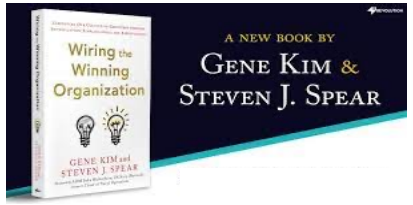
Frontiers of Generative AI

DEVOPS ENTERPRISE SUMMIT

A Shared Passion for AI

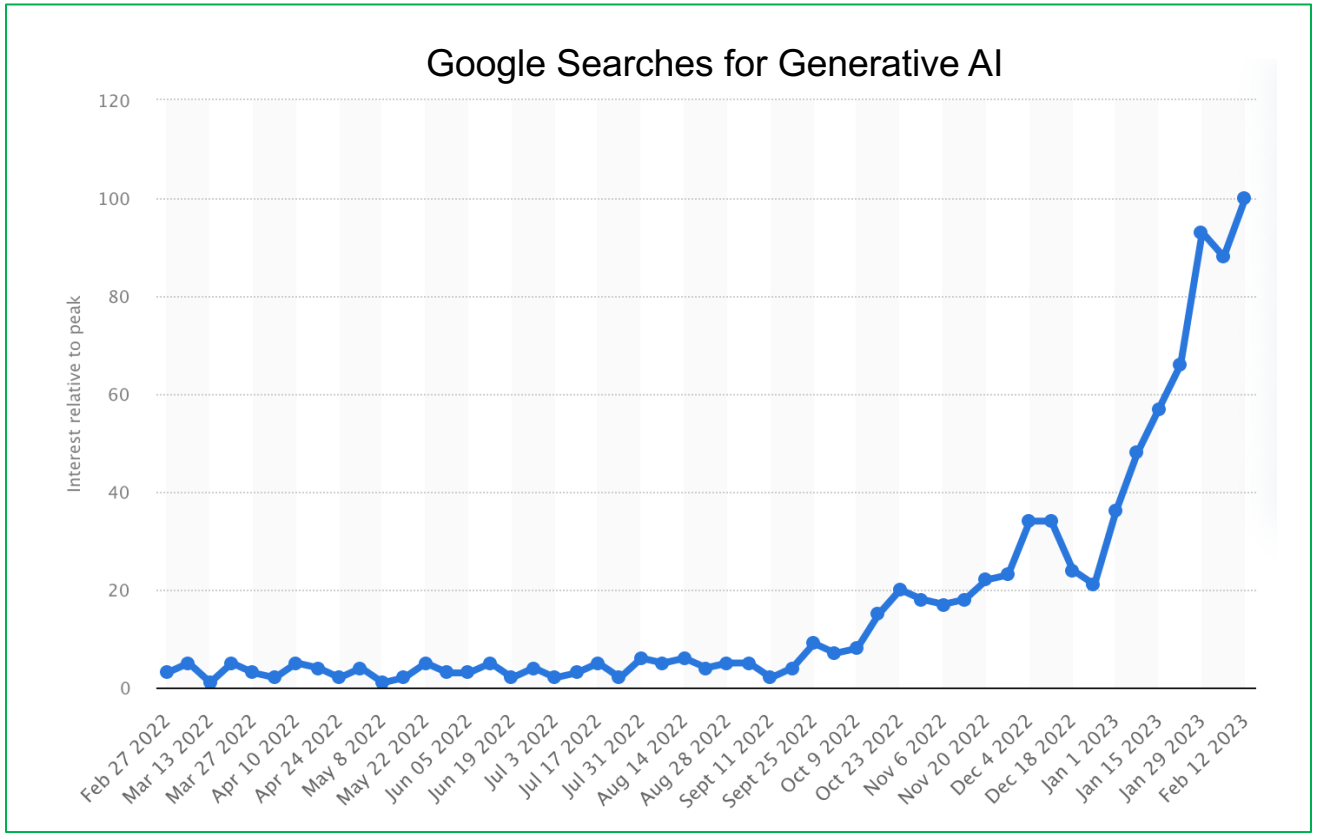
DEVOPS ENTERPRISE SUMMIT



The Growing Wave of AI Enthusiasm

DEVOPS ENTERPRISE SUMMIT

"Explain it to me like I'm a
OLD SCHOOL STORAGE GUY!"



A Brief Introduction

DEVOPS ENTERPRISE SUMMIT

EVT Why

*Be of Service to our Customers to **Enrich the Human Experience** every day.*

*We first **Understand our Customer's WHY** so that we can systematically understand the Challenges and Problems they are trying to solve and Partner with them to solve them.*

By adding value through People, Services, and Technologies

Dr. W. Edwards Deming, "**Every system is perfectly designed to get the result that it does.**"

PEOPLE + PROCESS + TECHNOLOGY

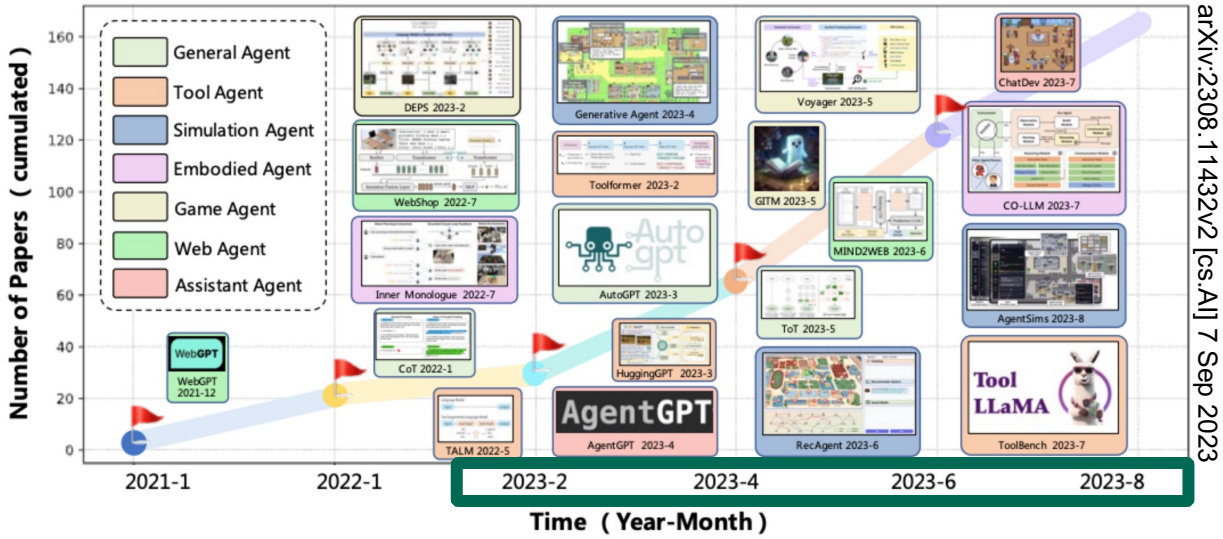
= RESULTS

ENTERPRISEVISIONTECHNOLOGIES

Accelerating Developments in AI

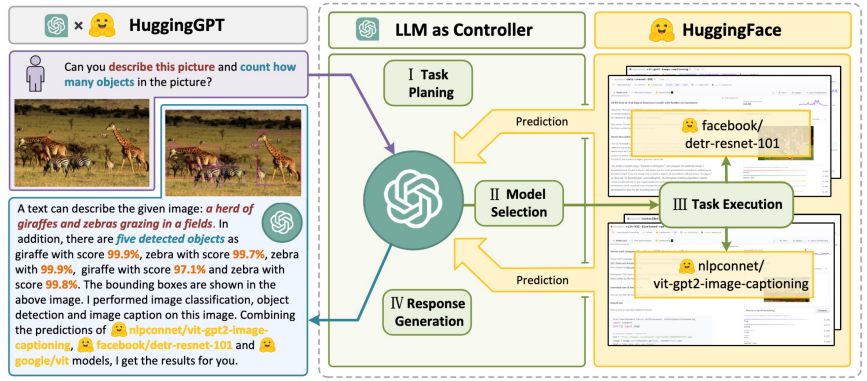
DEVOPS ENTERPRISE SUMMIT

LLM-based Autonomous Agents



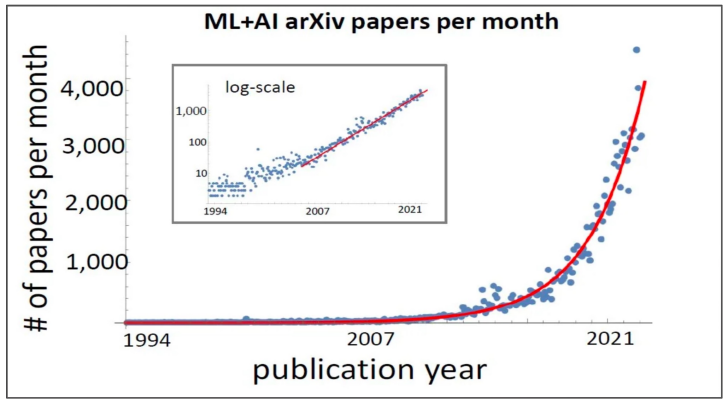
arXiv:2308.11432v2 [cs.AI] 7 Sep 2023

HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face

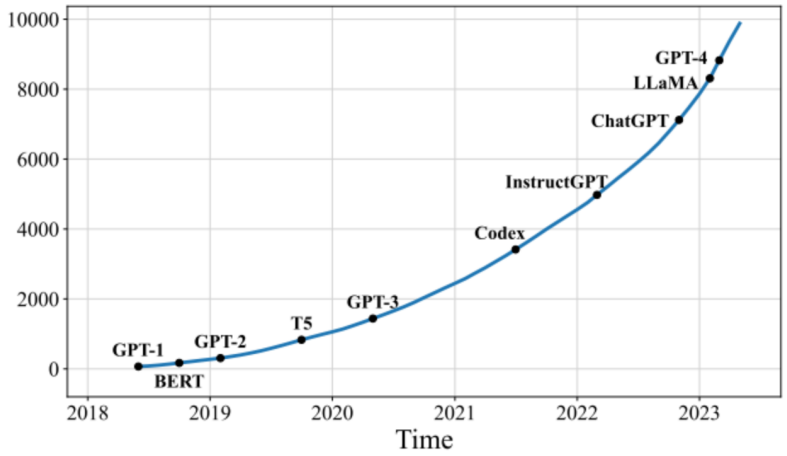


arXiv:2303.17580v3 [cs.CL] 25 May 2023

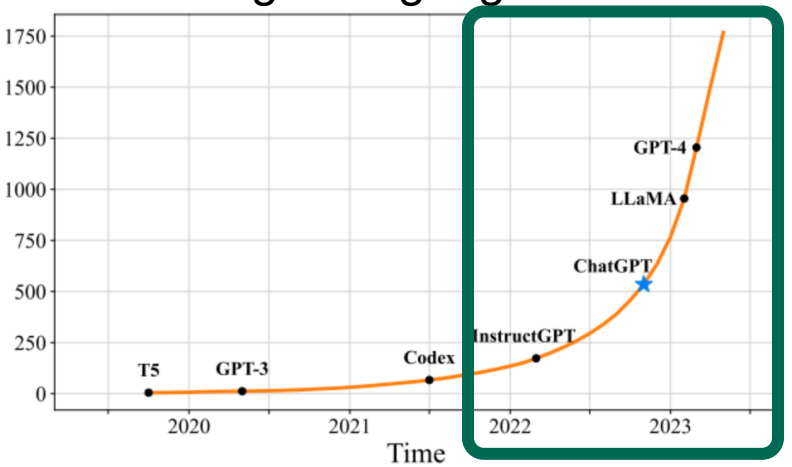
ML & AI



Language Models



Large Language Models



arXiv:2303.18223v12 [cs.CL] 11 Sep 2023

ENTERPRISEVISIONTECHNOLOGIES

Google "We Have No Moat, And Neither Does OpenAI"

DEVOPS ENTERPRISE SUMMIT

We Have No Moat

And neither does OpenAI

We've done a lot of looking over our shoulders at OpenAI. Who will cross the next milestone? What will the next move be?

But the uncomfortable truth is, *we aren't positioned to win this arms race and neither is OpenAI*. While we've been squabbling, a third faction has been quietly eating our lunch.

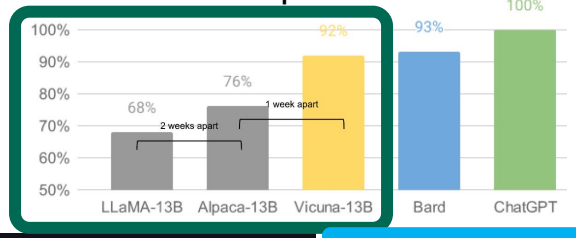
I'm talking, of course, about open source. Plainly put, they are lapping us. **Things we consider "major open problems" are solved and in people's hands today.** Just to name a few:

- **LLMs on a Phone:** People are running foundation models on a Pixel 6 at 5 tokens / sec.
- **Scalable Personal AI:** You can finetune a personalized AI on your laptop in an evening.
- **Responsible Release:** This one isn't "solved" so much as "obviated". There are entire websites full of art models with no restrictions whatsoever, and text is not far behind.
- **Multimodality:** The current multimodal ScienceQA SOTA was trained in an hour.

While our models still hold a slight edge in terms of quality, the gap is closing astonishingly quickly. Open-source models are faster, more customizable, more private, and pound-for-pound more capable. They are doing things with $100 and 13B params that we struggle with at $10M and 540B. And they are doing so in weeks, not months. This has profound implications for us:

- **We have no secret sauce.** Our best hope is to learn from and collaborate with what others are doing outside Google. We should prioritize enabling 3P integrations.
- **People will not pay for a restricted model when free, unrestricted alternatives are comparable in quality.** We should consider where our value add really is.
- **Giant models are slowing us down.** In the long run, the best models are the ones which can be iterated upon quickly. We should make small variants more than an afterthought, now that we know what is possible in the <20B parameter regime.

Model Comparison / Time



The Timeline

Feb 24, 2023 - LLaMA is Launched

Meta launches LLaMA, open sourcing the code, but not the weights. At this point, LLaMA is not instruction or conversation tuned. Like many current models, it is a relatively small model (available at 7B, 13B, 33B, and 65B parameters) that has been trained for a relatively large amount of time, and is therefore quite capable relative to its size.

March 3, 2023 - The Inevitable Happens

Within a week, LLaMA is leaked to the public. The impact on the community cannot be overstated. Existing licenses prevent it from being used for commercial purposes, but suddenly anyone is able to experiment. From this point forward, innovations come hard and fast.

March 12, 2023 - Language models on a Toaster

A little over a week later, Artem Andreenko gets the model working on a Raspberry P
At this point the model runs too slowly to be practical because the weights must be paged in and out of memory. Nonetheless, this sets the stage for an onslaught of minification efforts.

March 13, 2023 - Fine Tuning on a Laptop

The next day, Stanford releases Alpaca, which adds instruction tuning to LLaMA. More important than the actual weights, however, was Eric Wang's alpaca-lora repo, which used low rank fine-tuning to do this training "within hours on a single RTX 4090".

Suddenly, anyone could fine-tune the model to do anything, kicking off a race to the bottom on low-budget fine-tuning projects. Papers proudly describe their total spend of a few hundred dollars. What's more, the low rank updates can be distributed easily and separately from the original weights, making them independent of the original

March 18, 2023 - Now It's Fast

first "no GPU" solution that is fast enough to be practical.

March 19, 2023 - A 13B model achieves "parity" with Bard

The next day, a cross-university collaboration releases Vicuna, and uses GPT-4-powered eval to provide qualitative comparisons of model outputs. While the evaluation method is suspect, the model is materially better than earlier variants. **Training Cost: $300.**

Notably, they were able to use data from ChatGPT while circumventing restrictions on its API - They simply sampled examples of "impressive" ChatGPT dialogue posted on sites like ShareGPT.

March 25, 2023 - Choose Your Own Model

Nomic creates GPT4All, which is both a model and, more importantly, an ecosystem. For the first time, we see models (including Vicuna) being gathered together in one place. **Training Cost: $100.**

March 28, 2023 - Open Source GPT-3

Cerebras (not to be confused with our own Cerebra) trains the GPT-3 architecture using the optimal compute schedule implied by Chinchilla, and the optimal scaling implied by μ-parameterization. This outperforms existing GPT-3 clones by a wide margin, and represents the first confirmed use of μ-parameterization "in the wild". These models are trained from scratch, meaning the community is no longer dependent on LLaMA.

March 28, 2023 - Multimodal Training in One Hour

Using a novel Parameter Efficient Fine Tuning (PEFT) technique, LLaMA-Adapter introduces instruction tuning and multimodality in one hour of training. Impressively, they do so with just 1.2M learnable parameters. The model achieves a new SOTA on multimodal ScienceQA.

April 3, 2023 - Real Humans Can't Tell the Difference Between a 13B Open Model and ChatGPT

Berkeley launches Koala, a dialogue model trained entirely using freely available data.

They take the crucial step of measuring real human preferences between their model and ChatGPT. While ChatGPT still holds a slight edge, more than 50% of the time users either prefer Koala or have no preference. **Training Cost: $100.**

April 15, 2023 - Open Source RLHF at ChatGPT Levels

Open Assistant launches a model and, more importantly, a dataset for Alignment via RLHF. Their model is close (48.3% vs. 51.7%) to ChatGPT in terms of human preference. In addition to LLaMA, they show that this dataset can be applied to Pythia-12B, giving people the option to use a fully open stack to run the model. Moreover, because the dataset is publicly available, it takes RLHF from unachievable to cheap and easy for small experimenters.

Fine Tuning GPU Hours

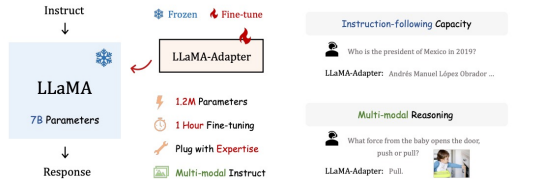


Table 1: **Efficiency Comparison** of different instruction-following methods. The training time is tested on 8 A100 GPUs.

Model	Tuned Params	Storage Space	Training Time
Alpaca [60]	7B	13G	3 hours
Alpaca-LoRA [1]	4.2M	16.8M	1.5 hours
LLaMA-Adapter	**1.2M**	**4.7M**	**1 hour**

https://www.semianalysis.com/p/google-we-have-no-moat-and-neither

arXiv:2303.16199v2 [cs.CV] 14 Jun 2023

Agent Experiment (Closed vs Open)

DEVOPS ENTERPRISE SUMMIT

Perplexity.ai – Closed agent

critical analysis of Google "We Have No Moat, And Neither Does OpenAI"
Leaked Internal Google Document Claims Open Source AI Will Outcompete Google and OpenAI

Copilot

Perplexity Millions of Dollars

Considering 15 sources

Sources

Google "We Have No Moat, And Neither Does... semianalysis · 1	Leaked Google document reveals open source AI... emergentmind · 2	Google: "We have no moat, and neither does... linkedin · 3	Leaked Internal Google Document Claims Open... hackaday · 4
Google "We have no moat, and neither does OpenAI... ycombinator · 5	[D] Google "We Have No Moat, And Neither Does... reddit · 6	Leaked Google Memo: "We Have No Moat, and... marketingaiinstitu... · 7	View 8 more

Answer

A leaked internal Google document claims that open-source AI models are advancing rapidly and may outpace both Google and OpenAI [1]. The document suggests that Google should embrace open-source AI, prioritize third-party integrations, and reconsider the focus on large models [2]. The anonymous Google researcher who leaked the document believes that neither Google nor OpenAI is positioned to win the AI "arms race" due to the rapid advancements made by open-source AI projects [8].

One of the key factors contributing to the success of open-source AI projects is Low-Rank Adaptation (LoRA), a technique that allows model fine-tuning at a fraction of the cost and time by reducing the size of update matrices [4]. The document argues that Google and OpenAI do not have a "secret sauce" that makes their approaches better than what the wider community can come up with [4]. The researcher suggests that Google should establish itself as a leader in the open-source community, cooperate with the broader conversation, and potentially publish model weights for small ULM variants [1].

The implications of this leaked document could significantly impact Google and OpenAI's future business strategies, as both companies may need to reevaluate their approach to AI research and development in light of the rapid advancements made by open-source AI projects [8].

GPT-Researcher – Open agent

Critical Analysis of Leaked Google Document: "We Have No Moat, And Neither Does OpenAI"

The contemporary field of artificial intelligence (AI) has been stirred by the recent leak of an internal Google document penned by a Google researcher. The memo, while not an official statement from the company, suggests a drastic shift in the AI landscape - the imminent threat from open-source AI. The document states that Google and its AI industry competitor, OpenAI, have no 'moat' in the AI technology domain, referring to the lack of a unique and constant competitive edge. This report provides a rigorous examination of the implications raised by the leaked document, the responses thus far, and the broader landscape of the AI industry.

The Clarion Call: No Moat in AI

The leaked memo alludes to the rapid advancements of open-source AI projects, outpacing proprietary models developed by tech giants like Google and OpenAI. The document also insinuates that the 'quality gap' between open and proprietary models is dwindling. Factors such as speed, customization, privacy, and capability have been spotlighted as areas where open-source projects are excelling. Tangible exemplifications of this have been noted with models like Meta's LLaMA language model and Vicuna-13B, an open-source chatbot. Both have evidenced the potential of open-source AI to challenge more established players. However, the text provides no quantifiable data to support these assertions, leaving the actual degree of this impact open to interpretation.

Response from Google and OpenAI

Contrarily, Demis Hassabis, CEO of Google's DeepMind, authenticated the leaked memo but disagreed with its conclusions. Hassabis appeared optimistic about Google's stand in the AI industry, pointing to the company's innovative researchers and the potential for progress through collaborations between Google Brain and Google DeepMind teams. No specific counter-arguments, numbers, or additional expert opinions were provided to refute the conte
the memo. Also, there has been no formal response from OpenAI, theret
adding a level of ambiguity to the situation.

Open-source AI: Boon and Bane

Interestingly, open-source AI is represented as both an opportunity and a challenge. The mounting momentum of open-source AI projects, with breakthroughs like the Low Rank Adaptation technique, and reduced entry barriers have emphasized the democratizing potential of open-source AI. Yet, there are also concerns about potential EU regulations on open-source models, which could impede the release of open-source AI in Europe. Here, the text lacks specifics on what these regulations could entail and how they might shift the AI dynamic.

On the flip side, while the open-source AI revolution poses potential threats to AI stalwarts like Google and OpenAI, it simultaneously prompts the necessity for these organizations to reassess their competitive strategies. The documents stress the need to embrace collaborative measures, learning from the innovations in the thriving open-source community, and staying open to external collaborations to retain their relevance in the AI market.

Beyond Openness: A Matter of Dominance

The Wired.com article brings a broader perspective to this discourse by highlighting the challenges of true openness in AI. The researchers argue that ownership of software frameworks, the high cost of computing power, secrecy around training data, and the lack of available expertise to fine-tune models concentrate power among large corporations. In this context, the open AI model may consolidate authority rather than lead to democratization. Memorably, it proposes the need for regulatory measures to avoid technological dominance by large corporations, especially in sensitive domains like healthcare, finance, education, and the workplace. Here again, further context or specifics could provide a more comprehensive understanding of these challenges.

Conclusion

The document leak may be seen as a hypothesis on the future trajectory of the AI race, highlighting that existing AI entities might not be as infallible as presumed. However, the underlying analysis remains somewhat speculative in the absence of concrete statistical data to support these assertions. Therefore, while it provokes thoughtful discourse on the future of AI, the suggestions of

Community + Us 48hrs

1. SemiAnalysis - Google "We Have No Moat, And Neither Does OpenAI"
2. Artisana - Leaked Google Memo Claiming "We Have No Moat, And Neither Does OpenAI" Shakes
3. LinkedIn - Google "We Have No Moat, And Neither Does Open AI"
4. The Guardian - Google Engineer Open Source Technology AI OpenAI
5. UX Design Bootcamp - Seizing The Open Source AI Opportunity
6. CMSWire - The Open Source Revolution Challenging AI Giants
7. The Verge - Google Memo Moat AI Leak
8. Datacamp - OpenAI Google AI Data Science
9. DeepChecks - OpenAI's ChatGPT vs. Google's Bard AI: A Comparative Analysis
10. Wired - The Myth of Open Source AI

Frontier and Open LLMs: A Snapshot

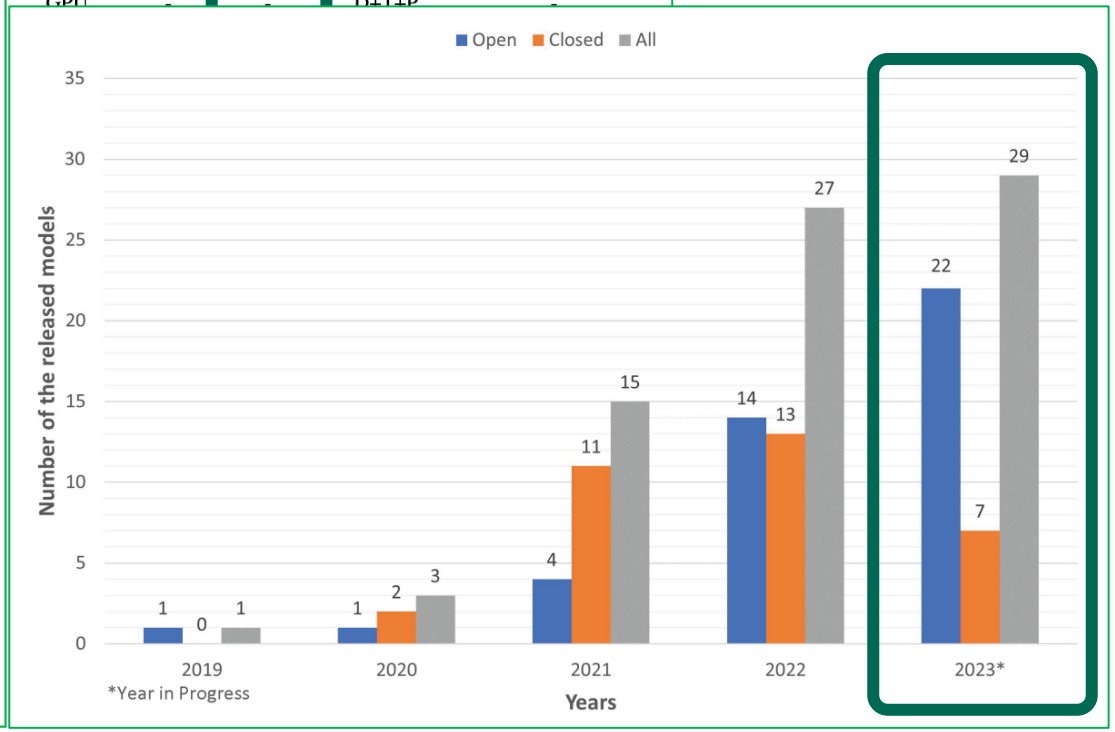
DEVOPS ENTERPRISE SUMMIT

Frontier (Closed) Models

Model	Lab	Parameters (B)	Tokens trained (B)	Announced
Gemini	Google DeepMind	8000	65000	TBA
Meta GPT	Meta AI	2000		TBA
GPT-5	OpenAI	2000		TBA
Claude-Next	Anthropic	2000		TBA
GPT-4 MathMix	OpenAI	1800	13000	May/2023
GPT-4	OpenAI	1800	13000	Mar/2023
GLaM	Google	1200		Dec/2021
PanGu-Sigma	Huawei	1085		Mar/2023
Switch	Google	1000	576	Jan/2021
Med-PaLM M	Google DeepMind	540	780	Jul/2023
Med-PaLM 1	Google DeepMind	540	780	Dec/2022
Flan-PaLM	Google	540	780	Oct/2022
U-PaLM	Google	540	780	Oct/2022
Minerva	Google	540	818.5	Jun/2022
PaLM-Coder	Google	540	780	Apr/2022
PaLM	Google	540	780	Apr/2022
MT-NLG	Microsoft/NVIDIA	530	270	Oct/2021
BERT-480	Google	480		Nov/2021
AudioPaLM	Google	340	3600	Jun/2023
PaLM 2	Google	340	3600	May/2023
Gopher	DeepMind	280	300	Dec/2021
ERNIE 3.0 Titan	Baidu	260		Dec/2021
Ajax GPT	Apple	200		TBA

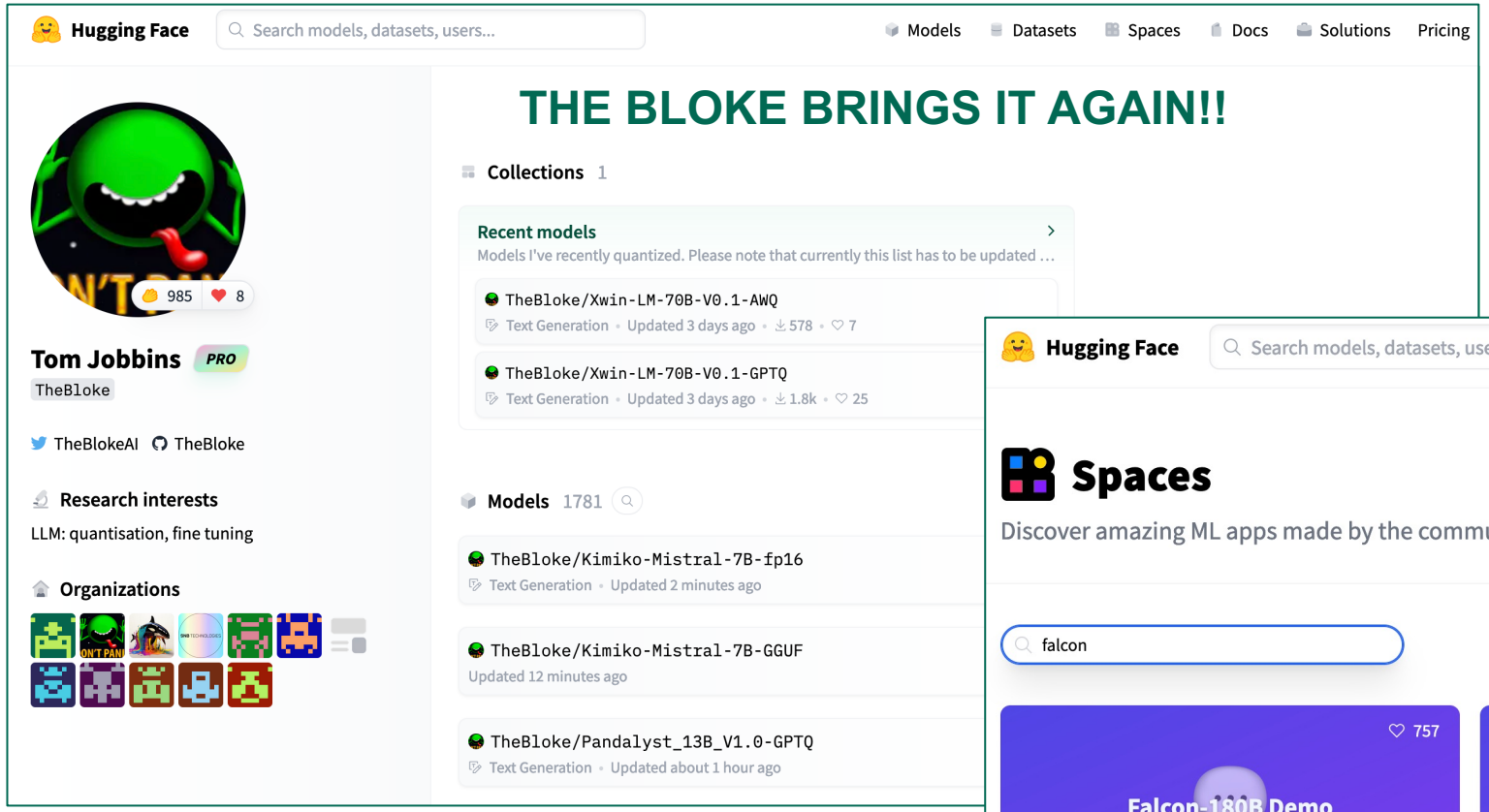
https://lifearchitect.ai/

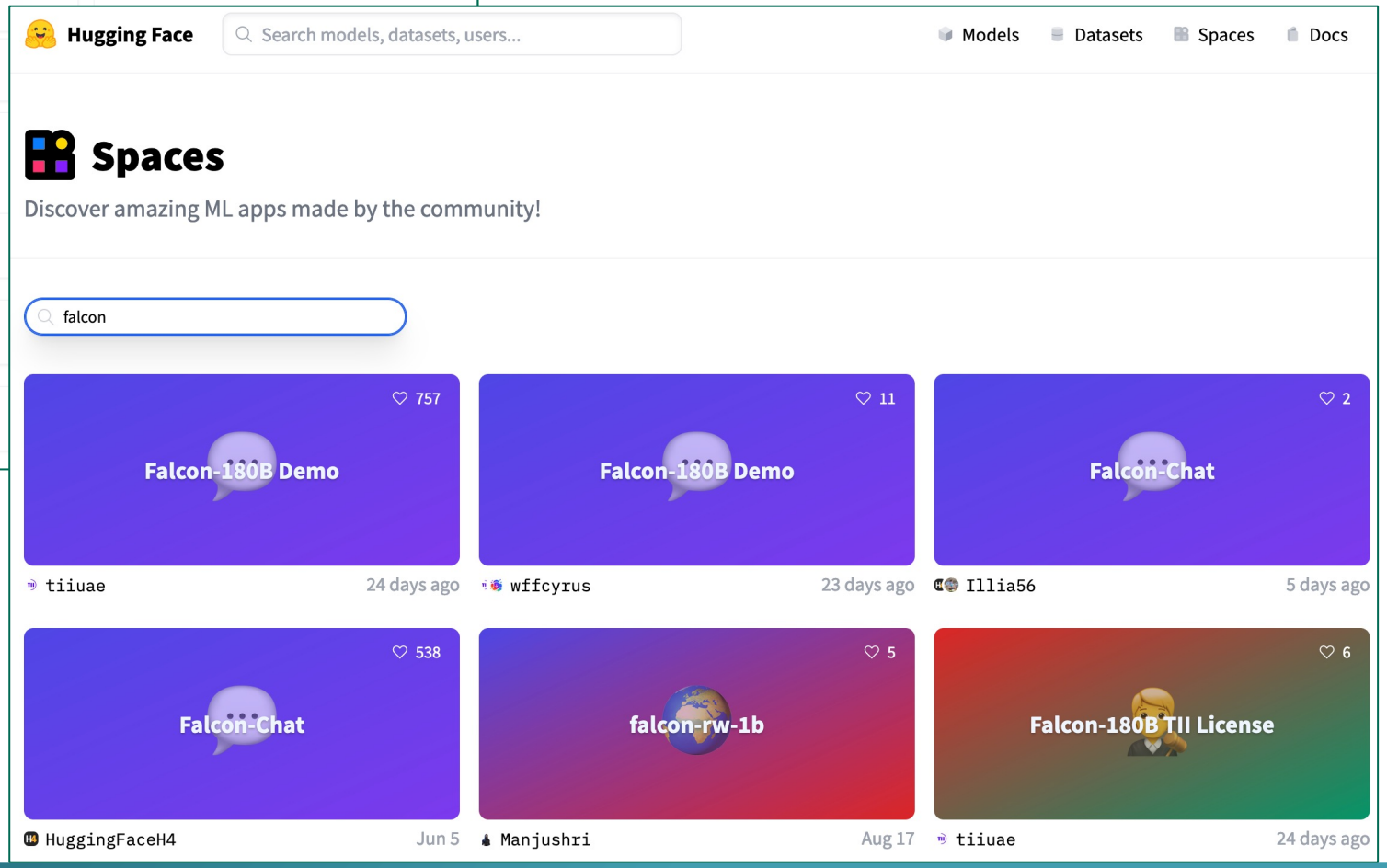
Processing Unit Type	Training Time	Calculated Train. Cost	Training Parallelism	Library
TPU v3	-	-	D+M	Mesh TensorFlow
V100	-	-	M	-
-	-	-	-	-
Ascend 910	-	-	)+OP+P+O+R	MindSpore
-	-	-	D+M	JAXFormer
-	-	-	-	-
V100	-	-	M*	PaddlePaddle
GPU	-	-	D+M+P	Megatron+DS
A100	321h	1.32 Mil	M	Megatron
GPU	-	-	D+T+P	-



Leaderboards, Benchmarks, Models & Spaces

DEVOPS ENTERPRISE SUMMIT



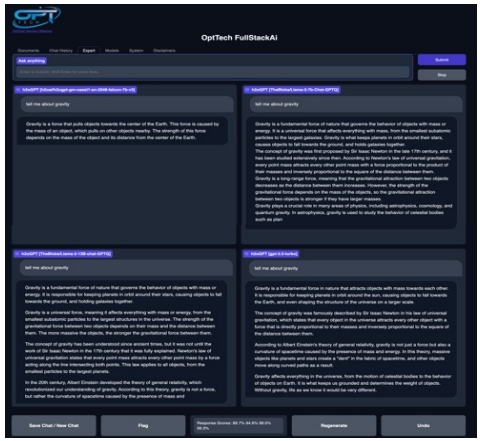


ENTERPRISEVISIONTECHNOLOGIES

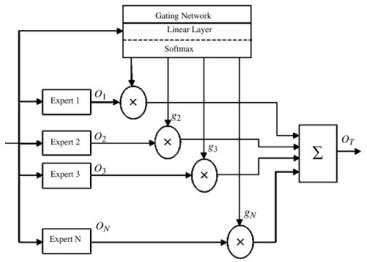
Navigating the AI Tool & Deployment Landscape

DEVOPS ENTERPRISE SUMMIT

Single LLM,
Ensemble-Mixture,
RAG and/or Agents



arXiv:1701.06538v1 Hinton, et al



Cloud Provider

aws
cohere
ANTHROPIC

Framework

databricks
mosaicML

Orchestration

LangChain
LlamaIndex

Vector Database

milvus
chroma
Pinecone

Hugging Face
H2O.ai

OpenAI
LLAMA 2

Foundation Model

Test Evaluate Debug Observe

LangSmith
truera
arize

Inferencing Datacenter Edge

vllm
Text Generation Inference
M2 ULTRA
NVIDIA DGX™
Hewlett Packard Enterprise
NVIDIA

ENTERPRISEVISIONTECHNOLOGIES

Primer on Tokens, Encoding, Embedding and Transformers

DEVOPS ENTERPRISE SUMMIT

Tokens & Encoding

My favorite color is red.

[3666, 4004, 3124, 318, 2266, 13]

My favorite color is Red.

[3666, 4004, 3124, 318, 2297, 13]

Red is my favorite color.

[7738 318, 616, 4004, 3124, 13]

Encodings

tiktoken

tiktoken is a fast BPE tokeniser for use with OpenAI's models.

Encodings specify how text is converted into tokens. Different models use different encodings.

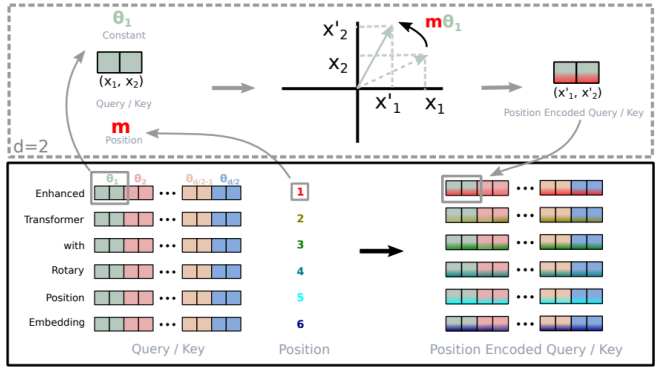
`tiktoken` supports three encodings used by OpenAI models:

Encoding name	OpenAI models
`cl100k_base`	`gpt-4`, `gpt-3.5-turbo`, `text-embedding-ada-002`
`p50k_base`	Codex models, `text-davinci-002`, `text-davinci-003`
`r50k_base` (or `gpt2`)	GPT-3 models like `davinci`

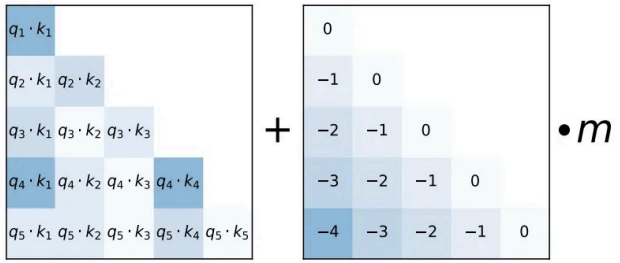
Embeddings - OpenAI API

Embedding Types

- Positional Embedding
- Relevant Embedding
- RoPE
- ALiBi



Rotary Positional Embeddings (RoPE)



Attention with Linear Biases (ALiBi)

Meta Platforms Position Interpolation
arXiv:2306.15595v2 [cs.CL] 28 Jun 2023

arXiv:2104.09864v4

arXiv:2108.12409v2

Transformers

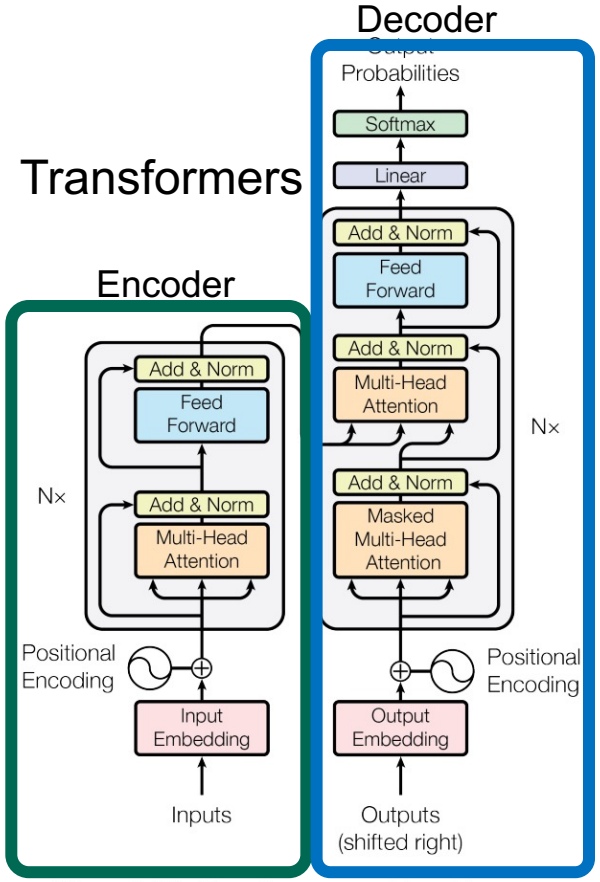
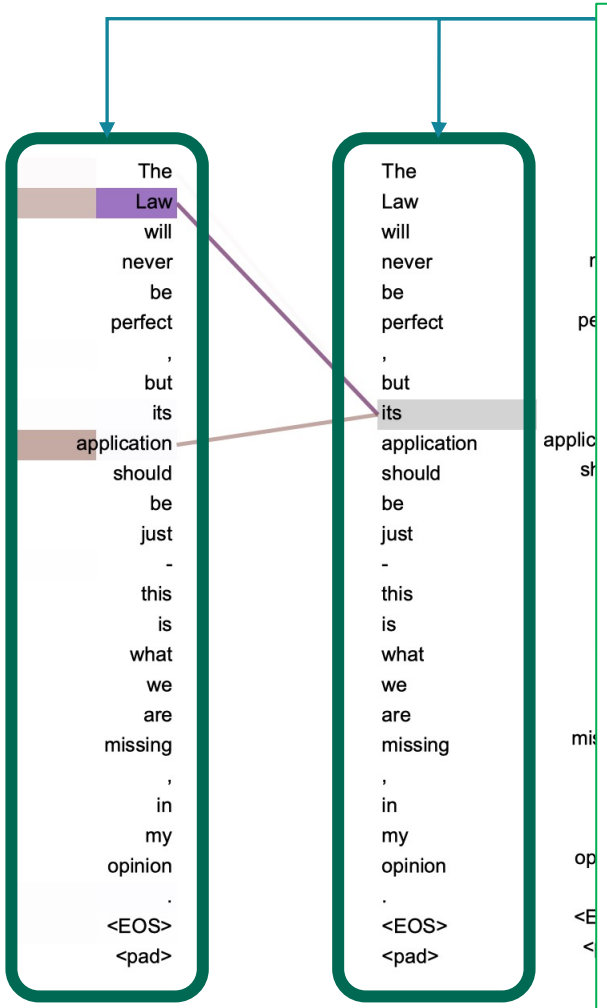


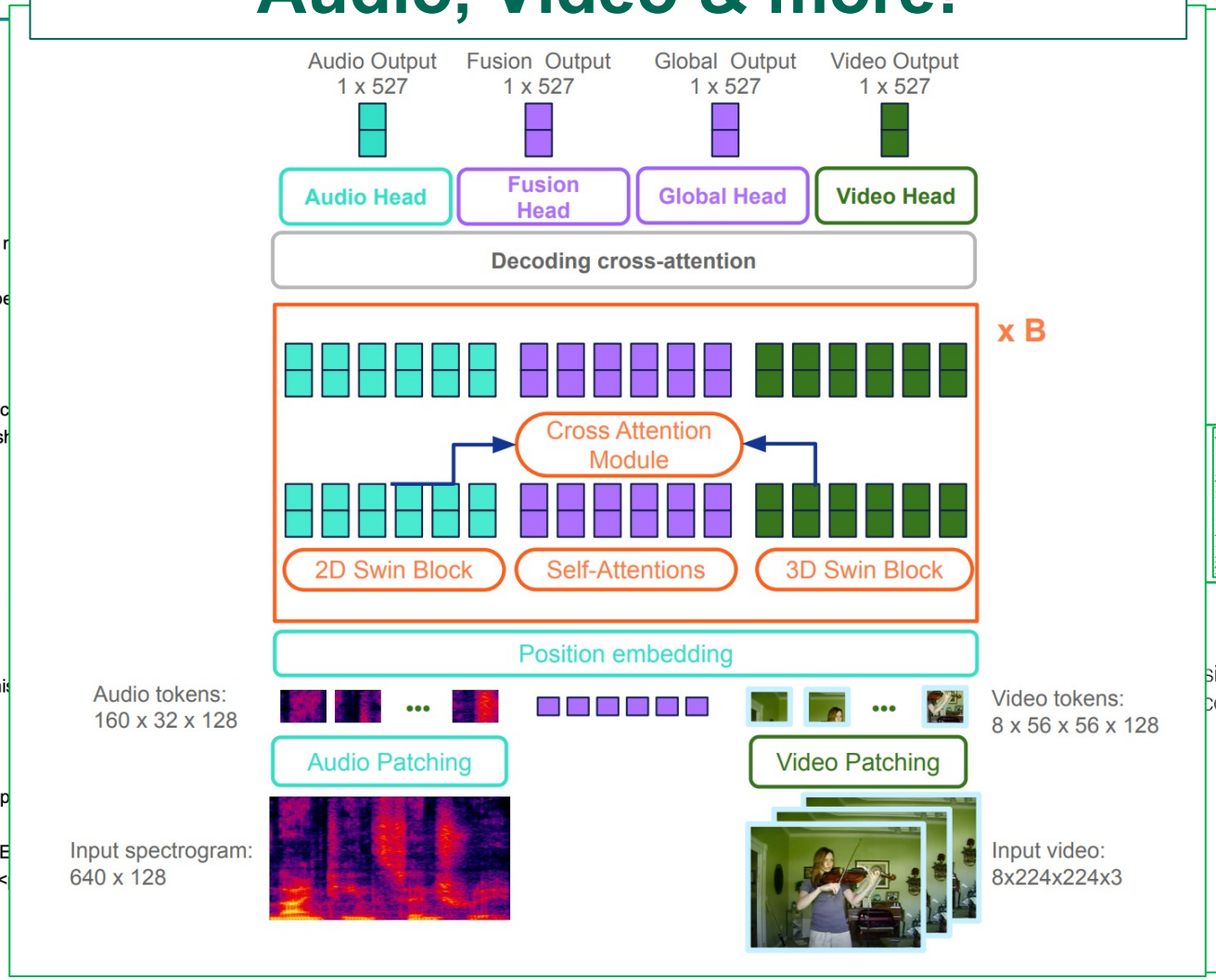
Figure 1: The Transformer - model architecture.

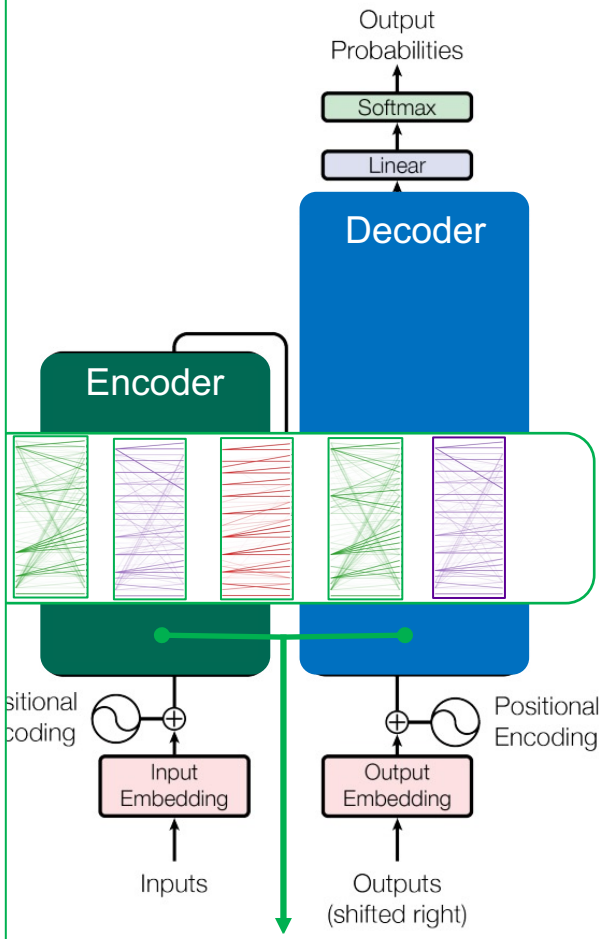
arXiv:1706.03762v7 [cs.CL] 2 Aug 2023

ENTERPRISEVISIONTECHNOLOGIES

Transformers can also learn Audio, Video & more!



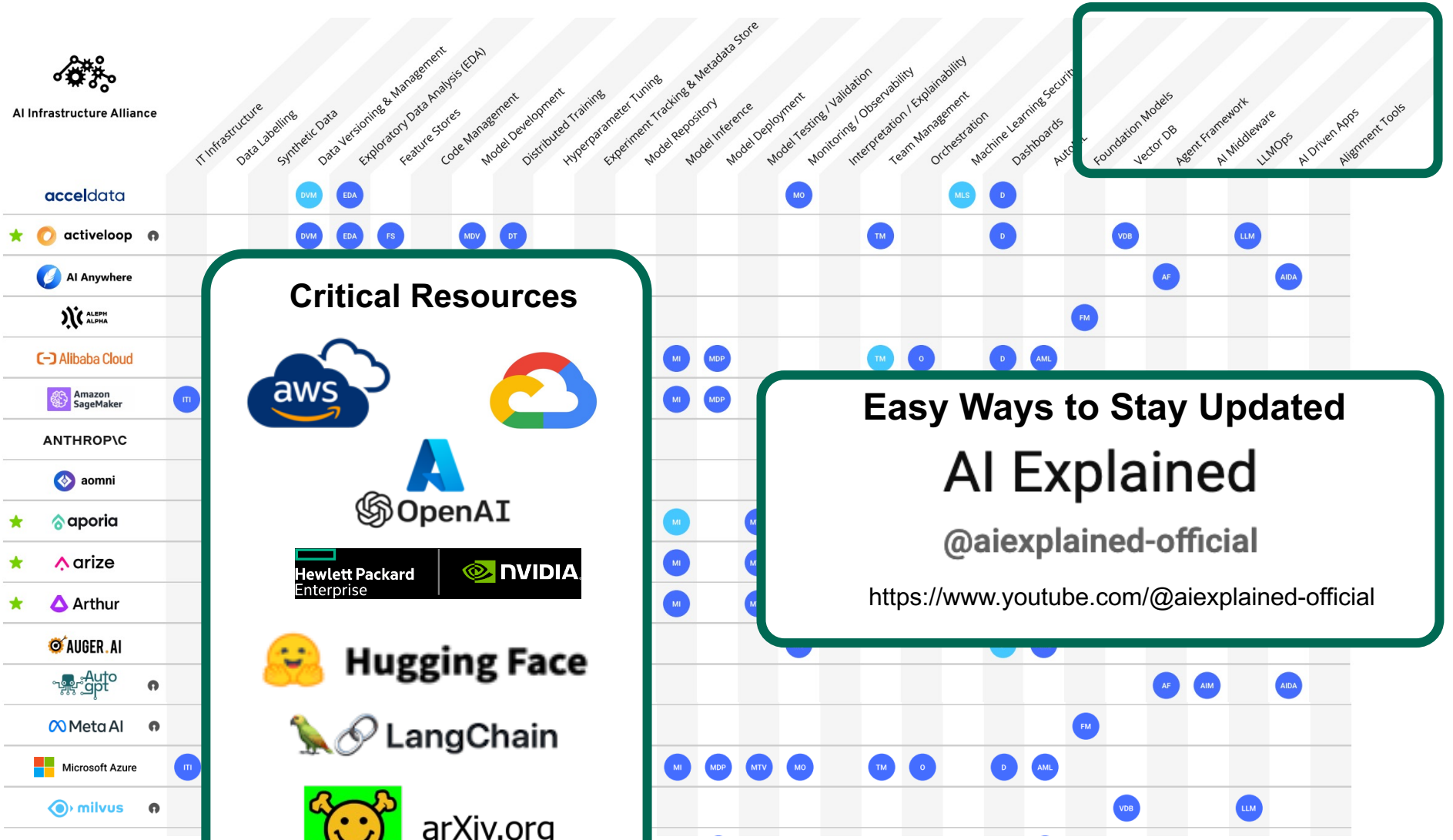




Multi Headed Attention

arXiv:1706.03762v7 [cs.CL] 2 Aug 2023

Strategic Decisions & Critical Resources



Critical Resources

aws

OpenAI

Hewlett Packard Enterprise | NVIDIA

Hugging Face

LangChain

arXiv.org

Easy Ways to Stay Updated

AI Explained

@aiexplained-official

https://www.youtube.com/@aiexplained-official

Home - AI Infrastructure Alliance (ai-infrastructure.org)

Risks and Insights

DEVOPS ENTERPRISE SUMMIT

Rapidly Changing Landscape

- Unprecedented pace of AI developments
- Evolution faster than ever before

Dangers of Outdated Understanding

- Risks of falling behind
- Prepare to navigate through paradigm shifts

Embrace Advancement

- Know impact of technological advancements
- Open-minded approach to new paradigms
- Foster a mindset of adaptability

Treat Data as a Treasure

- Establish importance of safeguarding data
- Encourage responsibility in data protection
- Maintain a sense of accountability

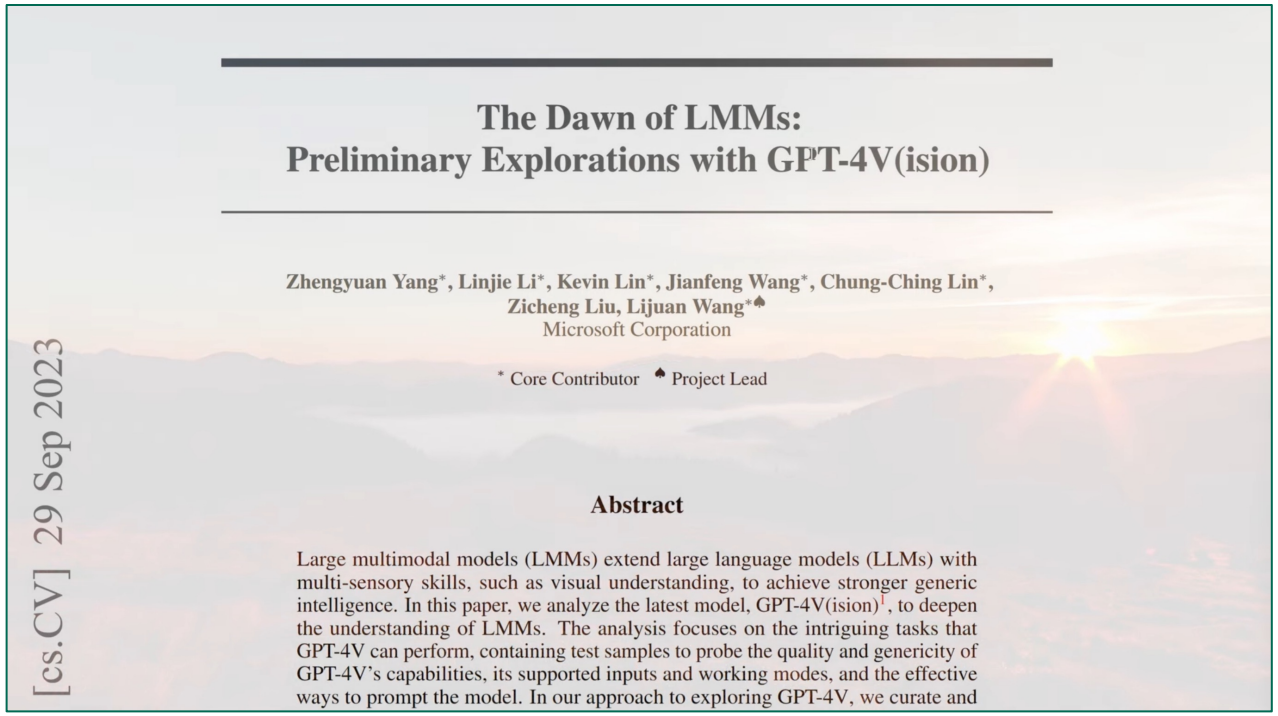
Ethical Components

- Consider ethical implications
- Mindful development and deployment of AI

Prepare For Expanded Personal Device Use

- Running advanced models on personal devices
- Highlight security, accessibility and portability

A Closer Look at ArXiv.org



[cs.CV] 29 Sep 2023

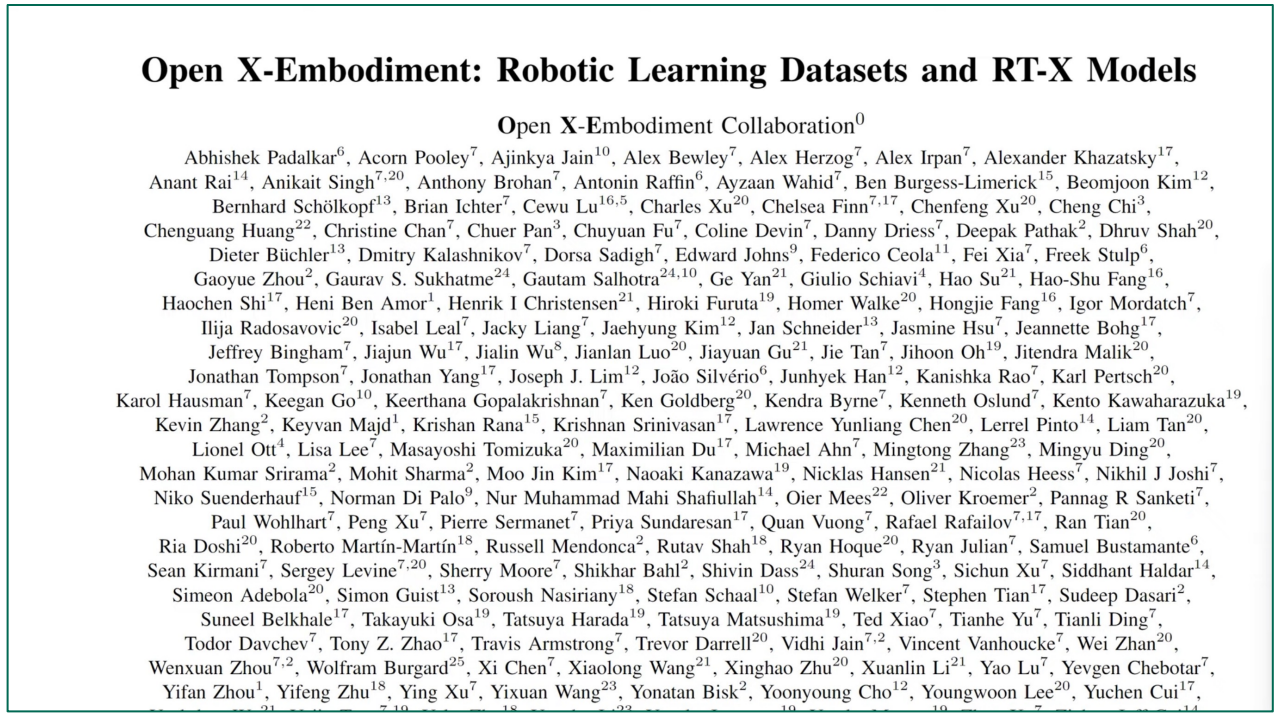
The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision)

Zhengyuan Yang*, Linjie Li*, Kevin Lin*, Jianfeng Wang*, Chung-Ching Lin*,
Zicheng Liu, Lijuan Wang*♠
Microsoft Corporation

* Core Contributor ♠ Project Lead

Abstract

Large multimodal models (LMMs) extend large language models (LLMs) with multi-sensory skills, such as visual understanding, to achieve stronger generic intelligence. In this paper, we analyze the latest model, GPT-4V(ision)[1], to deepen the understanding of LMMs. The analysis focuses on the intriguing tasks that GPT-4V can perform, containing test samples to probe the quality and genericity of GPT-4V's capabilities, its supported inputs and working modes, and the effective ways to prompt the model. In our approach to exploring GPT-4V, we curate and



Open X-Embodiment: Robotic Learning Datasets and RT-X Models

Open X-Embodiment Collaboration[0]

Abhishek Padalkar[6], Acorn Pooley[7], Ajinkya Jain[10], Alex Bewley[7], Alex Herzog[7], Alex Irpan[7], Alexander Khazatsky[17], Anant Rai[14], Anikait Singh[7,20], Anthony Brohan[7], Antonin Raffin[6], Ayzaan Wahid[7], Ben Burgess-Limerick[15], Beomjoon Kim[12], Bernhard Schölkopf[13], Brian Ichter[7], Cewu Lu[16,5], Charles Xu[20], Chelsea Finn[7,17], Chenfeng Xu[20], Cheng Chi[3], Chenguang Huang[22], Christine Chan[7], Chuer Pan[3], Chuyuan Fu[7], Coline Devin[7], Danny Driess[7], Deepak Pathak[2], Dhruv Shah[20], Dieter Büchler[13], Dmitry Kalashnikov[7], Dorsa Sadigh[7], Edward Johns[9], Federico Ceola[11], Fei Xia[7], Freek Stulp[6], Gaoyue Zhou[2], Gaurav S. Sukhatme[24], Gautam Salhotra[24,10], Ge Yan[21], Giulio Schiavi[4], Hao Su[21], Hao-Shu Fang[16], Haochen Shi[17], Heni Ben Amor[1], Henrik I Christensen[21], Hiroki Furuta[19], Homer Walke[20], Hongjie Fang[16], Igor Mordatch[7], Ilija Radosavovic[20], Isabel Leal[7], Jacky Liang[7], Jaehyung Kim[12], Jan Schneider[13], Jasmine Hsu[7], Jeannette Bohg[17], Jeffrey Bingham[7], Jiajun Wu[17], Jialin Wu[8], Jianlan Luo[20], Jiayuan Gu[21], Jie Tan[7], Jihoon Oh[19], Jitendra Malik[20], Jonathan Tompson[7], Jonathan Yang[17], Joseph J. Lim[12], João Silvério[6], Junhyek Han[12], Kanishka Rao[7], Karl Pertsch[20], Karol Hausman[7], Keegan Go[10], Keerthana Gopalakrishnan[7], Ken Goldberg[20], Kendra Byrne[7], Kenneth Oslund[7], Kento Kawaharazuka[19], Kevin Zhang[2], Keyvan Majd[1], Krishan Rana[15], Krishnan Srinivasan[17], Lawrence Yunliang Chen[20], Lerrel Pinto[14], Liam Tan[20], Lionel Ott[4], Lisa Lee[7], Masayoshi Tomizuka[20], Maximilian Du[17], Michael Ahn[7], Mingtong Zhang[23], Mingyu Ding[20], Mohan Kumar Srirama[2], Mohit Sharma[2], Moo Jin Kim[17], Naoaki Kanazawa[19], Nicklas Hansen[21], Nicolas Heess[7], Nikhil J Joshi[7], Niko Suenderhauf[15], Norman Di Palo[9], Nur Muhammad Mahi Shafiullah[14], Oier Mees[22], Oliver Kroemer[2], Pannag R Sanketi[7], Paul Wohlhart[7], Peng Xu[7], Pierre Sermanet[7], Priya Sundaresan[17], Quan Vuong[7], Rafael Rafailov[7,17], Ran Tian[20], Ria Doshi[20], Roberto Martín-Martín[18], Russell Mendonca[2], Rutav Shah[18], Ryan Hoque[20], Ryan Julian[7], Samuel Bustamante[6], Sean Kirmani[7], Sergey Levine[7,20], Sherry Moore[7], Shikhar Bahl[2], Shivin Dass[24], Shuran Song[3], Sichun Xu[7], Siddhant Haldar[14], Simeon Adebola[20], Simon Guist[13], Soroush Nasiriany[18], Stefan Schaal[10], Stefan Welker[7], Stephen Tian[17], Sudeep Dasari[2], Suneel Belkhale[17], Takayuki Osa[19], Tatsuya Harada[19], Tatsuya Matsushima[19], Ted Xiao[7], Tianhe Yu[7], Tianli Ding[7], Todor Davchev[7], Tony Z. Zhao[17], Travis Armstrong[7], Trevor Darrell[20], Vidhi Jain[7,2], Vincent Vanhoucke[7], Wei Zhan[20], Wenxuan Zhou[7,2], Wolfram Burgard[25], Xi Chen[7], Xiaolong Wang[21], Xinghao Zhu[20], Xuanlin Li[21], Yao Lu[7], Yevgen Chebotar[7], Yifan Zhou[1], Yifeng Zhu[18], Ying Xu[7], Yixuan Wang[23], Yonatan Bisk[2], Yoonyoung Cho[12], Youngwoon Lee[20], Yuchen Cui[17],

Adapting to Change: A DevOps Enterprise Tale

- Over 1300 DevOps talks
- One of the things that Gene wanted to do was have an AI help with generating one- or two-page executive summaries for DevOps Enterprise talks
  - “I’ve watched most of them, but not all of them — but I could definitely read every one-pager, that focused on the business problem they were trying to solve, the metrics they reported, and testimonials”
  - For years, they’ve generated transcripts (using rev.ai) — each talk typically is around 6K tokens
- In beginning, we needed to build tools to get around the fact that context windows were 2K — so there’s a set of techniques you need to chop up documents to do summarization. which I spent weeks building.
- by Aug, OpenAI updated GPT-3.5-turbo to 16K — which made it no longer necessary to use the tool. We could pass it (mostly) to OpenAI. Like I said, life moves fast in this space.

AI Adventures: AIforOPS Hackathon!!

DEVOPS ENTERPRISE SUMMIT

Exploring GenAI Together!

DEVOPS ENTERPRISE SUMMIT

Foundation Models

Large-scale models trained on diverse datasets that serve a
tuning. Foundation Models (FM) can learn a wide range o

Document Quest

Efficiently load, manage, and query data using various ve
query options and advanced information retrieval. Found

Utility Agents & A

Facilitates seamless integration with external APIs such a
messaging systems and more. Promotes efficient task ex

Were looking for help from the Community!!

Multi-Step Task Agents

Assists in defining high-level objectives, organizing tasks, and accomplishing complex, multi-step tasks. Provides robust support for context accuracy and action planning. Supports **Multi-Path Reasoning**.

Hyper-Personalized Agents

Offers advanced personalization options, with interactions tailored to individual user personas. Integrates real-time environmental elements supported by secure and accurate information.

Thank You!

DEVOPS ENTERPRISE SUMMIT

MidJourney Prompt: Whimsical illustration of a spaceship launching into the unknown, representing the beginning of the Odyssey. Day time, highly detailed Las Vegas strip in the background. --ar 16:9