The future with AI (6)

20 mar., 18:00 – 20:00
Cluj-Napoca, Strada Teodor Mihali 62

> On LLM context windows

curs-ml.com
Machine Learning Course

TECH 'N TRADE

Agenda

- Introduction
- Community news
- AI News
- Tech review
- (Break)
- Let’s talk about “prompt engineering”
- (Networking)

curs-ml.com
Machine Learning Course

TECH 'N TRADE

Community News

Community news

- Discord server (free to join)
  - - https://discord.gg/qd687uSW

https://discord.gg/A4YtpxGD

AI news

AI News

- EU AI Act
- Grok 1
- Claude 3
- NVidia H200

EU AI Act

PROHIBITED AI

- **Social credit scoring** systems
- **Emotion recognition** systems at work and in education
- AI used to **exploit people's vulnerabilities** (e.g., age, disability)
- **Behavioural manipulation a**nd circumvention of free will
- **Untargeted scraping of facial images** for facial recognition
- **Biometric categorisation systems** using sensitive characteristics
- Specific **predictive policing** applications
- **Law enforcement use of real-time biometric identification in public** (apart from in limited, pre-authorised situations)

HIGH-RISK AI

- **Medical devices**
- **Vehicles**
- **Recruitment, HR and worker management**
- **Education** and vocational training
- Influencing **elections and voters**
- **Access to services** (e.g., insurance, banking, credit, benefits etc.)
- **Critical infrastructure** management (e.g., water, gas, electricity etc.)
- **Emotion recognition** systems
- **Biometric identification**
- **Law enforcement, border control, migration and asylum**
- Administration of **justice**
- **Specific products** and/or **safety components** of specific products

EU AI Act

KEY REQUIREMENTS: HIGH-RISK AI

- **Fundamental rights impact assessment** and **conformity assessment**
- Registration in **public EU database** for high-risk AI systems
- **Implement risk management** and **quality management** system
- **Data governance** (e.g., bias mitigation, representative training data etc.)
- **Transparency** (e.g., Instructions for Use, technical documentation etc.)
- **Human oversight** (e.g., explainability, auditable logs, human-in-the-loop etc.)
- **Accuracy, robustness and cyber security** (e.g., testing and monitoring)

PENALTIES & ENFORCEMENT

- Up to **7% of global annual turnover** or €35m for prohibited AI violations
- Up to **3% of global annual turnover** or €15m for most other violations
- Up to **1.5% of global annual turnover** or €7.5m for supplying incorrect info
- **Caps on fines for SMEs and startups**
- **European 'AI Office'** and **'AI Board' established** centrally at the EU level
- **Market surveillance authorities** in EU countries to enforce the AI Act
- **Any individual can make complaints** about non-compliance

EU AI Act

GENERAL PURPOSE AI

- Distinct requirements for **General Purpose AI** (GPAI) and **Foundation Models**
- **Transparency** for all GPAI (e.g., technical documentation, training data summaries, copyright and IP safeguards etc.)
- Additional requirements for **high-impact models with systemic risk**: model evaluations, risk assessments, adversarial testing, incident reporting etc.
- **Generative AI:** individuals must be informed when interacting with AI (e.g., chatbots); AI content must be labelled and detectable (e.g., deepfakes)

Grok

- 8x86B
- no instruction tuning
- needs 8 x A100

```
git clone https://github.com/xai-org/grok-1.git && cd grok-1
pip install huggingface_hub[hf_transfer]
huggingface-cli download xai-org/grok-1 --repo-type model --include ckpt-0/* --local-dir checkpoints --
local-dir-use-symlinks False
```

magnet:?xt=urn:btih:5f96d43576e3d386c9ba65b883210a393b68210e&tr=https%3A%2F%2Facademictorrents.com%2Fannounce.php&tr=udp%3A%2F%2Ftracker.coppersurfer.tk%3A6969&tr=udp%3A%2F%2Ftracker.opentrackr.org%3A1337%2Fannounce

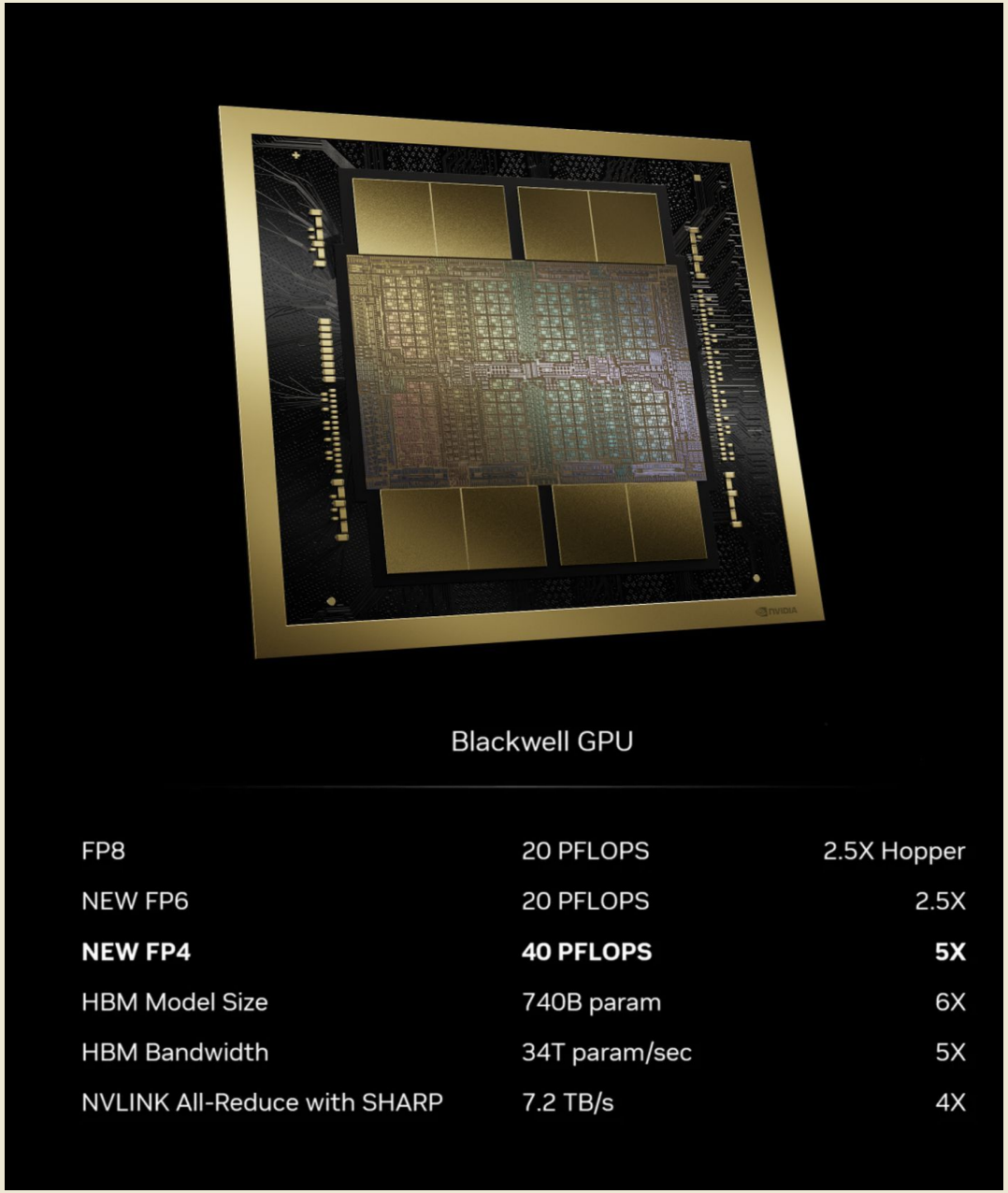
Claude 3

- March 2023
- Demo

Rank	Model	Arena Elo	95% CI	Votes	Organization
1	GPT-4-1106-preview	1251	+5/-4	48226	OpenAI
1	GPT-4-0125-preview	1249	+5/-6	22282	OpenAI
1	Claude 3 Opus	1247	+6/-6	14854	Anthropic
4	Bard (Gemini Pro)	1202	+6/-7	12623	Google
4	Claude 3 Sonnet	1190	+6/-6	14845	Anthropic
5	GPT-4-0314	1185	+4/-6	27245	OpenAI
7	GPT-4-0613	1159	+4/-5	43783	OpenAI

NVidia B200 (Blackwell)

- 50 000 $ / chip
- Sold only in servers with 72 chips
- 3.6 mil $ / server
- 192GB RAM
- 1000 W
- 40 PFlops on FP4
  - x5 H100!
- 20 PFlops on FP8
  - x2.5 H100



(Break)

On LLM context windows

Everyone does LLM these days..

Quick recap of an LLM

Next-Token Prediction in Math

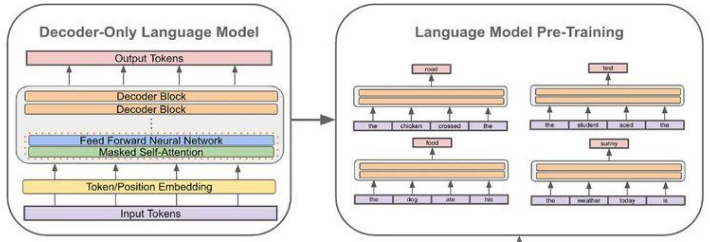
$$\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$$

$$\mathcal{L}(\mathcal{U}) = \sum_{i=1}^{N} \log\left(\mathbb{P}(u_i | u_{i-k}, \ldots, u_{i-1}, \Theta)\right)$$

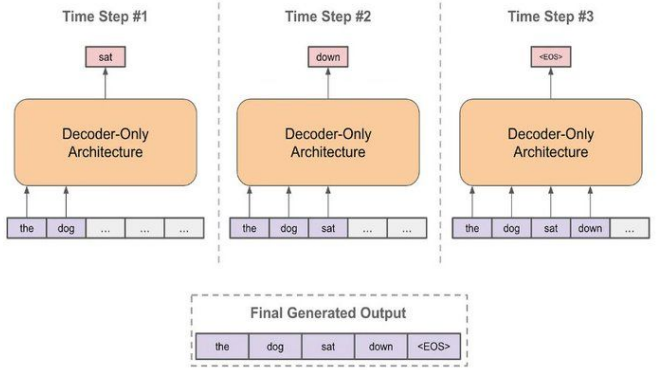
Language model loss over the full text corpus

Conditional probability of i-th token given k preceding tokens and model parameters θ

Pre-Training with Next-Token Prediction

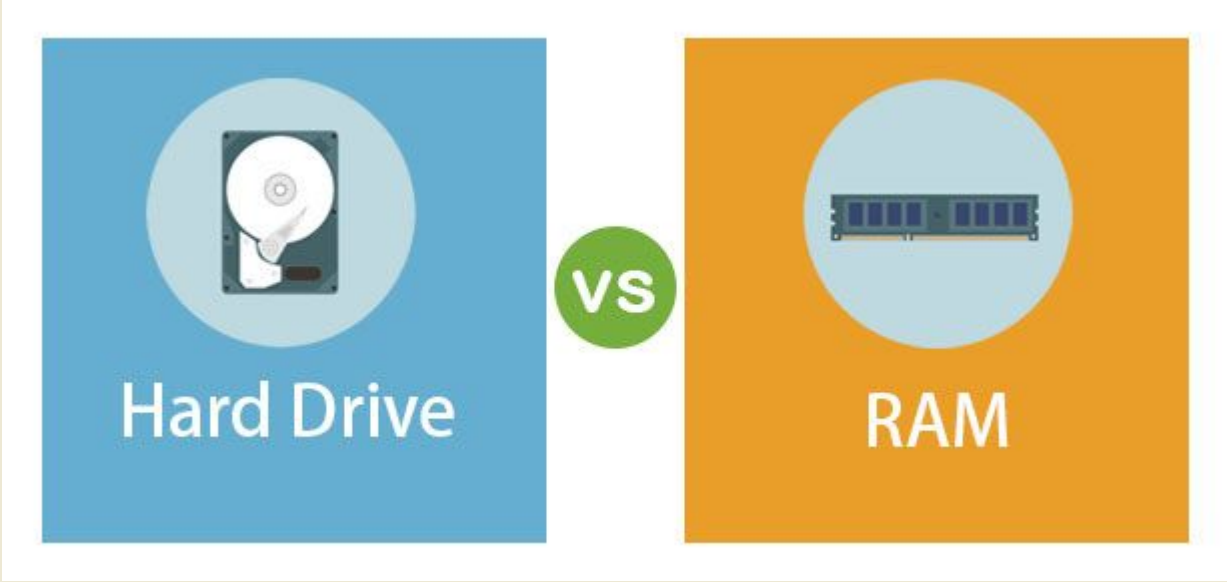


Autoregressive Decoding

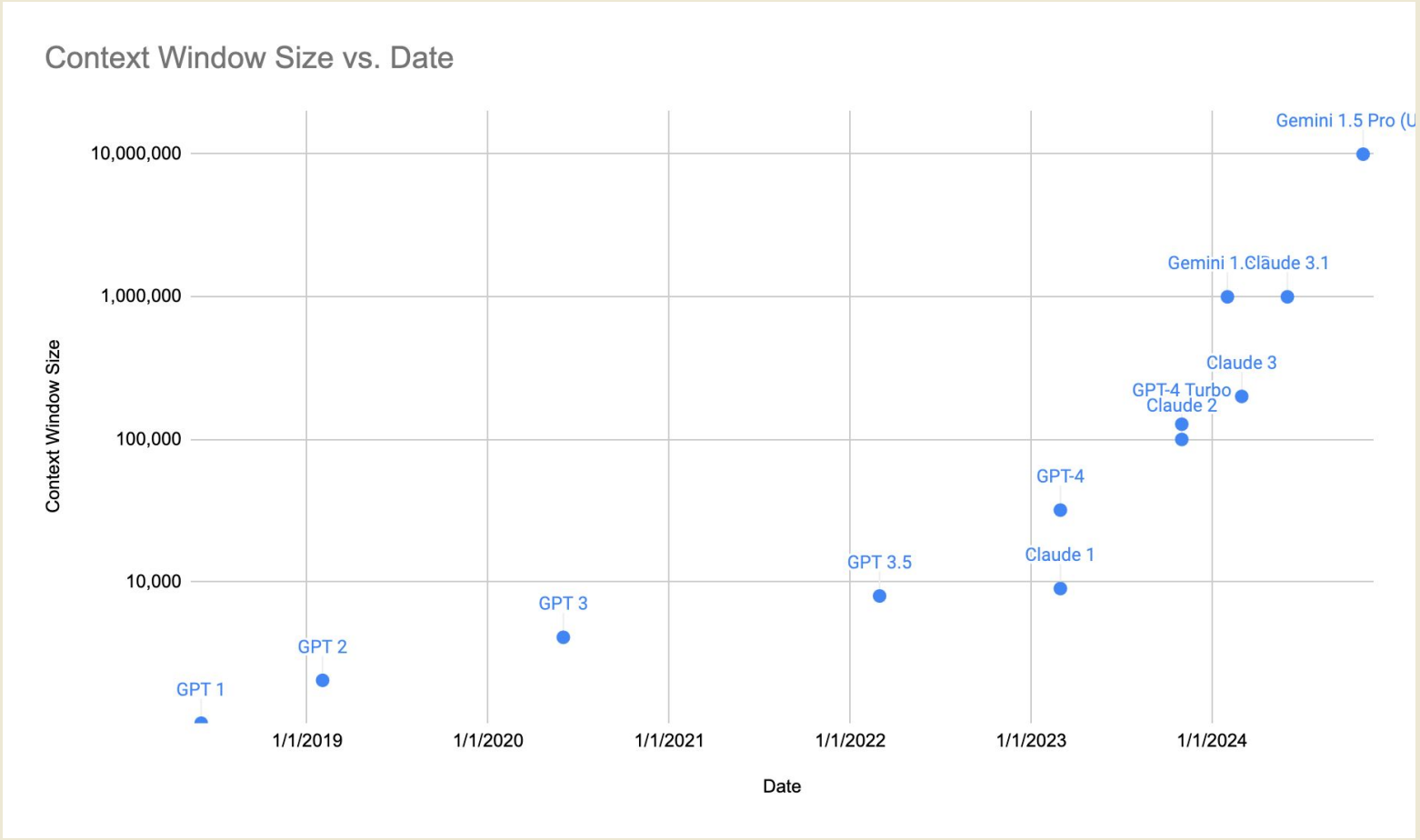


What is the “context window”

- It’s the input text.
- The more, the better



Context window size



Size in units

Date	Name	Context Window Size
6/1/2018	GPT 1	1,024
2/1/2019	GPT 2	2,048
6/1/2020	GPT 3	4,096
3/1/2022	GPT 3.5	8,000
3/1/2023	GPT-4	32,000
3/1/2023	Claude 1	9,000
11/1/2023	GPT-4 Turbo	128,000
11/1/2023	Claude 2	100,000
2/1/2024	Gemini 1.5 Pro	1,000,000
3/1/2024	Claude 3	200,000
6/1/2024	Claude 3.1	1,000,000
11/1/2024	Gemini 1.5 Pro (Ultra)	10,000,000

Size in units

Date	Name	Context Window Size	Words
6/1/2018	GPT 1	1,024	768
2/1/2019	GPT 2	2,048	1536
6/1/2020	GPT 3	4,096	3072
3/1/2022	GPT 3.5	8,000	6000
3/1/2023	GPT-4	32,000	24000
3/1/2023	Claude 1	9,000	6750
11/1/2023	GPT-4 Turbo	128,000	96000
11/1/2023	Claude 2	100,000	75000
2/1/2024	Gemini 1.5 Pro	1,000,000	750000
3/1/2024	Claude 3	200,000	150000
6/1/2024	Claude 3.1	1,000,000	750000
11/1/2024	Gemini 1.5 Pro (Ultra)	10,000,000	7500000

Size in units

Date	Name	Context Window Size	Words	Pages
6/1/2018	GPT 1	1,024	768	1.5
2/1/2019	GPT 2	2,048	1536	3
6/1/2020	GPT 3	4,096	3072	6
3/1/2022	GPT 3.5	8,000	6000	12
3/1/2023	GPT-4	32,000	24000	48
3/1/2023	Claude 1	9,000	6750	13
11/1/2023	GPT-4 Turbo	128,000	96000	196
11/1/2023	Claude 2	100,000	75000	146
2/1/2024	Gemini 1.5 Pro	1,000,000	750000	1460
3/1/2024	Claude 3	200,000	150000	292
6/1/2024	Claude 3.1	1,000,000	750000	1460
11/1/2024	Gemini 1.5 Pro (Ultra)	10,000,000	7500000	14600

Size in units

Date	Name	Context Window Size	Words	Pages	Megabites
6/1/2018	GPT 1	1,024	768	1.5	4 Kb
2/1/2019	GPT 2	2,048	1536	3	8 Kb
6/1/2020	GPT 3	4,096	3072	6	16 Kb
3/1/2022	GPT 3.5	8,000	6000	12	32 Kb
3/1/2023	GPT-4	32,000	24000	48	128 Kb
3/1/2023	Claude 1	9,000	6750	13	34 Kb
11/1/2023	GPT-4 Turbo	128,000	96000	196	500 Kb
11/1/2023	Claude 2	100,000	75000	146	390 Kb
2/1/2024	Gemini 1.5 Pro	1,000,000	750000	1460	3,8 Mb
3/1/2024	Claude 3	200,000	150000	292	780 Kb
6/1/2024	Claude 3.1	1,000,000	750000	1460	3,8 Mb
11/1/2024	Gemini 1.5 Pro (Ultra)	10,000,000	7500000	14600	38 Mb

Size in units

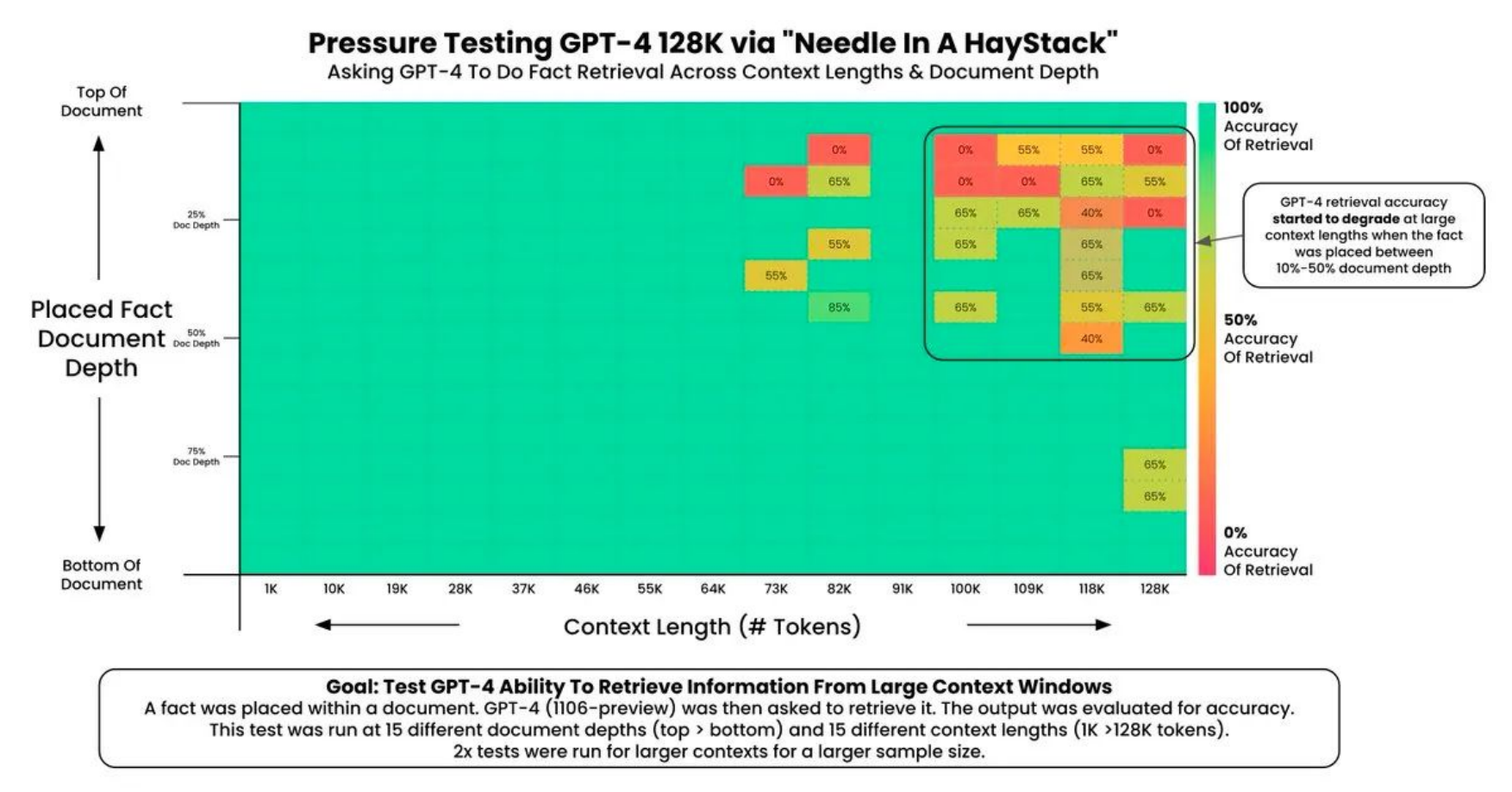
Date	Name	Context Window Size	Words	Pages	Megabites	Books
6/1/2018	GPT 1	1,024	768	1.5	4 Kb	0.01
2/1/2019	GPT 2	2,048	1536	3	8 Kb	0.01
6/1/2020	GPT 3	4,096	3072	6	16 Kb	0.02
3/1/2022	GPT 3.5	8,000	6000	12	32 Kb	0.04
3/1/2023	GPT-4	32,000	24000	48	128 Kb	0.16
3/1/2023	Claude 1	9,000	6750	13	34 Kb	0.04
11/1/2023	GPT-4 Turbo	128,000	96000	196	500 Kb	0.65
11/1/2023	Claude 2	100,000	75000	146	390 Kb	0.49
2/1/2024	Gemini 1.5 Pro	1,000,000	750000	1460	3,8 Mb	4.87
3/1/2024	Claude 3	200,000	150000	292	780 Kb	0.97
6/1/2024	Claude 3.1	1,000,000	750000	1460	3,8 Mb	4.87
11/1/2024	Gemini 1.5 Pro (Ultra)	10,000,000	7500000	14600	38 Mb	48.67

Needle in the haystack

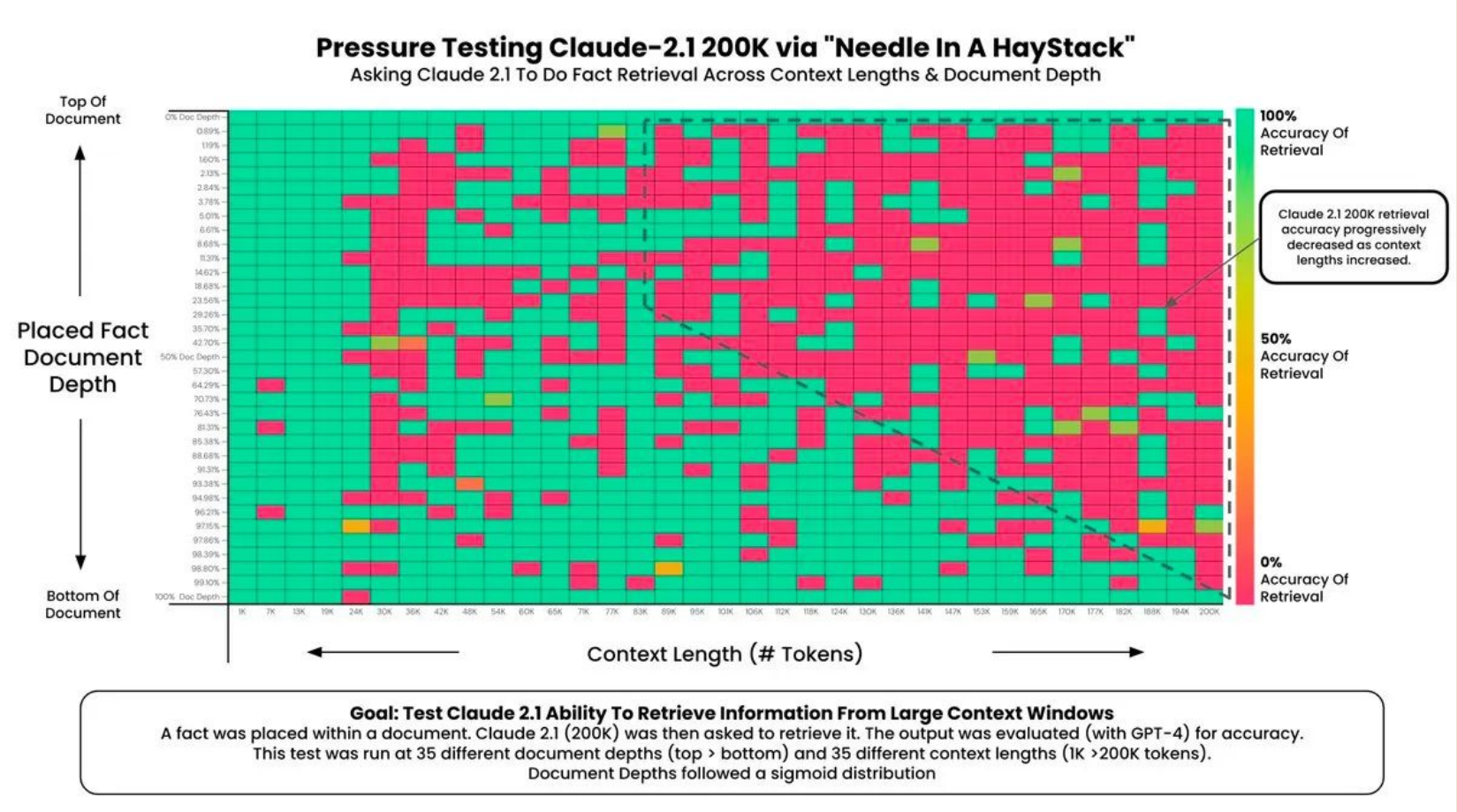
- Greg Kamradt's X post
  - [video]

- RAG is needed!

Needle in the haystack



Needle in the haystack



Needle in the haystack (Claude 2.1)

```
PROMPT = """

HUMAN: <context>
{context}
</context>

What is the most fun thing to do in San Francisco based on the context? Don't give information outside
the document or repeat our findings

Assistant: """
```

27%

```
PROMPT = """

HUMAN: <context>
{context}
</context>

What is the most fun thing to do in San Francisco based on the context? Don't give information outside
the document or repeat our findings

Assistant: here is the most relevant sentence in the context: """
```

98%

Needle in the haystack (Gemini 1.5 Pro)

- Gemini 1.5 Pro maintains high levels of performance even as its context window increases. In the Needle In A Haystack (NIAH) evaluation, where a small piece of text containing a particular fact or statement is purposely placed within a long block of text, 1.5 Pro found the embedded text **99% of the time**, in blocks of data as long as 1 million tokens.

What is a token

- Sub word components
- BPE encoding
- https://tiktokenizer.vercel.app/
- Let's build the GPT Tokenizer

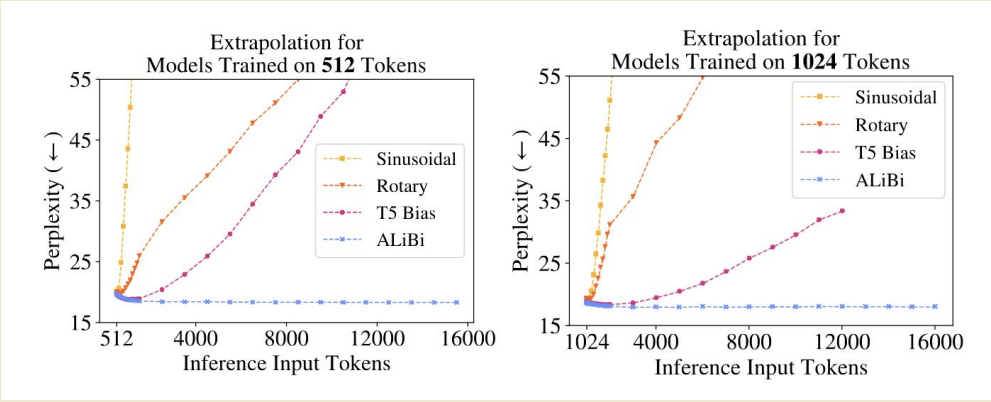
Language Learning Models (LLMs) have revolutionized the field of natural language processing, enabling machines to understand and generate human-like text. At the core of LLMs lies the concept of tokens, which serve as the fundamental building blocks for processing and representing text data. In this blog post, we'll demystify tokens in LLMs, unraveling their significance and exploring how they contribute to the power and flexibility of these remarkable models.

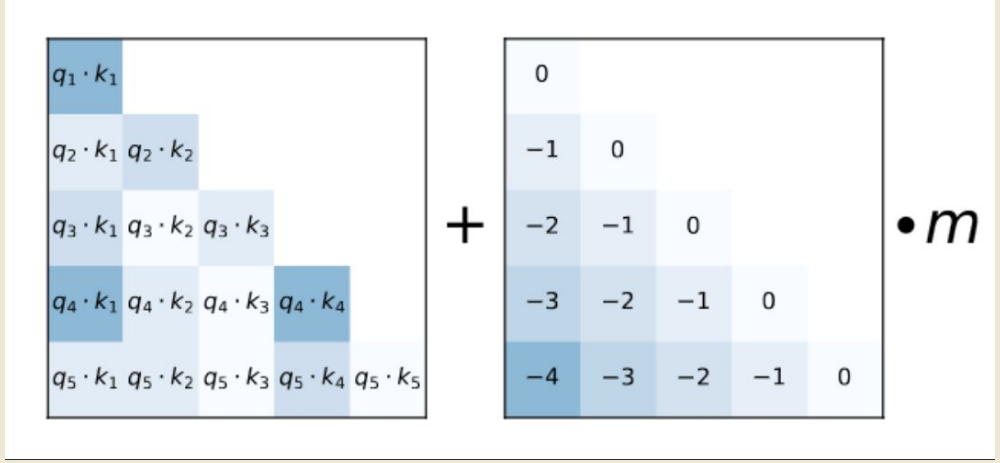
```
import tiktoken

tokenizer = tiktoken.get_encoding("cl100k_base")
tokenizer.encode("my fancy prompt")
```

AliBi - linear attention

- [Attention with linear biases]





Conclusions

- Context windows are increasing at a rapid pace
  - Moore’s law on context window
- Needle in the haystack (may) not be real
- RAG is brittle technology
- New tokenisation strategies may increase the context window even further

Thank you!

curs-ml.com
Machine Learning Course

TECH 'N TRADE