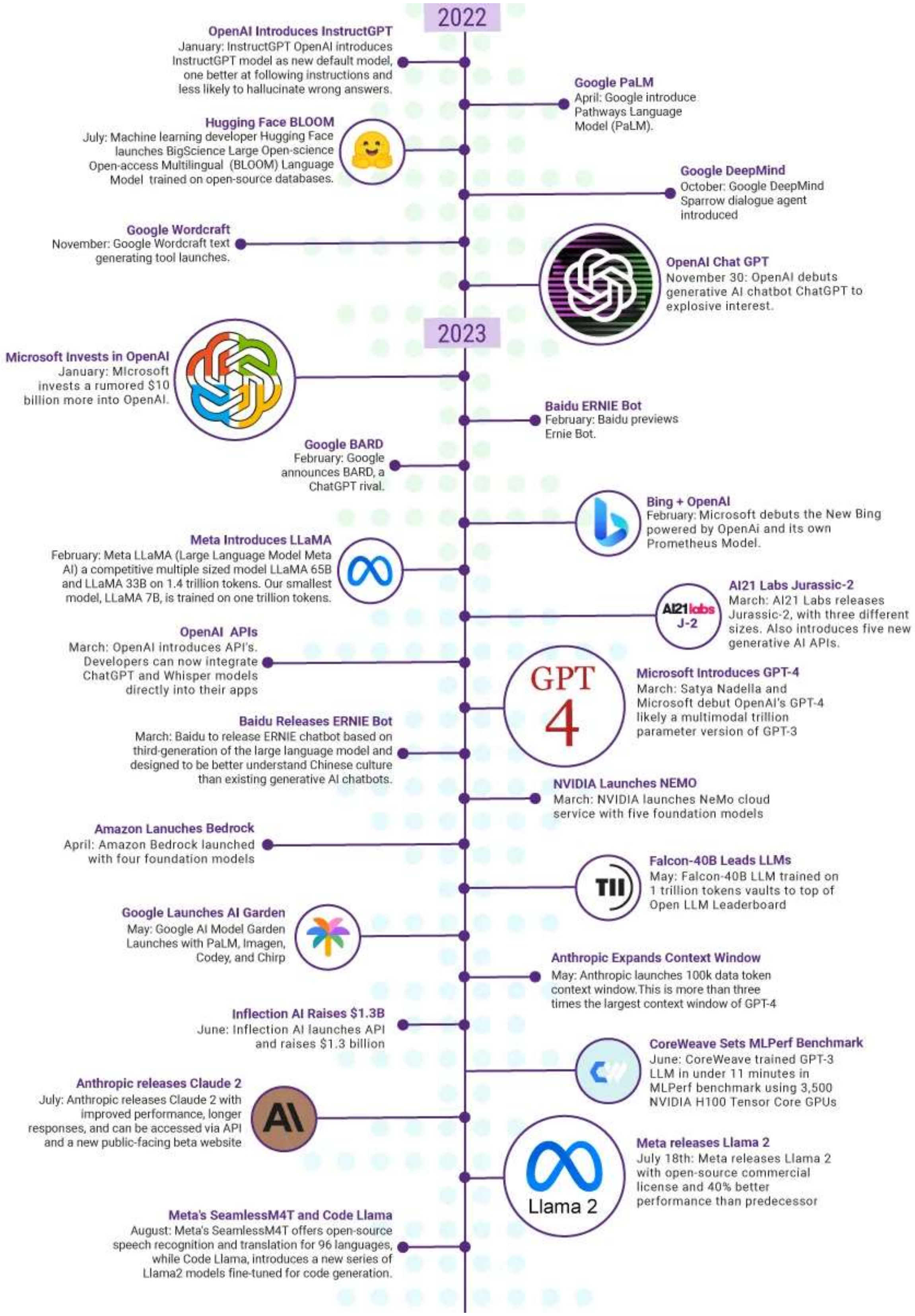
Google LaMDA
May: Google introduces Language Model for Dialogue Applications (LaMDA) neural language models.

Github Copilot
June: GitHub announces Copilot, an AI pair-programmer for coding.

OpenAI Codex
July: OpenAI launches Codex, which translates natural language into code.

AI21 Labs Releases Jurassic-1
August: AI21 Labs releases Jurassic-1 large language model in two sizes as well as the AI21 Studio developer platform for creating generative AI tools and apps.

OpenAI Releases WebGPT
December: OpenAI releases WebGPT, a fine-tuned GPT-3 model for web browsers that can answer open-ended questions with citations and links to sources.



2022

OpenAI Introduces InstructGPT
January: InstructGPT OpenAI introduces InstructGPT model as new default model, one better at following instructions and less likely to hallucinate wrong answers.

Google PaLM
April: Google introduce Pathways Language Model (PaLM).

Hugging Face BLOOM
July: Machine learning developer Hugging Face launches BigScience Large Open-science Open-access Multilingual (BLOOM) Language Model trained on open-source databases.

Google DeepMind
October: Google DeepMind Sparrow dialogue agent introduced

Google Wordcraft
November: Google Wordcraft text generating tool launches.

OpenAI Chat GPT
November 30: OpenAI debuts generative AI chatbot ChatGPT to explosive interest.

2023

Microsoft Invests in OpenAI
January: MIcrosoft invests a rumored $10 billion more into OpenAI.

Baidu ERNIE Bot
February: Baidu previews Ernie Bot.

Google BARD
February: Google announces BARD, a ChatGPT rival.

Bing + OpenAI
February: Microsoft debuts the New Bing powered by OpenAi and its own Prometheus Model.

Meta Introduces LLaMA
February: Meta LLaMA (Large Language Model Meta AI) a competitive multiple sized model LLaMA 65B and LLaMA 33B on 1.4 trillion tokens. Our smallest model, LLaMA 7B, is trained on one trillion tokens.

AI21 Labs Jurassic-2
March: AI21 Labs releases Jurassic-2, with three different sizes. Also introduces five new generative AI APIs.

OpenAI APIs
March: OpenAI introduces API's. Developers can now integrate ChatGPT and Whisper models directly into their apps

Microsoft Introduces GPT-4
March: Satya Nadella and Microsoft debut OpenAI's GPT-4 likely a multimodal trillion parameter version of GPT-3

Baidu Releases ERNIE Bot
March: Baidu to release ERNIE chatbot based on third-generation of the large language model and designed to be better understand Chinese culture than existing generative AI chatbots.

NVIDIA Launches NEMO
March: NVIDIA launches NeMo cloud service with five foundation models

Amazon Lanuches Bedrock
April: Amazon Bedrock launched with four foundation models

Falcon-40B Leads LLMs
May: Falcon-40B LLM trained on 1 trillion tokens vaults to top of Open LLM Leaderboard

Google Launches AI Garden
May: Google AI Model Garden Launches with PaLM, Imagen, Codey, and Chirp

Anthropic Expands Context Window
May: Anthropic launches 100k data token context window.This is more than three times the largest context window of GPT-4

Inflection AI Raises $1.3B
June: Inflection AI launches API and raises $1.3 billion

CoreWeave Sets MLPerf Benchmark
June: CoreWeave trained GPT-3 LLM in under 11 minutes in MLPerf benchmark using 3,500 NVIDIA H100 Tensor Core GPUs

Anthropic releases Claude 2
July: Anthropic releases Claude 2 with improved performance, longer responses, and can be accessed via API and a new public-facing beta website

Meta releases Llama 2
July 18th: Meta releases Llama 2 with open-source commercial license and 40% better performance than predecessor

Meta's SeamlessM4T and Code Llama
August: Meta's SeamlessM4T offers open-source speech recognition and translation for 96 languages, while Code Llama, introduces a new series of Llama2 models fine-tuned for code generation.

Large language models (LLM) are the workhorse behind much of the generative AI frenzy, but not everyone followed the development progress before GPT-3 in 2020 or even ChatGPT in 2022.

Source: https://synthedia.substack.com/p/how-the-cloud-hyperscalers-are-shaping