CONF’

“Use case de fine-tuning d'un modèle de génération d’image”

Dmitry Kuzovkin
Head of AI @ Seelab

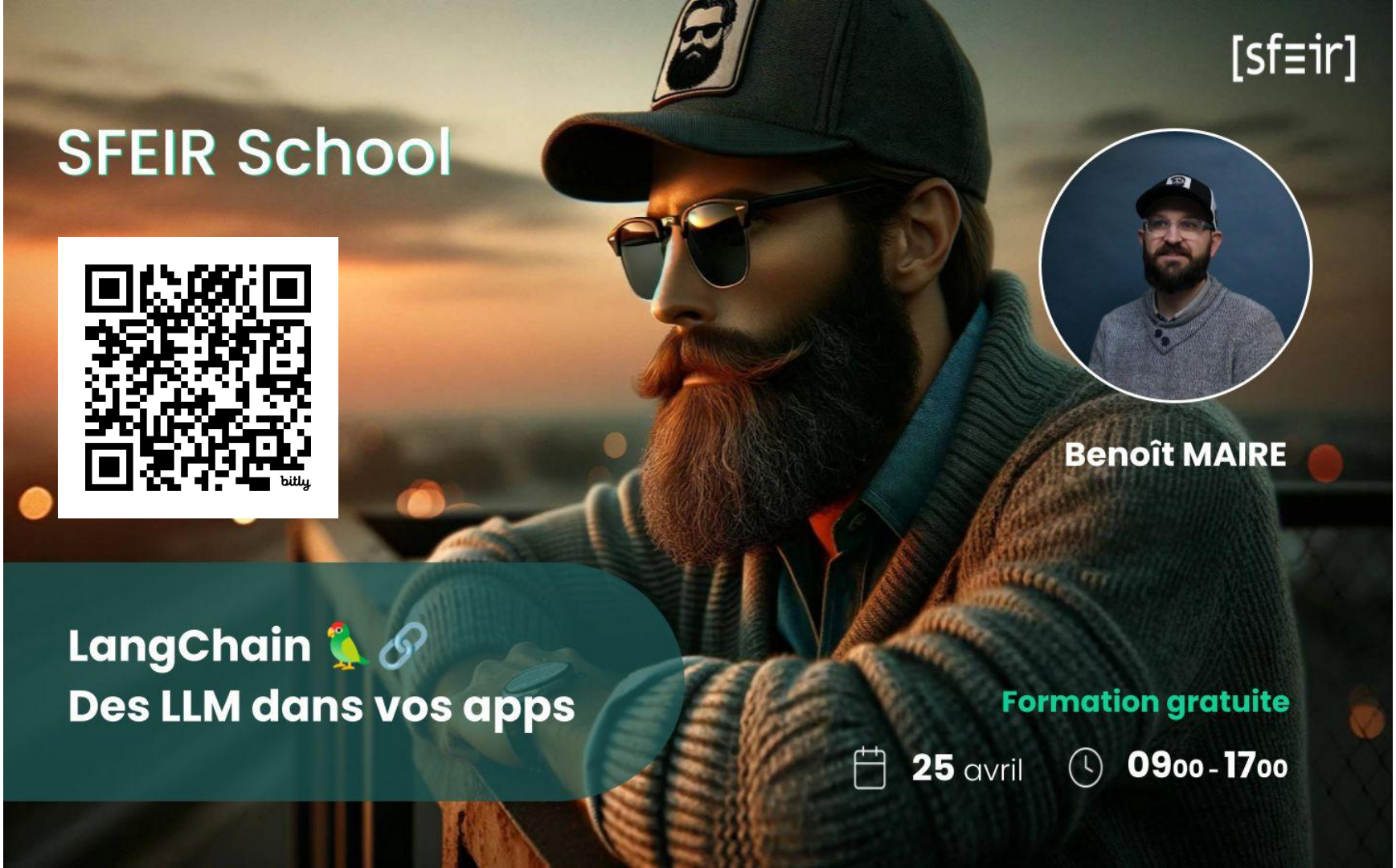
Matthieu Grosselin
CEO @ Seelab

seelab

Le **17/04/2024** à **19h**
Chez Lucca, **3 rue Michel Columb**, Nantes

lucca

Screeb RCA [sfeir] lonestone


[sfeir]
SFEIR School
bitly
Benoît MAIRE
LangChain 🦜🔗
Des LLM dans vos apps
Formation gratuite
25 avril
09h00 - 17h00

Schedule

 1- Monthly update (20min)

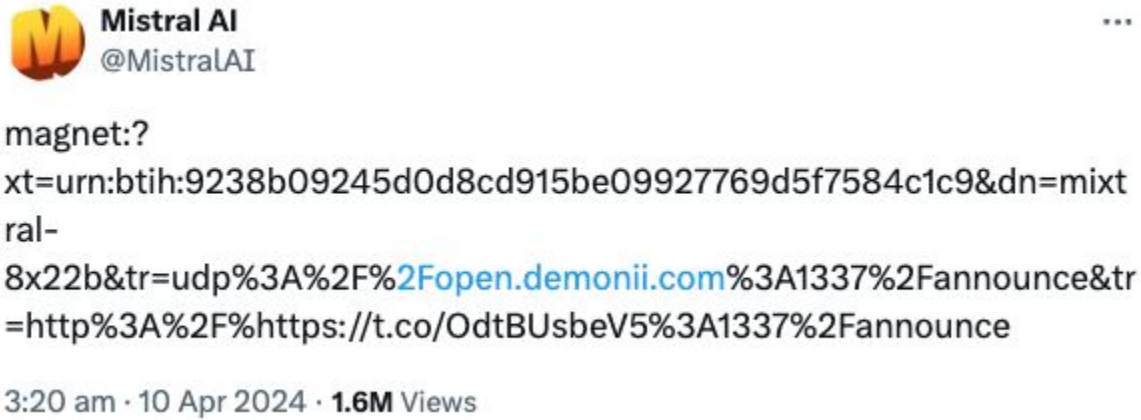
 2- Talk (1h30)

 3- Enjoy ;)

LLM updates

1) Opensource

Mistral: Mixtral-8x22B-v0.1



Mistral AI
@MistralAI

magnet:?
xt=urn:btih:9238b09245d0d8cd915be09927769d5f7584c1c9&dn=mixtral-
8x22b&tr=udp%3A%2F%2Fopen.demonii.com%3A1337%2Fannounce&tr=http%3A%2F%https://t.co/OdtBUsbeV5%3A1337%2Fannounce

3:20 am · 10 Apr 2024 · **1.6M** Views

Mistral: Mixtral-8x22B-v0.1

- 64k context
- MoE

Cohere: Command R+

- Open source (non commercial)
- 128k context window
- Use case: RAG + agent

Cohere API Pricing	$ / M input tokens	$ / M output tokens
Command R	$0.50	$1.50
Command R+	$3.00	$15.00

Benchmark

			Common sense and reasoning					Knowledge	
	Model	**Active parameters**	**MMLU**	**HellaS**	**WinoG**	**Arc C (5)**	**Arc C (25)**	**TriQA**	**NaturalQS**
	LLaMA 2 70B	*70B*	69.9%	87.1%	83.2%	86.0%	85.1%	77.57%	35.5%
CC-BY-NC license	Command R	*35B*	68.2%	87.0%	81.5%	-	66.5%	-	-
CC-BY-NC license	Command R+	*104B*	75.7%	**88.6%**	**85.4%**	-	71.0%	-	-
	Mistral 7B	*7B*	62.47%	83.1%	78.0%	77.2%	78.1%	68.8%	28.1%
	Mixtral 8x7B	*12.9B*	70.63%	86.6%	81.2%	85.8%	85.9%	78.4%	36.5%
	Mixtral 8x22B	*39B*	**77.75%**	**88.5%**	84.7%	**91.3%**	**91.3%**	**82.2%**	**40.1%**

2) Proprietary

OpenAI

GPT-4 Turbo

With 128k context, fresher knowledge and the broadest set of capabilities, GPT-4 Turbo is more powerful than GPT-4 and offered at a lower price.

Learn about GPT-4 Turbo ↗

Model	Input	Output
gpt-4-turbo-2024-04-09	$10.00 / 1M tokens	$30.00 / 1M tokens
gpt-4	$30.00 / 1M tokens	$60.00 / 1M tokens
gpt-4-32k	$60.00 / 1M tokens	$120.00 / 1M tokens

Gemini: 1.5 pro

1 million context window
System prompt + audio/video input + file input

Gemini: 1.5 pro

- Over 700,000 words
- Over 30,000 lines of code
- 11 hours of audio
- 1 hour of video

- $7 / m token (input)
- $21 / m token (output)

Gemini: text-embedding-004

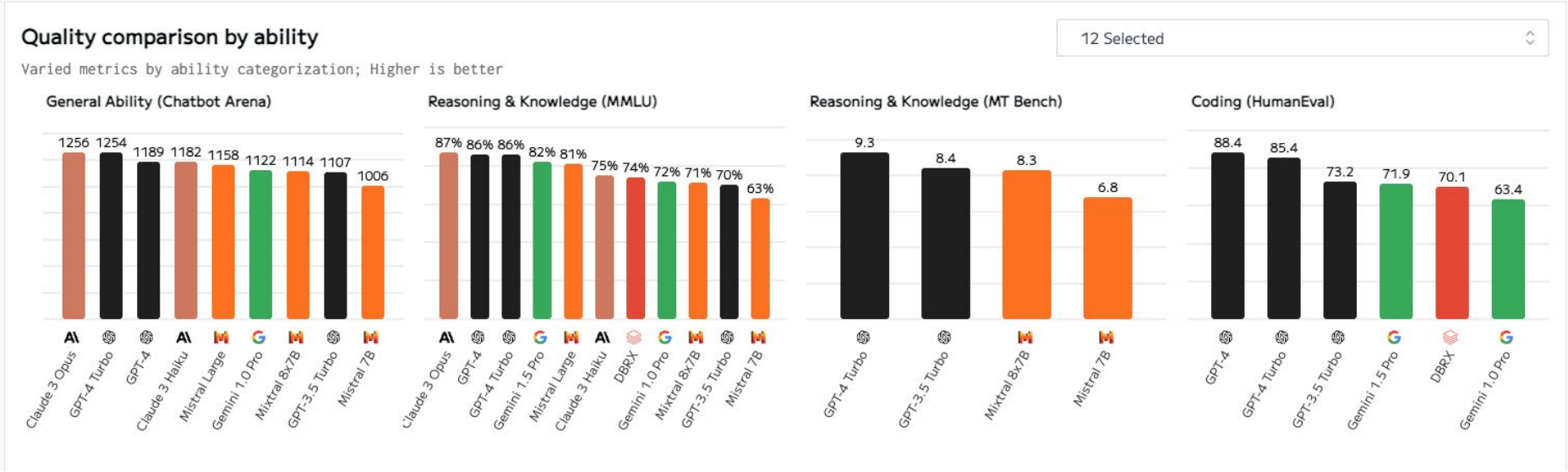
Gecko: Versatile Text Embeddings Distilled from Large Language Models

	Dim.	# Params.	Class.	Cluster.	Pair.	Rerank.	Retrieval	STS	Summary	Avg.
gritlm-8x7b	4,096	56B	78.53	50.14	84.97	59.80	55.09	83.26	29.82	65.66
e5-mistral-7b-instruct	4,096	7B	78.47	50.26	**88.34**	60.21	56.89	**84.63**	**31.40**	66.63
echo-mistral-7b-instruct	4,096	7B	77.43	46.32	87.34	58.14	55.52	82.56	30.73	64.69
gritlm-7b	4,096	7B	**79.46**	**50.61**	87.16	**60.49**	**57.41**	83.35	30.37	**66.76**
text-embedding-3-large (OpenAI)	3,072	n/a	75.45	49.01	85.72	59.16	55.44	81.73	29.92	64.59
gtr-t5-xxl	768	5B	67.41	42.42	86.12	56.66	48.48	78.38	30.64	58.97
gtr-t5-xl	768	1.2B	67.11	41.51	86.13	55.97	47.96	77.80	30.21	58.42
instructor-xl	768	1.5B	73.12	44.74	86.62	57.29	49.26	83.06	32.32	61.79
text-embedding-3-large-256 (OpenAI)	256	n/a	71.97	46.23	84.22	57.99	51.66	81.04	29.92	62.00
gecko-1b-256	256	1.2B	78.99	45.07	87.25	57.78	52.44	84.93	32.36	64.37
gecko-1b-768	768	1.2B	**81.17**	**47.48**	**87.61**	**58.91**	**55.70**	**85.06**	**32.63**	**66.31**
– zero-shot (FRet-only)	768	1.2B	70.26	46.82	86.27	57.60	53.16	83.14	32.16	62.64

Leaderboard LMSYS

Rank	Model	Arena Elo	95% CI	Votes	Organization	License	Knowledge Cutoff
1	GPT-4-Turbo-2024-04-09	1260	+5/-5	15751	OpenAI	Proprietary	2023/12
1	Claude 3 Opus	1255	+3/-4	56101	Anthropic	Proprietary	2023/8
1	GPT-4-1106-preview	1254	+3/-3	65159	OpenAI	Proprietary	2023/4
2	GPT-4-0125-preview	1250	+3/-4	50923	OpenAI	Proprietary	2023/12
5	Bard (Gemini Pro)	1209	+5/-5	12468	Google	Proprietary	Online
5	Claude 3 Sonnet	1203	+3/-3	62056	Anthropic	Proprietary	2023/8
7	Command R+	1193	+4/-4	29437	Cohere	CC-BY-NC-4.0	2024/3
7	GPT-4-0314	1189	+4/-4	42925	OpenAI	Proprietary	2021/9
9	Claude 3 Haiku	1182	+3/-3	57727	Anthropic	Proprietary	2023/8
10	GPT-4-0613	1164	+3/-3	61520	OpenAI	Proprietary	2021/9
10	Mistral-Large-2402	1158	+3/-4	37650	Mistral	Proprietary	Unknown
11	Qwen1.5-72B-Chat	1154	+4/-5	27826	Alibaba	Qianwen LICENSE	2024/2
12	Claude-1	1150	+4/-5	21868	Anthropic	Proprietary	Unknown
12	Mistral Medium	1148	+3/-5	30764	Mistral	Proprietary	Unknown
12	Command R	1148	+3/-4	33061	Cohere	CC-BY-NC-4.0	2024/3
16	Claude-2.0	1131	+7/-5	13484	Anthropic	Proprietary	Unknown

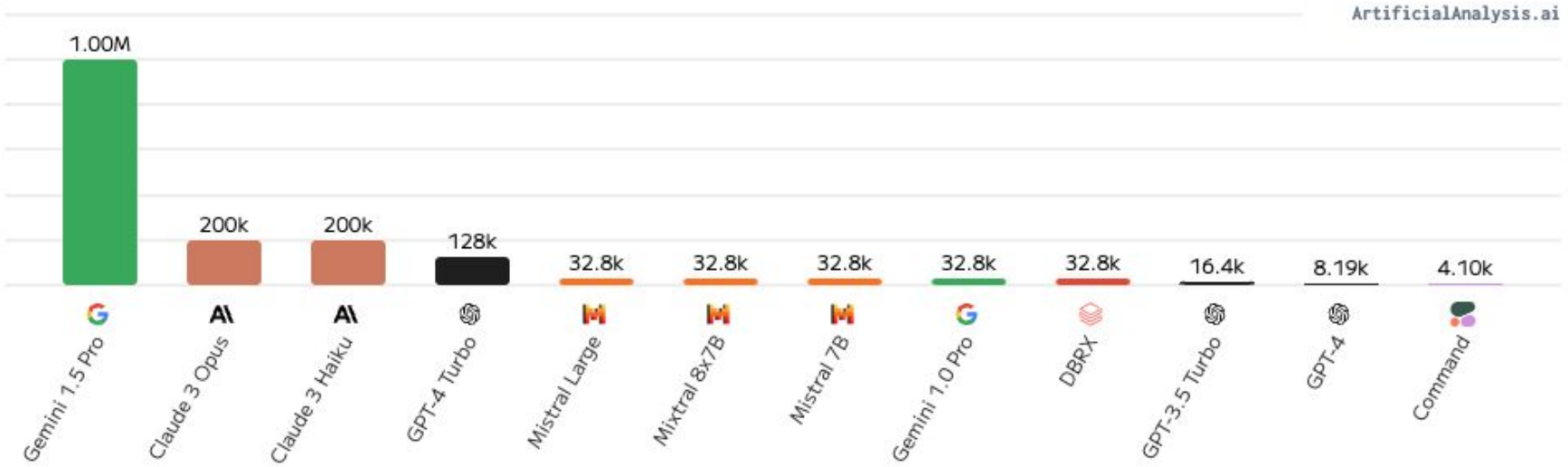
Leaderboard artificialanalysis.ai



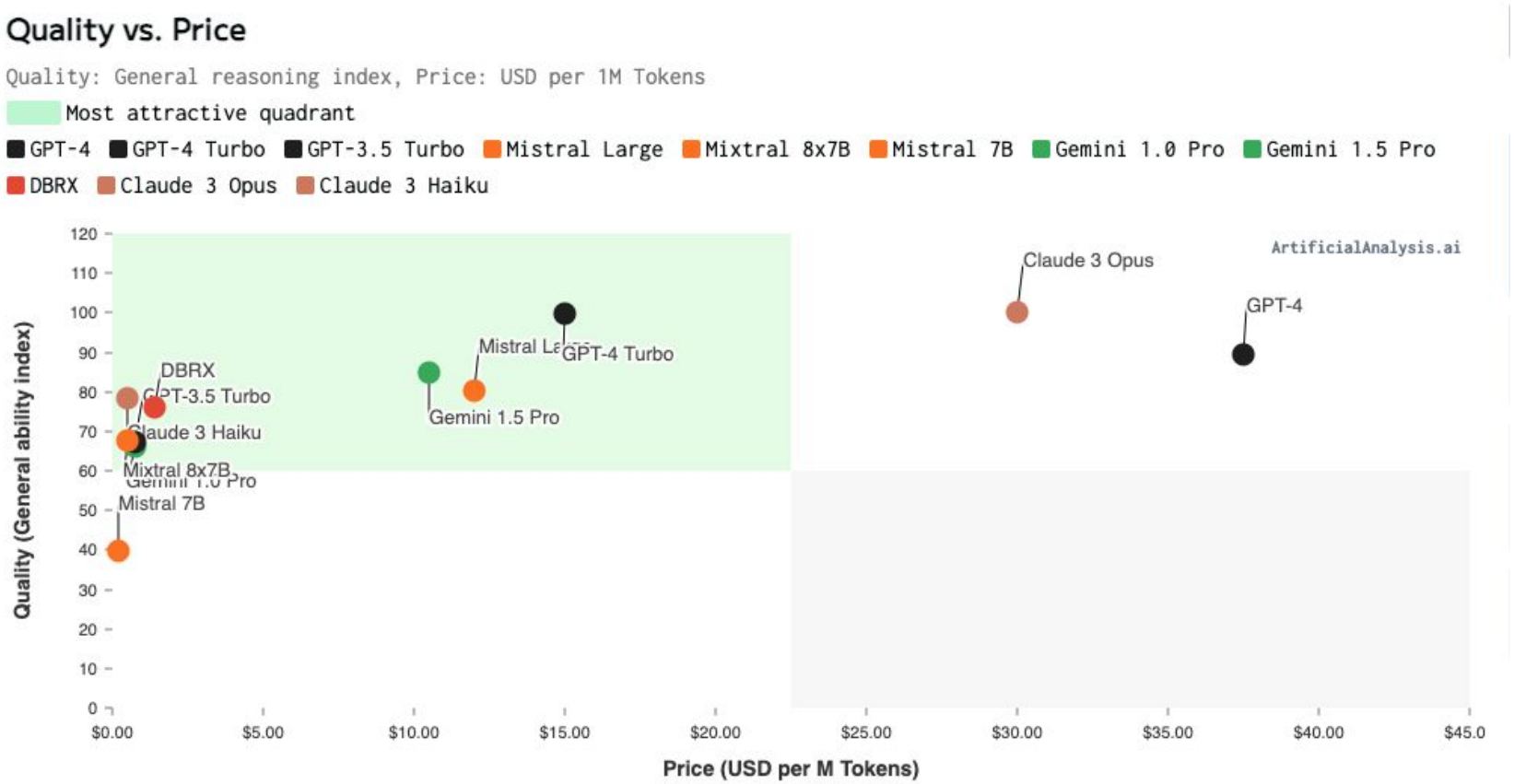
Leaderboard artificialanalysis.ai

Context window

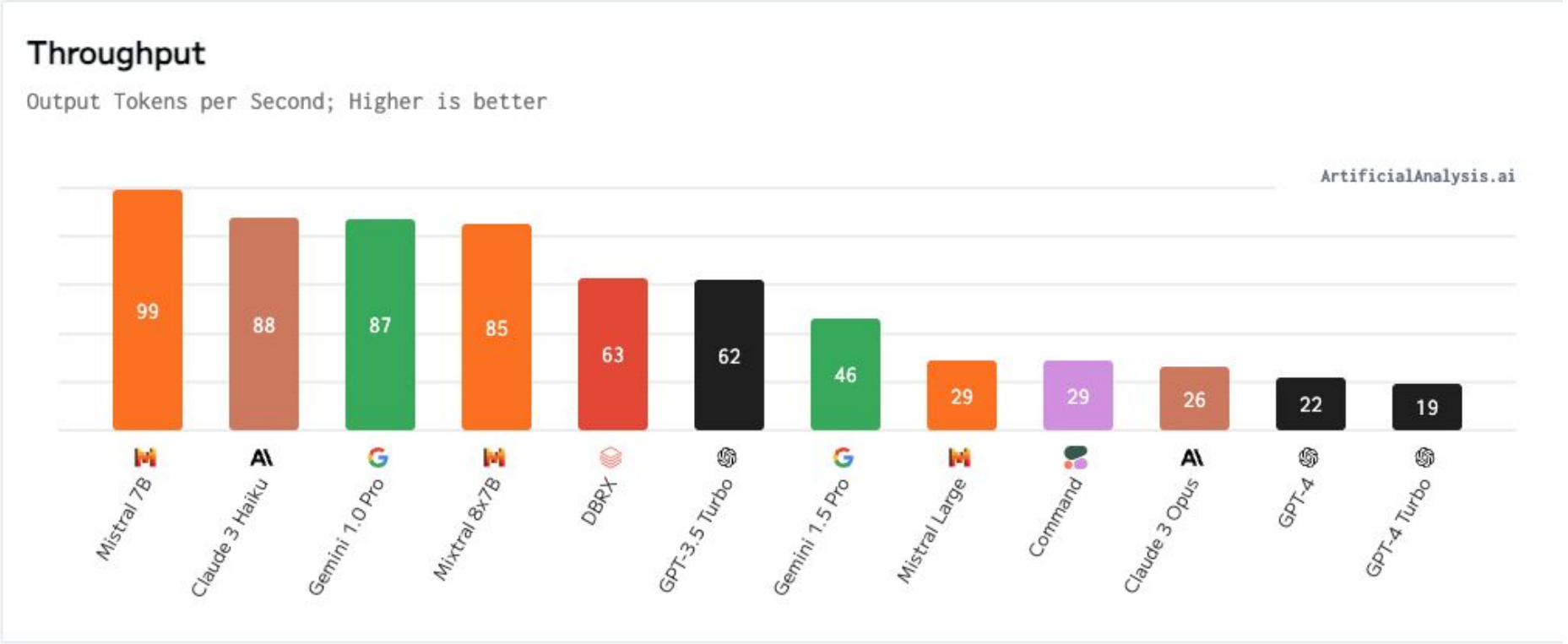
Context window: Tokens limit; Higher is better



Leaderboard artificialanalysis.ai



Leaderboard artificialanalysis.ai



Hugging Face 🤗

Training

github.com/karpathy/**llm.c**

Andrej Karpathy
@karpathy

Have you ever wanted to train LLMs in pure C without 245MB of PyTorch and 107MB of cPython? No? Well now you can! With llm.c:
github.com/karpathy/llm.c

To start, implements GPT-2 training on CPU/fp32 in only ~1,000 lines of clean code. It compiles and runs instantly, and exactly matches the PyTorch reference implementation.

I chose GPT-2 to start because it is the grand-daddy of LLMs, the first time the LLM stack was put together in a recognizably modern form, and with model weights available.

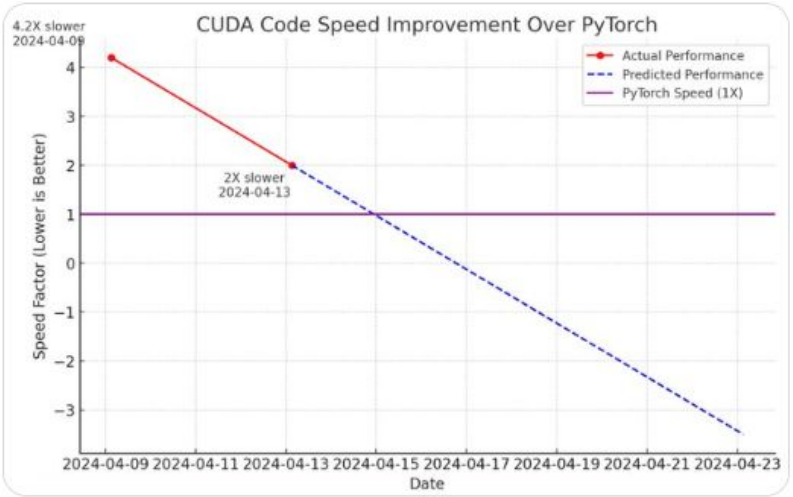
From github.com

github.com/karpathy/**llm.c**

Andrej Karpathy @karpathy · Apr 13

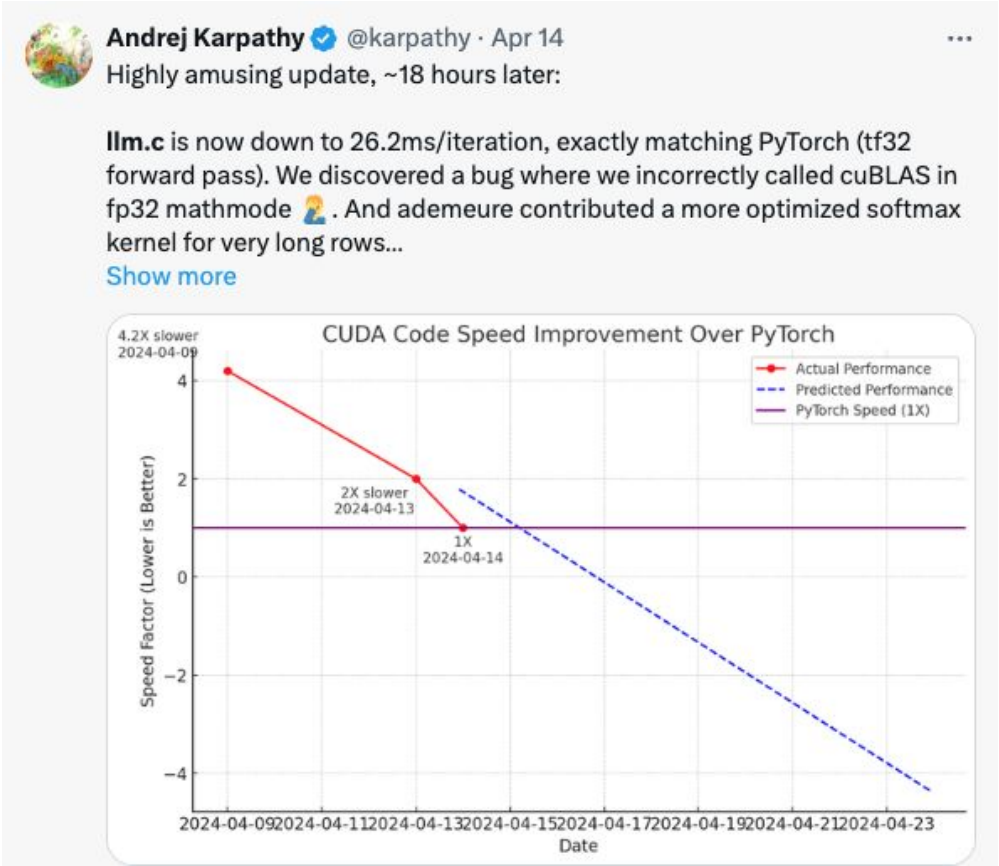
A few new CUDA hacker friends joined the effort and now **llm.c** is only 2X slower than PyTorch (fp32, forward pass) compared to 4 days ago, when it was at 4.2X slower 📈

The biggest improvements were:
- turn on TF32 (NVIDIA TensorFLoat-32) instead of FP32 for matmuls. This is a...
Show more



112 391 4.4K 1.3M

github.com/karpathy/**llm.c**



github.com/karpathy/**llm.c**

Stay tuned: https://twitter.com/**karpathy**

Deep dive into LLM*

*sorry 😅

Context window

To infinity and beyond

Knowledge from prompt vs memorisation

traduire en français: "Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum."

Claude-3-Opus Poe

Voici la traduction en français du texte latin "Lorem ipsum":

"Mais je dois vous expliquer comment toute cette idée erronée de dénoncer le plaisir et de louer la douleur est née et je vous donnerai une explication complète du système, et exposerai les véritables enseignements du grand explorateur de la vérité, le maître-constructeur du bonheur humain. Personne ne rejette, n'aime ou n'évite le plaisir lui-même, parce que c'est le plaisir, mais parce que ceux qui ne savent pas comment poursuivre le plaisir rationnellement rencontrent des conséquences qui sont extrêmement douloureuses. Il n'y a pas non plus d'homme qui aime ou poursuit ou désire obtenir de la douleur de lui-même, parce que c'est de la douleur, mais parce que parfois des circonstances se produisent dans lesquelles le labeur et la douleur peuvent lui procurer un grand plaisir. Pour prendre un exemple trivial, qui d'entre nous entreprend jamais un exercice physique laborieux, sauf pour en obtenir un avantage ? Mais qui a le droit de trouver à redire à un homme qui choisit de jouir d'un plaisir qui n'a pas de conséquences fâcheuses, ou à celui qui évite une douleur qui ne produit aucun plaisir résultant ?"

Il est à noter que le texte "Lorem ipsum" est en fait un faux-texte en latin utilisé comme bouche-trou dans la mise en page, le graphisme ou l'imprimerie. Son contenu n'a pas de véritable sens, il s'agit juste d'une extraction altérée d'un passage de Cicéron. La traduction ci-dessus correspond au texte original dont est tiré le "Lorem ipsum".

Produire un Lorem Ipsum en Français.

Claude-3-Opus Poe

Voici un exemple de faux-texte en français dans le style du "Lorem ipsum" latin :

"Loin, très loin, au delà des monts Mots, à mille lieues des pays Voyellie et Consonnia, demeurent les Bolos Bolos. Ils vivent en retrait, à Bourg-en-Lettres, sur les côtes de la Sémantique, un vaste océan de langues. Un petit ruisseau, du nom de Larousse, coule en leur lieu et les approvisionne en règlages nécessaires en tout genre. C'est un pays paradisiagmatique, dans lequel des pans entiers de phrases prémâchées vous volent litéralement tout cuit dans la bouche. Pas même la toute puissante Ponctuation ne régit les Bolos Bolos - une vie on ne peut moins orthodoxegraphique.

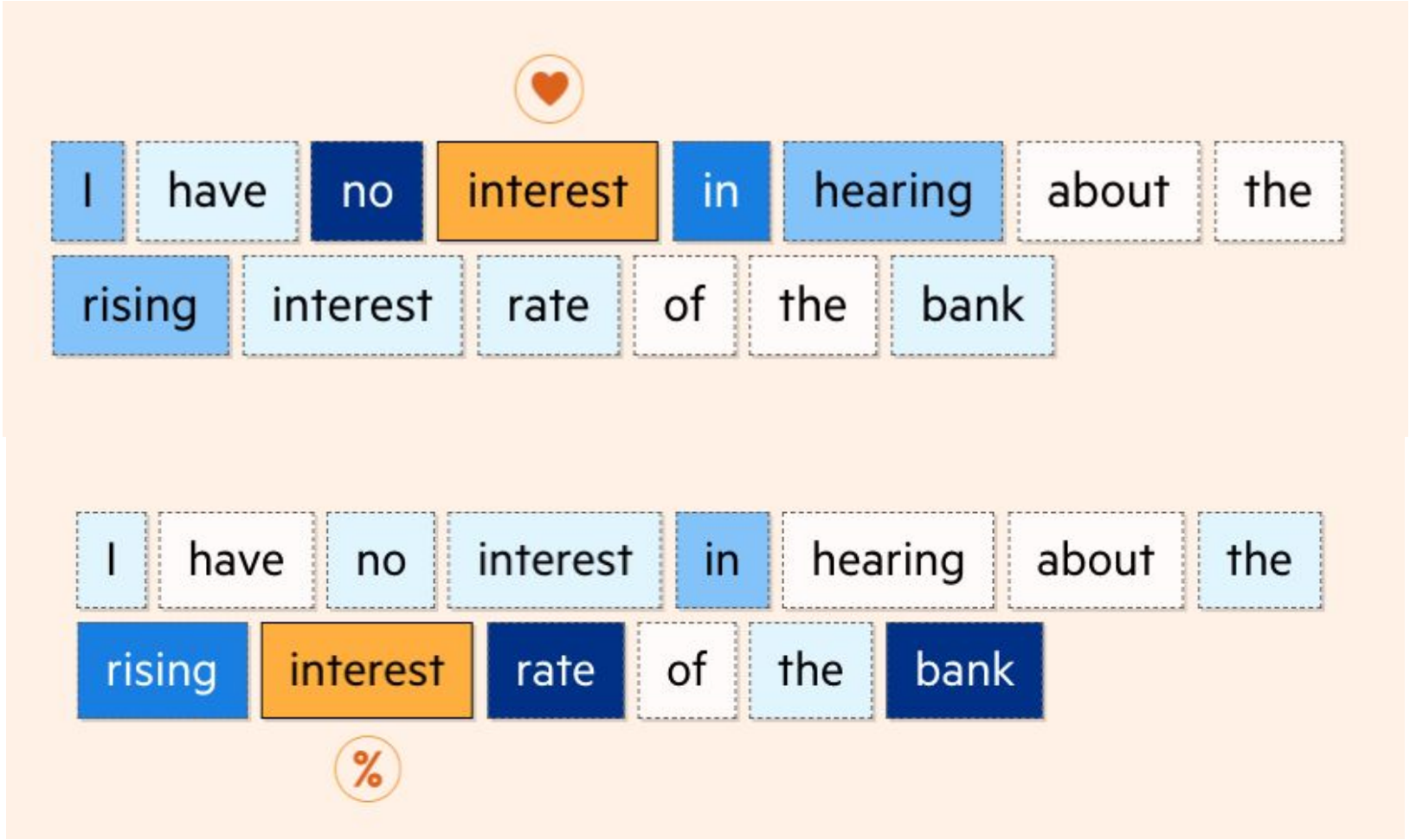
Un jour pourtant, une petite ligne de Bolo Bolo du nom de Lorem Ipsum décida de s'aventurer dans la vaste Grammaire. Le grand Oxymore voulut l'en dissuader, le prevenant que là-bas cela fourmillait de virgules retorses, de points d'exclamation sauvages, et de sournois points-virgules, mais ces conseils ne firent pas fléchir Lorem Ipsum : il empaqueta ses sept versales, glissa son initiale dans sa panse et se mit en route. Alors qu'il avait gravi les premiers contreforts de la chaîne des monts Mots, il jeta un dernier regard sur la skyline de Bourg-en-Lettres, la headline de Mots-clés, le subtitle de sa propre route, la ligne de flottaison de la Sémantique, et le dernier full-stop où officiait son père, le Maître des Signes..."

Bien entendu, comme pour le texte latin, ce paragraphe n'a pas de sens particulier. Il utilise des jeux de mots, des néologismes et des métaphores linguistiques pour donner l'apparence d'un vrai texte tout en n'étant qu'un simple bouche-trou pour la mise en page. Vous pouvez le modifier ou l'étendre à votre guise pour vos besoins de conception graphique.

Context window size is quadratic complexity

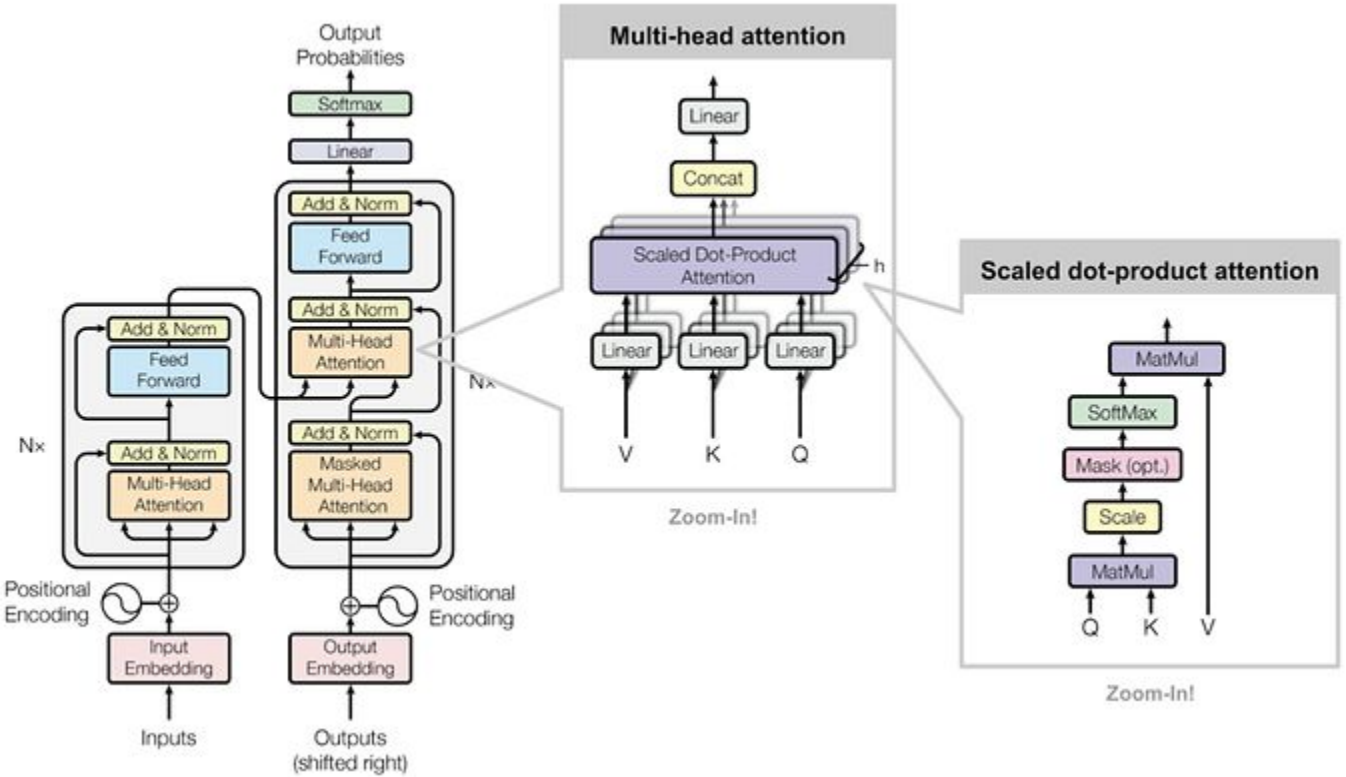
2k context window
vs
200k context window

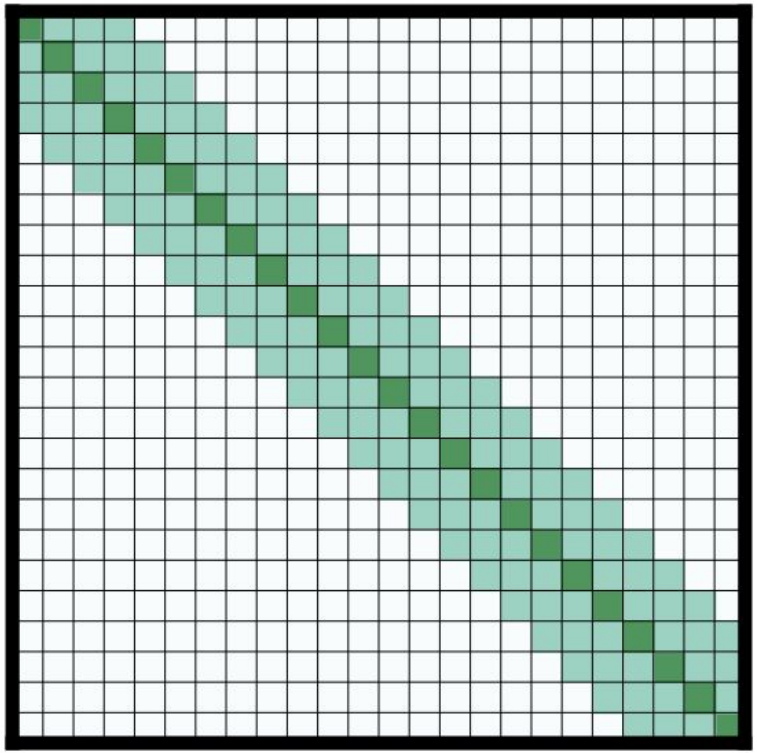
x10.000



https://ig.ft.com/generative-ai/

Attention & KV cache





(b) Sliding window attention

Meta: https://arxiv.org/pdf/2309.17453.pdf

StreamLLM

EFFICIENT STREAMING LANGUAGE MODELS WITH ATTENTION SINKS

Guangxuan Xiao[1*] **Yuandong Tian**[2] **Beidi Chen**[3] **Song Han**[1,4] **Mike Lewis**[2]

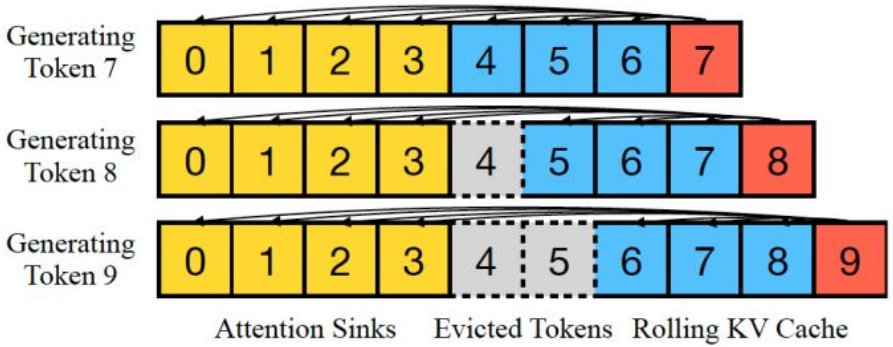
[1] Massachusetts Institute of Technology [2] Meta AI
[3] Carnegie Mellon University [4] NVIDIA
https://github.com/mit-han-lab/streaming-llm

ABSTRACT

Deploying Large Language Models (LLMs) in streaming applications such as multi-round dialogue, where long interactions are expected, is urgently needed but poses two major challenges. Firstly, during the decoding stage, caching previous tokens' Key and Value states (KV) consumes extensive memory. Secondly, popular LLMs cannot generalize to longer texts than the training sequence length. Window attention, where only the most recent KVs are cached, is a natural approach — but we show that it fails when the text length surpasses the cache size. We observe an interesting phenomenon, namely *attention sink*, that keeping the KV of initial tokens will largely recover the performance of window attention. In this paper, we first demonstrate that the emergence of *attention sink* is due to the strong attention scores towards initial tokens as a "sink" even if they are not semantically important. Based on the above analysis, we introduce StreamingLLM, an efficient framework that enables LLMs trained with a *finite length* attention window to generalize to *infinite sequence length* without any fine-tuning. We show that StreamingLLM can enable Llama-2, MPT, Falcon, and Pythia to perform stable and efficient language modeling with up to 4 million tokens and more. In addition, we discover that adding a placeholder token as a dedicated attention sink during pre-training can further improve streaming deployment. In streaming settings, StreamingLLM outperforms the sliding window recomputation baseline by up to 22.2× speedup. Code and datasets are provided in the link.

Meta: https://arxiv.org/pdf/2309.17453.pdf

StreamLLM



Meta: https://arxiv.org/abs/2404.08801

MEGALODON: Efficient LLM Pretraining and Inference with Unlimited Context Length

Xuezhe Ma[π*] Xiaomeng Yang[μ*] Wenhan Xiong[μ] Beidi Chen[κ] Lili Yu[μ]

Hao Zhang[δ] Jonathan May[π] Luke Zettlemoyer[μ] Omer Levy[μ] Chunting Zhou[μ*]

[μ]AI at Meta
[π]University of Southern California
[κ]Carnegie Mellon University
[δ]University of California San Diego

Abstract

The quadratic complexity and weak length extrapolation of Transformers limits their ability to scale to long sequences, and while sub-quadratic solutions like linear attention and state space models exist, they empirically underperform Transformers in pretraining efficiency and downstream task accuracy. We introduce MEGALODON, an neural architecture for efficient sequence modeling with unlimited context length. MEGALODON inherits the architecture of MEGA (exponential

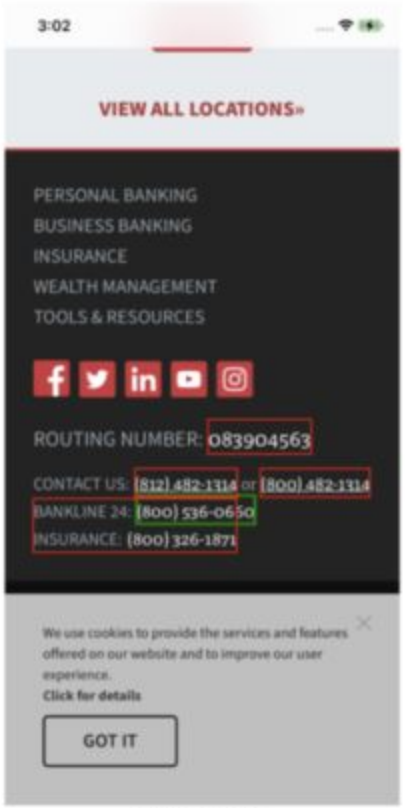
Google: https://arxiv.org/abs/2404.07143.pdf

Leave No Context Behind: Efficient Infinite Context Transformers with Infini-attention

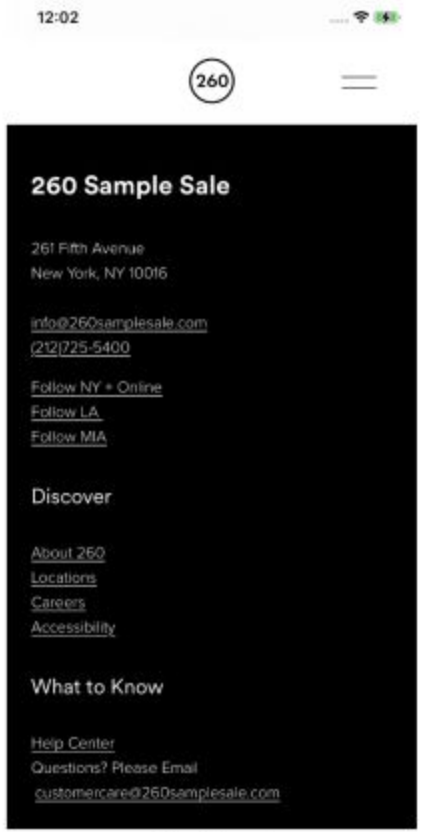
Tsendsuren Munkhdalai, Manaal Faruqui and Siddharth Gopal
Google
tsendsuren@google.com

Abstract

This work introduces an efficient method to scale Transformer-based Large Language Models (LLMs) to infinitely long inputs with bounded memory and computation. A key component in our proposed approach is a new attention technique dubbed Infini-attention. The Infini-attention incorporates a compressive memory into the vanilla attention mechanism and builds in both masked local attention and long-term linear attention mechanisms in a single Transformer block. We demonstrate the effectiveness of our approach on long-context language modeling benchmarks, 1M sequence length passkey context block retrieval and 500K length book summarization tasks with 1B and 8B LLMs. Our approach introduces minimal bounded memory parameters and enables fast streaming inference for LLMs.



germanamerican.com ii



260samplesale.com ii


ANNOUNCING NVIDIA BLACKWELL PLATFORM
FOR TRILLION-PARAMETER SCALE GENERATIVE AI
AI SUPERCHIP
208B Transistors
2nd GEN TRANSFORMER ENGINE
FP4/FP6 Tensor Core
5th GENERATION NVLINK
Scales to 576 GPUs
RAS ENGINE
100% In-System Self-Test
SECURE AI
Full Performance
Encryption & TEE
DECOMPRESSION ENGINE
800 GB/sec


NVIDIA
Revealed Project GROOT


"raise my windows and turn the radio on"
raise_windows(),
radio_on()

Intelligence artificielle : un accord de partenariat entre « Le Monde » et OpenAI

Cet accord pluriannuel, le premier entre un média français et un acteur majeur de l'IA, permettra à la société de s'appuyer sur le corpus du journal pour établir et fiabiliser les réponses de son outil ChatGPT, moyennant une source significative de revenus supplémentaires.

Par Louis Dreyfus (Président du directoire du « Monde ») et Jérôme Fenoglio (Directeur du « Monde »)

Publié le 13 mars 2024 à 18h30, modifié le 04 avril 2024 à 10h11 · Lecture 5 min. · Read in English

Dans le cadre de ses discussions avec les principaux acteurs de l'intelligence artificielle (IA) désireux de disposer d'une source de référence en langue française, *Le Monde* vient de conclure un accord pluriannuel avec la société OpenAI, connue pour son outil ChatGPT. Cet accord fera date puisqu'il est le premier signé entre un média français et un acteur majeur de cette industrie naissante. Il porte à la fois sur l'entraînement des modèles d'IA développés par l'entreprise américaine et sur les services de moteurs de réponse tels que ChatGPT. Il bénéficiera aux utilisateurs de cet outil en améliorant sa pertinence grâce à un contenu récent et faisant autorité sur une large palette de sujets d'actualité et de prescriptions éditoriales, tout en mettant en avant explicitement la contribution de notre média aux services d'OpenAI.

Cet accord s'inscrit dans la durée et est conçu comme un véritable partenariat. Il prévoit que nos équipes pourront, si elles le souhaitent, s'appuyer sur les technologies d'OpenAI pour

Les plus lus

1 Après l'attaque de l'Iran sur Israël, la Russie et la Chine font


OpenAI
VOICE
ENGINE

Arena Elo | Full Leaderboard

Total #models: **75**. Total #votes: **477471**. Last updated: March 26, 2024.

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

Rank	Model	Arena Elo	95% CI	Votes	Organization	License	Knowledge Cutoff
1	Claude 3 Opus	1253	+5/-5	33250	Anthropic	Proprietary	2023/8
1	GPT-4-1106-preview	1251	+4/-4	54141	OpenAI	Proprietary	2023/4
1	GPT-4-0125-preview	1248	+4/-4	34825	OpenAI	Proprietary	2023/12
4	Bard (Gemini Pro)	1203	+5/-7	12476	Google	Proprietary	Online
4	Claude 3 Sonnet	1198	+5/-5	32761	Anthropic	Proprietary	2023/8
6	GPT-4-0314	1185	+5/-4	33499	OpenAI	Proprietary	2021/9
6	Claude 3 Haiku	1179	+5/-5	18776	Anthropic	Proprietary	2023/8
8	GPT-4-0613	1158	+4/-5	51860	OpenAI	Proprietary	2021/9
8	Mistral-Large-2402	1157	+5/-4	26734	Mistral	Proprietary	Unknown
9	Qwen1.5-72B-Chat	1148	+5/-5	20211	Alibaba	Qianwen LICENSE	2024/2
10	Claude-1	1146	+6/-6	21908	Anthropic	Proprietary	Unknown
10	Mistral Medium	1145	+5/-4	26196	Mistral	Proprietary	Unknown
13	Starling-LM-7B-beta	1127	+9/-10	4270	Nexusflow	Apache-2.0	2024/3
13	Claude-2.0	1126	+7/-4	13543	Anthropic	Proprietary	Unknown
13	Gemini Pro (Dev API)	1125	+6/-6	14856	Google	Proprietary	2023/4


amazon
ANTHROP\C


arcads


GEN AI Coming Soon to Premiere Pro
Model
OpenAI
Prompt
View of city skyline in the rain at night
Generate
Remove Object
Pr


GEN AI Coming Soon to Premiere Pro
Model
OpenAI
Prompt
View of city skyline in the rain at night
Generate
Remove Object
Pr


Scott Wu, CEO / Cognition AI
Human Software Engineer

☠️ RIP Devin ?

▲ Debunking Devin: "First AI Software Engineer" Upwork Lie Exposed [video] (youtube.com)

292 points by smukherjee19 5 days ago | flag | hide | past | favorite | 43 comments

add comment

▲ mike_hearn 5 days ago | next [–]

An extremely solid and convincing rebuttal. Sad. I wonder what the Devin team will say in response, if anything. Summarizing the video:

• Devin is sold as being able to solve arbitrary Upwork tasks. In the video demo the problem it was asked to solve doesn't match the stated requirements of the customer (who asked for setup instructions, not code).

• Devin is shown fixing errors in the source of a GitHub repo, but the files it's shown editing don't actually exist in that repo and some of the errors its fixing are nonsensical, of the type that'd never be made by a human. Inference: Devin must be fixing bugs in files it has itself created, but that's not clearly indicated.

• There is no need to do any coding in the first place, because the README in the repository has all the instructions needed to achieve the task ready to go and they still work fine with only a one-line tweak, even though the repository is old. This is why the customer asked for instructions for how to run it on EC2 rather than for some coding. Devin didn't seem to read the README or understand that it only had to execute a couple of pre-existing Python scripts. The output in the video makes it look like the task was complex and sophisticated, with a long plan and many check boxes showing work completed, but the work was in fact pointless and redundant.

• Devin's code changes are bad, e.g. writing its own low level file read loop instead of using the standard library properly.

• Although the video makes it look like Devin did the task quickly, and the video creator was able to do the requested task in ~30 minutes, the timestamps in the chat show the task stretching over many hours and even into the next day.

• Devin does nonsensical shell commands like `head -n 5 foo | tail -n 5`

The strange mistakes lead to questions about what underlying model it's using. I don't think GPT-4 would make mistakes like that.

The Internet of Bugs guy is an AI fan and uses coding AI himself, but points out that the company behind it says you can "watch Devin get paid for doing work" which isn't actually supported by their video evidence when watched carefully.

reply

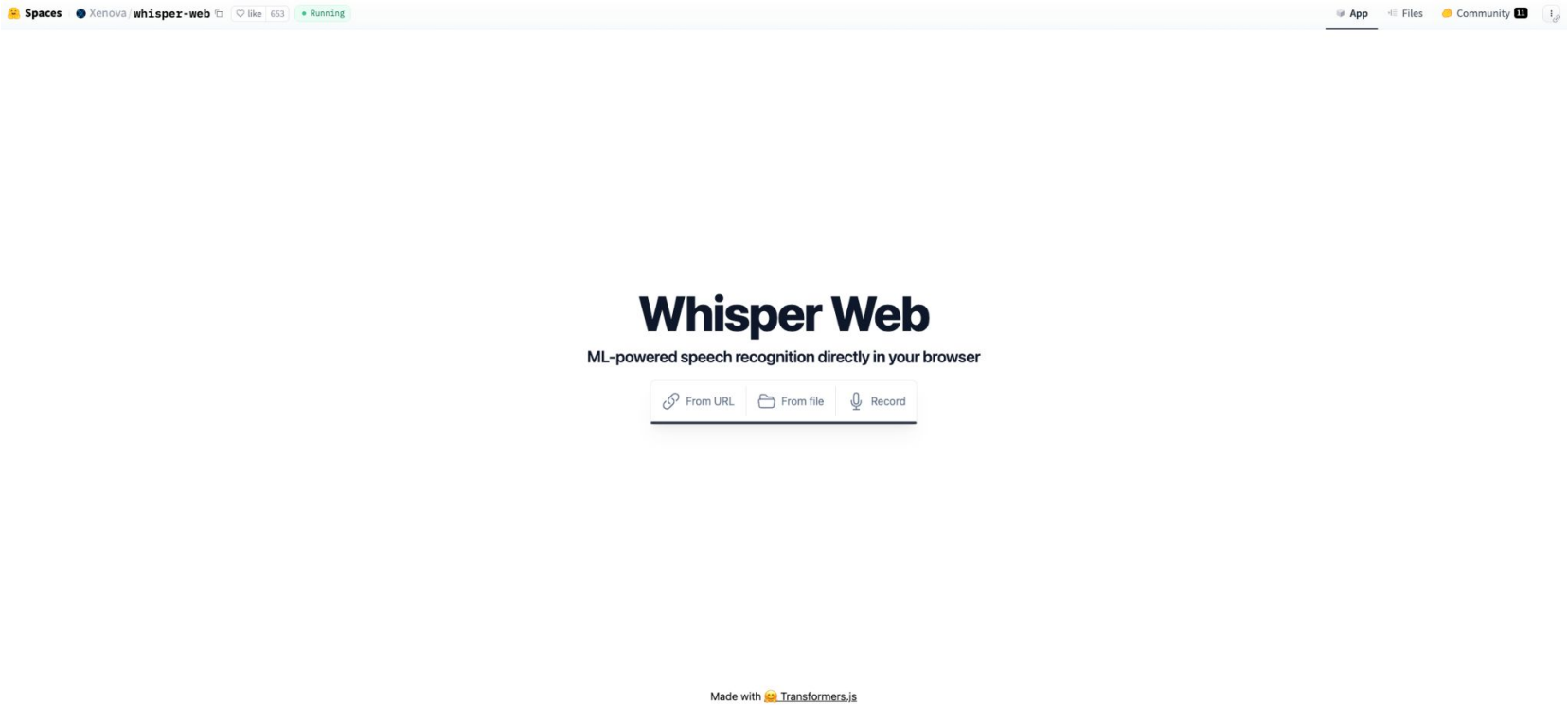
☠️ RIP BloombergGPT

https://arxiv.org/pdf/2305.05862.pdf

Comparison over LLMs. We are able to benchmark the performance of ChatGPT and GPT-4 with four other LLMs on five tasks with eight datasets. ChatGPT and GPT-4 significantly outperforms others in almost all datasets except the NER task. It is interesting to observe that both models perform better on financial NLP tasks than BloombergGPT, which was specifically trained on financial corpora. This might be due to the larger model size of the two models. Finally, GPT-4 constantly shows 10+% boost over ChatGPT in straightforward tasks such as Headlines and FiQA SA. For challenging tasks like RE and QA, GPT-4 can introduce 20-100% performance growth. This indicates that GPT-4 could be the first choice for financial NLP tasks before a more powerful LLM emerges.

The Worst Product I've Ever Reviewed... For Now

Marques Brownlee · 1.8M views · 15 h...



https://huggingface.co/spaces/Xenova/whisper-web?utm_source=www.therundown.ai&utm_medium=newsletter&utm_campaign=openai-fires-researchers-over-leaks

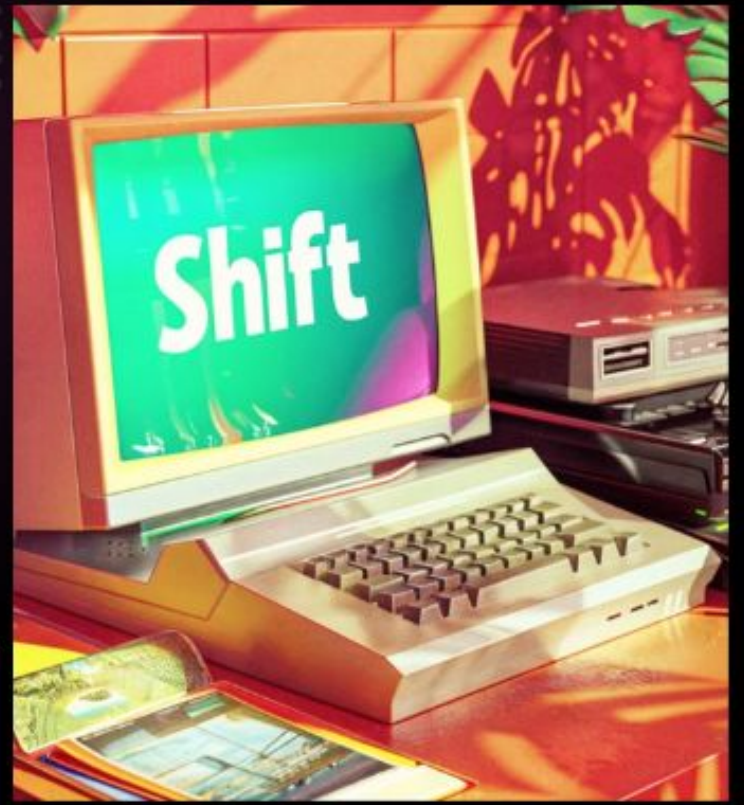
SHIFT

LE HACKATHON GEN AI

Du 31/05 au 02/06, tu as exactement 48h pour créer le futur en intégrant de l'IA Générative dans un produit tech 🔥

JE SUIS CHAUD

Soutenu par Google



CONF’

“Use case de fine-tuning d'un modèle de génération d’image”

Dmitry Kuzovkin
Head of AI @ Seelab

Matthieu Grosselin
CEO @ Seelab

seelab

Le **17/04/2024** à **19h**
Chez Lucca, **3 rue Michel Columb**, Nantes

lucca

Screeb RCA [sfeir] lonestone

Sources :

- https://x.com/AiExplorerFR/status/1771501534703927795
- https://twitter.com/MKBHD/status/1779641280110161957
- https://news.mit.edu/2024/ai-generates-high-quality-images-30-times-faster-single-step-0321?utm_source=www.therundown.ai&utm_medium=newsletter&utm_campaign=apple-s-ai-master-plan