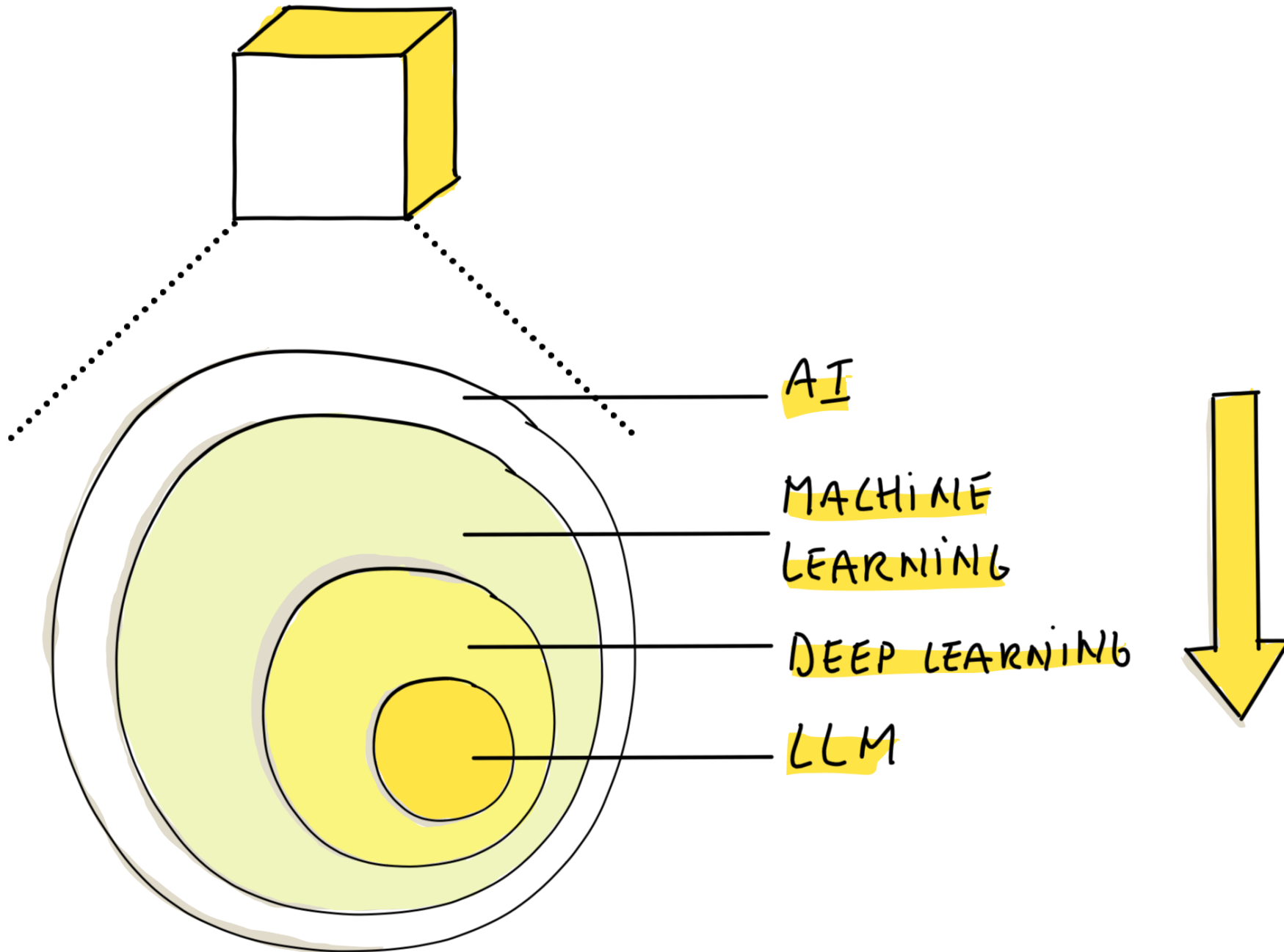


DEEL 5

LLMs





LLMS

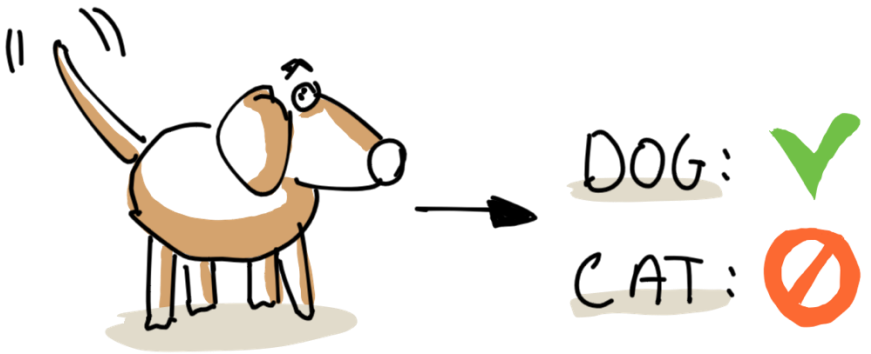
- Introduction
- LLM Training
- LLM Quality
- LLM Adoption in Belgium
- Popular LLMs



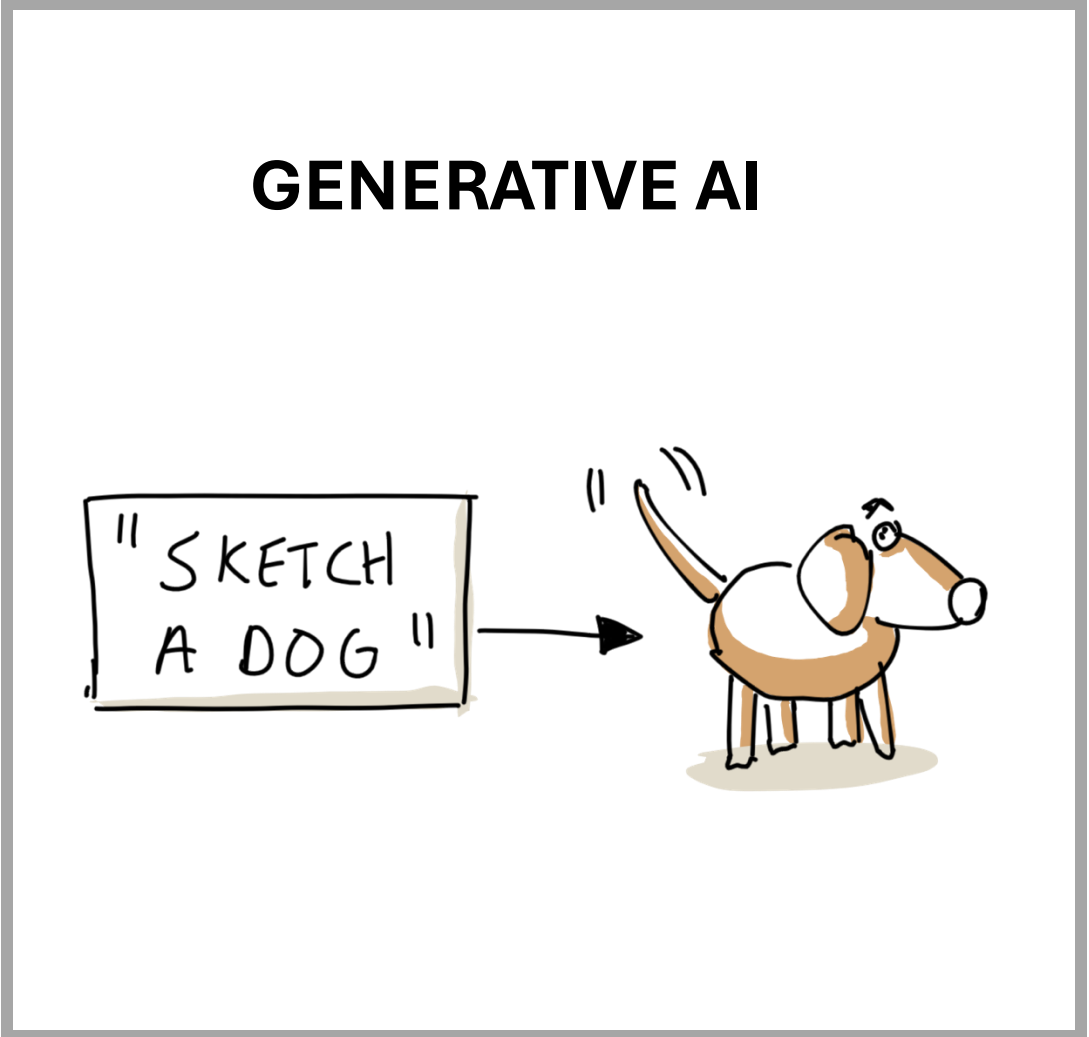
LLMS

- **Introduction**
- LLM Training
- LLM Quality
- LLM Adoption in Belgium
- Popular LLMs

DISCRIMINATIVE AI

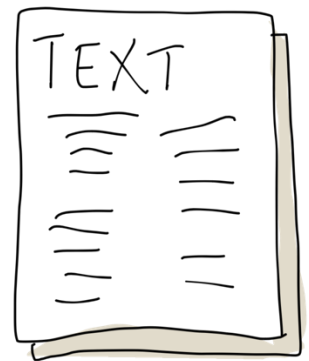


GENERATIVE AI



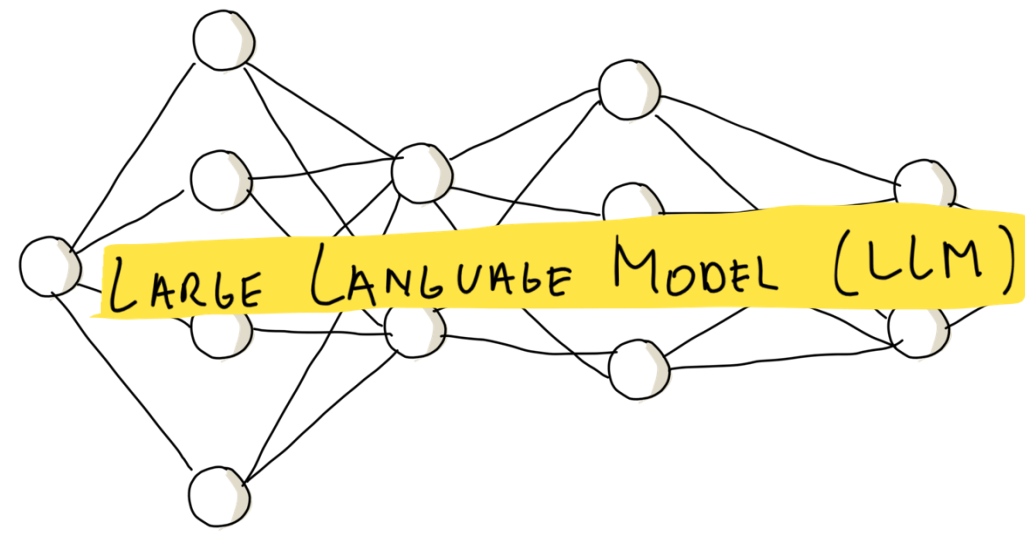


ANN



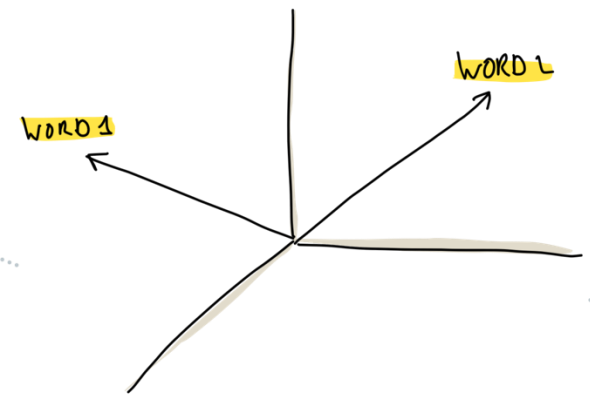
NUMBERS

$[1, 5, \dots]$

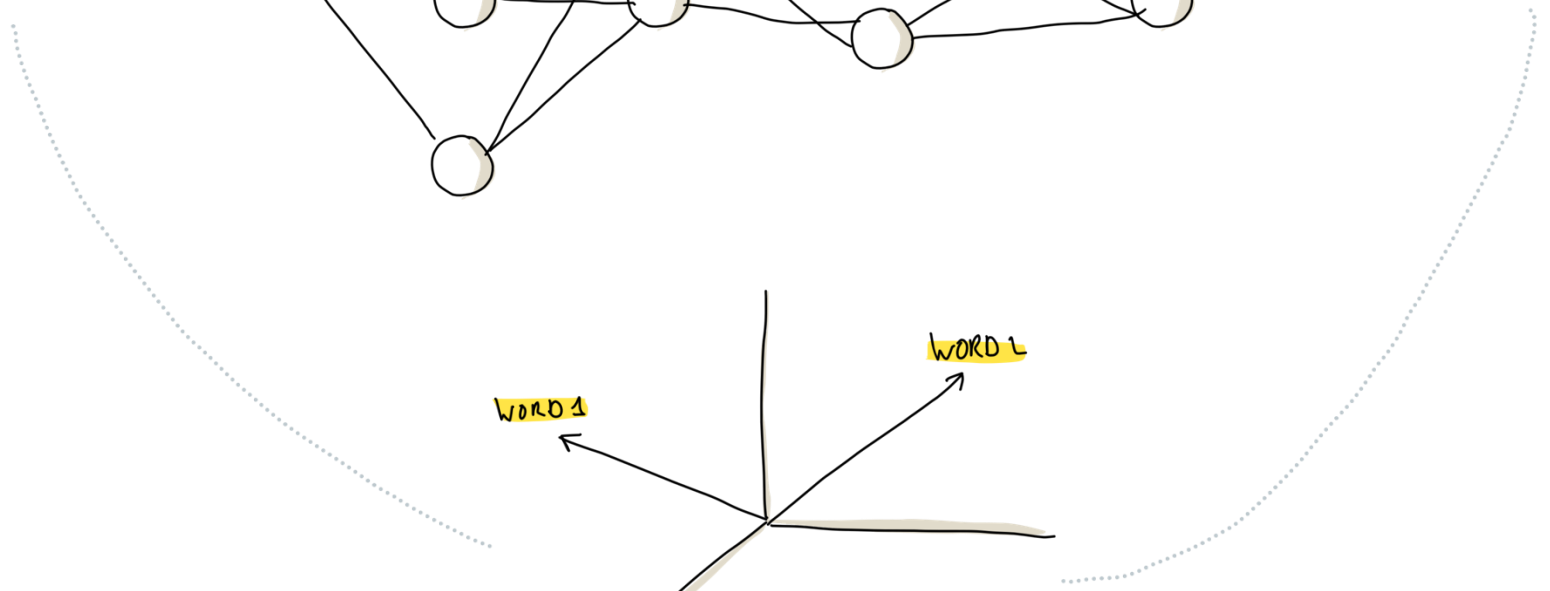


NUMBERS

$[8, 6, \dots]$



VECTOR SPACE





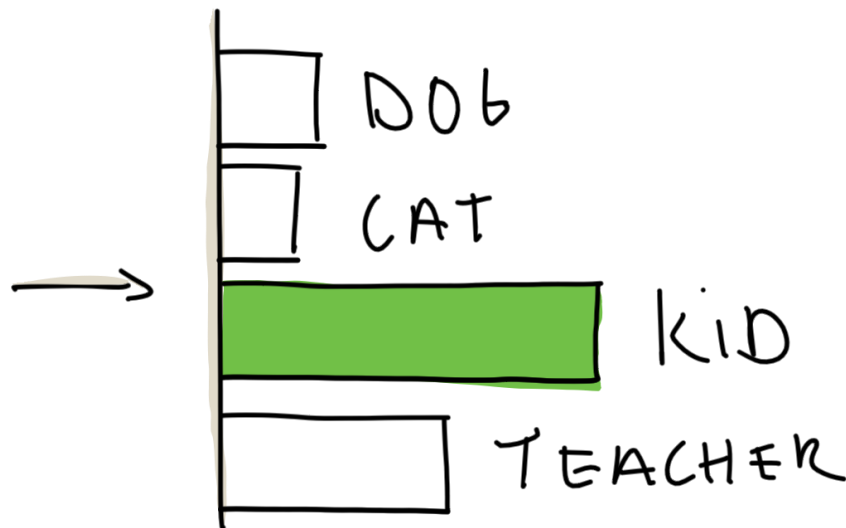
Predicting The Missing Word

THE ??
WENT TO THE
PLAYGROUND

SENTENCE



LLM

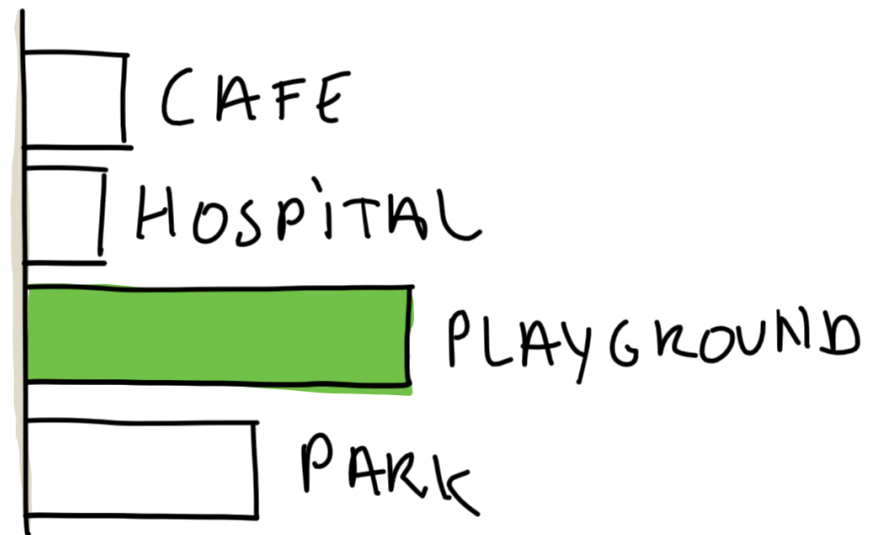
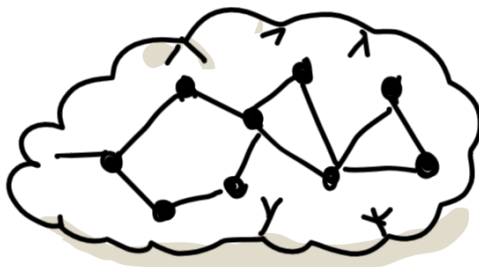


PREDICTED WORDS



Predicting The Next Word

THE KID
WENT TO THE



PREVIOUS WORDS

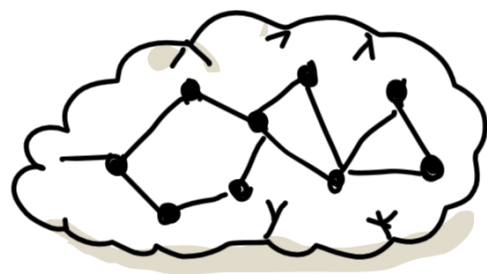
LLM

PREDICTED WORDS



Predicting The Next Word Sentence Paragraph ...

THE KID
WENT TO THE



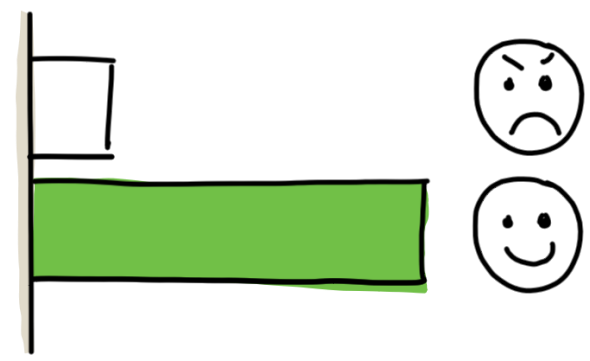
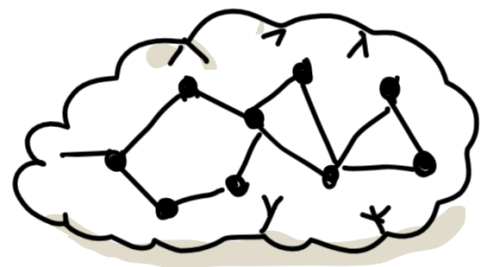
PREVIOUS WORDS

LLM

PREDICTED WORDS

Sentiment Analysis

MY EXPERIENCE
SO FAR HAS
BEEN FANTASTIC



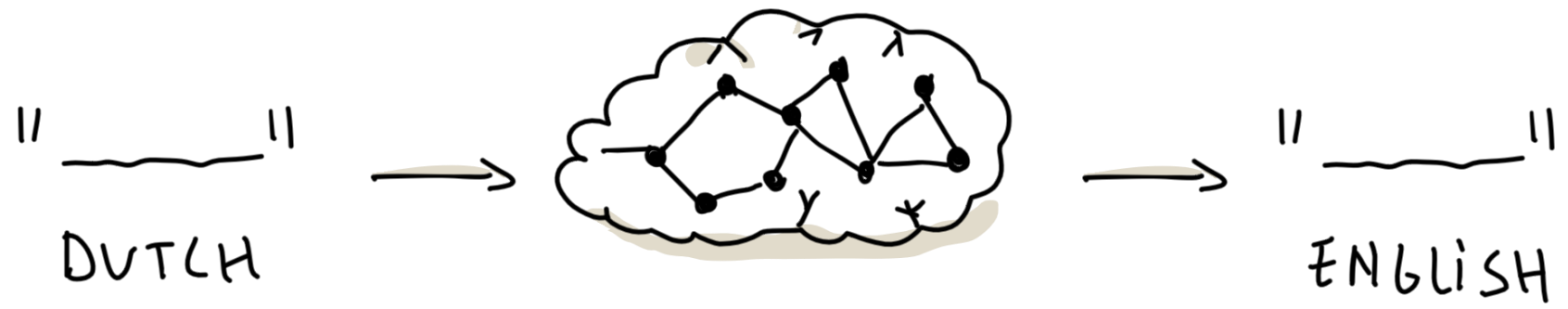
SENTENCE

LLM

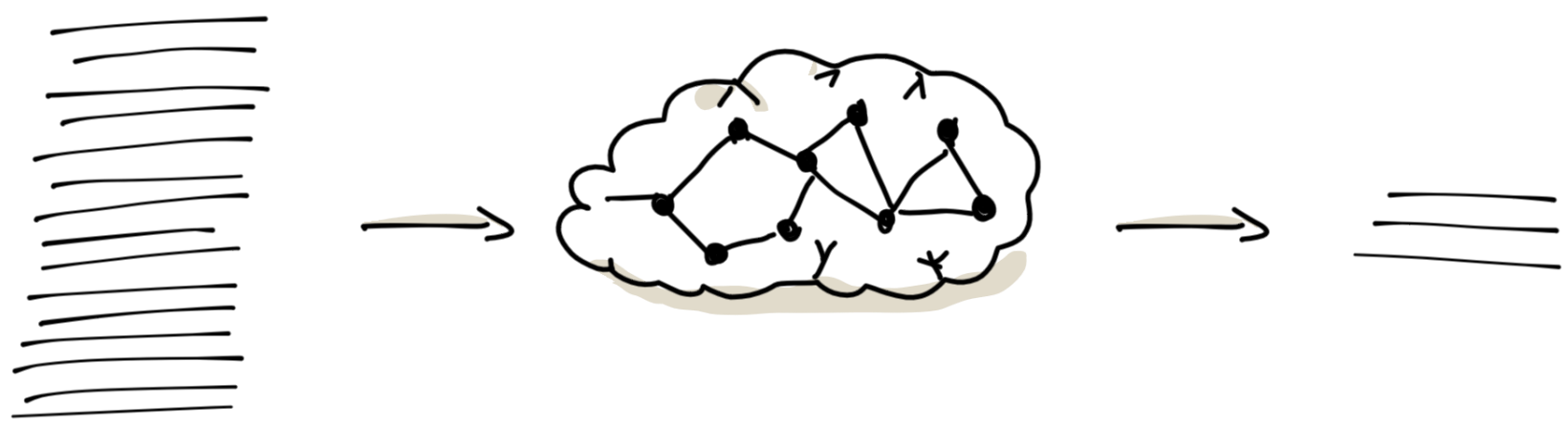
PREDICTED SENTIMENT



Translation



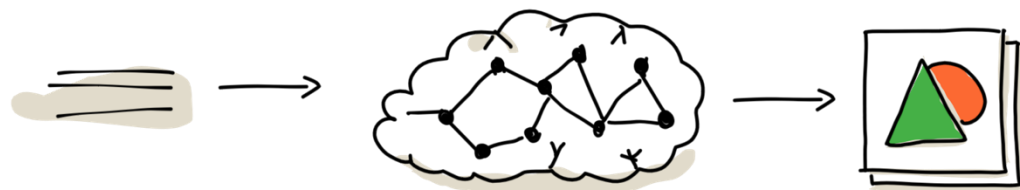
Summarization





Key Concept Behind the Scenes: **Vectorisation**







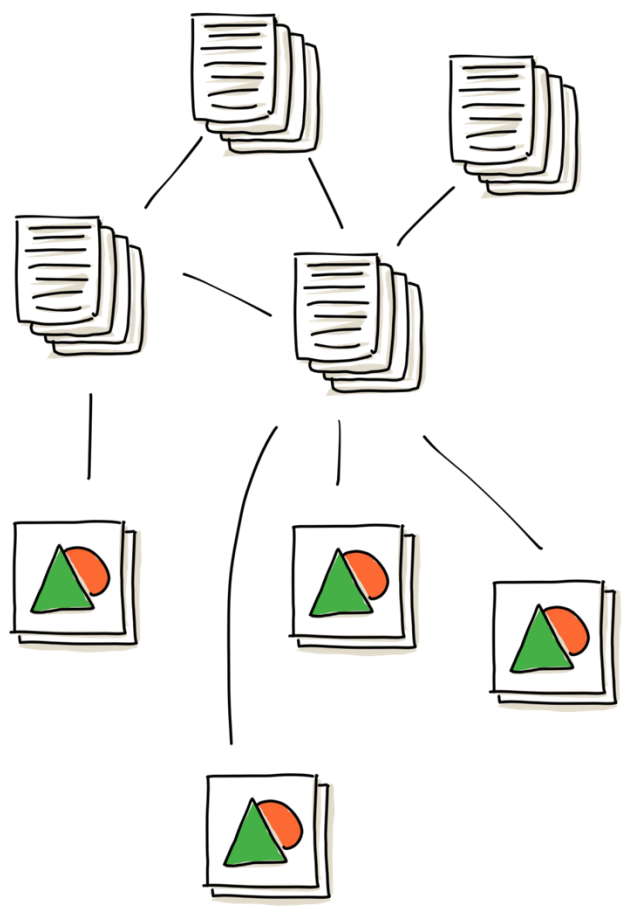
LLMS

- Introduction
- **LLM Training**
- LLM Quality
- LLM Adoption in Belgium
- Popular LLMs

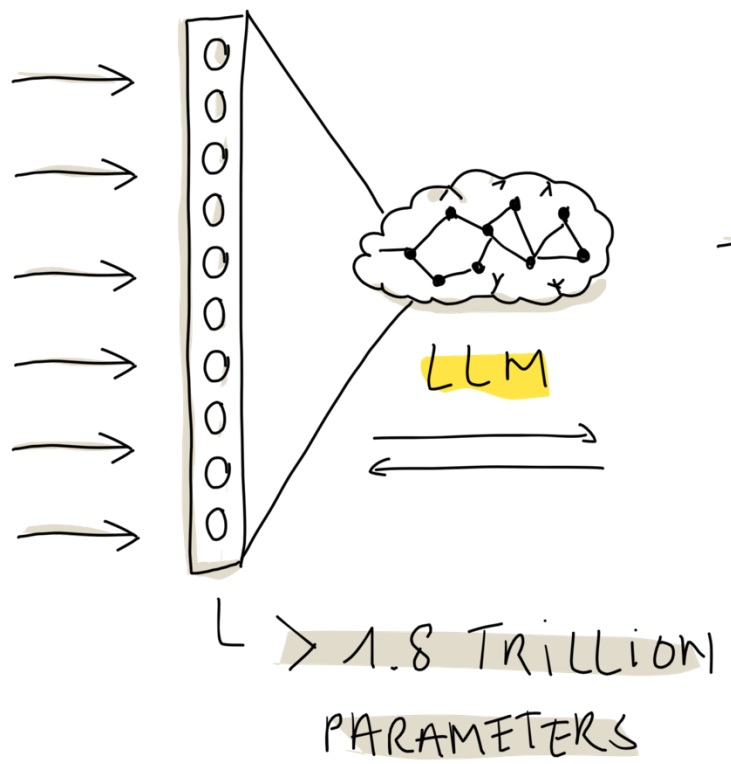


LLM Training?

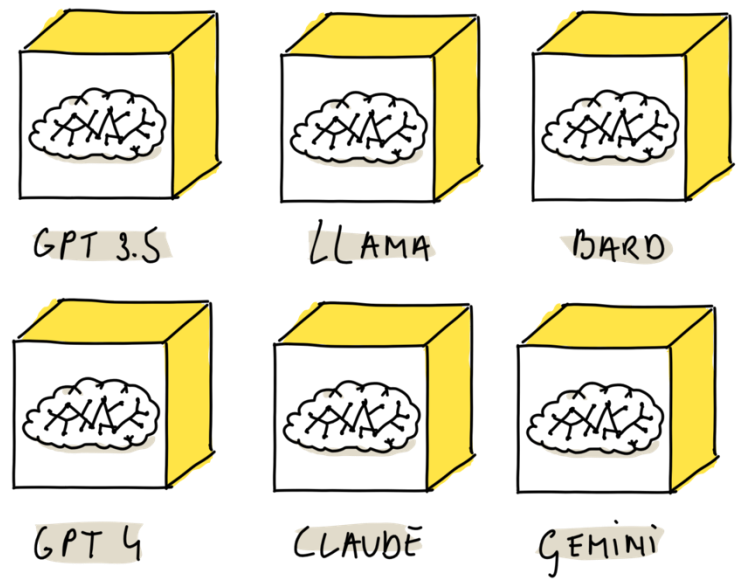
DATA



TRAINING



LLM MODEL





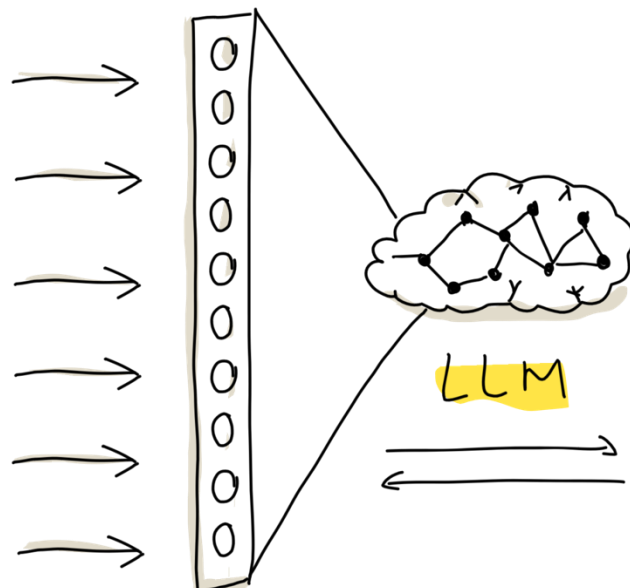
Training your own LLM (NanoGPT)

DATA



SHAKESPEARE
TEXT

TRAINING



$L < 1B$
PARAMETERS

LLM MODEL



NANO GPT-JM-0.1



Input Data: Shakespeare Text

First Citizen:
Before we proceed any further, hear me speak.

All:
Speak, speak.

First Citizen:
You are all resolved rather to die than to famish?

All:
Resolved. resolved.

First Citizen:
First, you know Caius Marcius is chief enemy to the people.

All:
We know't, we know't.

First Citizen:
Let us kill him, and we'll have corn at our own price.
Is't a verdict?

All:
No more talking on't; let it be done: away, away!

Second Citizen:
One word, good citizens.

First Citizen:
We are accounted poor citizens, the patricians good.
What authority surfeits on would relieve us: if they
would yield us but the superfluity, while it were
wholesome, we might guess they relieved us humanely;
but they think we are too dear: the leanness that
afflicts us, the object of our misery, is as an
inventory to particularise their abundance; our
sufferance is a gain to them Let us revenge this with
our pikes, ere we become rakes: for the gods know I
speak this in hunger for bread, not in thirst for revenge.

Second Citizen:
Would you proceed especially against Caius Marcius?

All:
Against him first: he's a very dog to the commonalty.

Second Citizen:
Consider you what services he has done for his country?

First Citizen:
Very well; and could be content to give him good
report fort, but that he pays himself with being proud.

Second Citizen:
Nay, but speak not maliciously.

First Citizen:
Well, sir, what answer made the belly?

MENENIUS:
Sir, I shall tell you. With a kind of smile,
Which ne'er came from the lungs, but even thus--
For, look you, I may make the belly smile
As well as speak--it tauntingly replied
To the discontented members, the mutinous parts
That envied his receipt; even so most fitly
As you malign our senators for that
They are not such as you.

First Citizen:
Your belly's answer? What!
The kingly-crowned head, the vigilant eye,
The counsellor heart, the arm our soldier,
Our steed the leg, the tongue our trumpeter.
With other muniments and petty helps
In this our fabric, if that they--

MENENIUS:
What then?
'Fore me, this fellow speaks! What then? what then?

First Citizen:
Should by the cormorant belly be restrain'd,
Who is the sink o' the body,--

MENENIUS:
Well, what then?

First Citizen:
The former agents, if they did complain,
What could the belly answer?

MENENIUS:
I will tell you
If you'll bestow a small--of what you have little--
Patience awhile, you'll hear the belly's answer.

First Citizen:
Ye're long about it.

BRUTUS:
Fame, at the which he aims,
In whom already he's well graced, can not
Better be held nor more attain'd than by
A place below the first: for what miscarries
Shall be the general's fault, though he perform
To the utmost of a man, and giddy censure
Will then cry out of Marcius 'O if he
Had borne the business!'

SICINIUS:
Besides, if things go well,
Opinion that so sticks on Marcius shall
Of his demerits rob Cominius.

BRUTUS:
Come:
Half all Cominius' honours are to Marcius.
Though Marcius earned them not, and all his faults
To Marcius shall be honours, though indeed
In aught he merit not.

SICINIUS:
Let's hence, and hear
How the dispatch is made, and in what fashion,
More than his singularity, he goes
Upon this present action.

BRUTUS:
Let's along.

First Senator:
So, your opinion is, Aufidius,
That they of Rome are entered in our counsels
And know how we proceed.

AUFIDIUS:
Is it not yours?
What ever have been thought on in this state,
That could be brought to bodily act ere Rome
Had circumvention? 'Tis not four days gone
Since I heard thence; these are the words: I think
I have the letter here; yes, here it is.
'They have press'd a power, but it is not known
Whether for east or west: the dearth is great;
The people mutinous; and it is rumour'd,
Cominius, Marcius your old enemy,
Who is of Rome worse hated than of you,
And Titus Lartius, a most valiant Roman,
These three lead on this preparation
Whither 'tis bent: most likely 'tis for you:
Consider of it.'

-- -- --



Training: After 200 iterations

```
Overriding: out_dir = out-shakespeare-char  
Overriding: device = mps  
number of parameters: 0.80M  
Loading meta from data/shakespeare_char/meta.pkl...
```

```
Upastat don to che withe be upornie hid.  
  
HUke CORCHe tage to to she ear or cund to of havr  
The kncoous of in prey wall, the ous,  
Thas inds beke afl-hices it of woul see awile,  
Is le strel with shem of to the live;  
I mor ther toon he muse and's of son yeret f,  
She gien your, to your and to heur is lat rukin  
ike your mad, with thour my thoun yougme,  
ear thou that frown the ling, have all won im to,  
The man wee dap to hin and strreatan thous lath's wert yere,  
Non to of son, Kame torr ear it to f  
-----
```



Training: After 3000 iterations

```
Overriding: out_dir = out-shakespeare-char  
number of parameters: 10.65M  
Loading meta from data/shakespeare_char/meta.pkl...
```

```
DUKE VINCENTIO:  
I prove thy hands as a villain,  
And nor I am nod nurse! I was be so with  
him. A prison Padua abroad is the poor and soon a cursed buckle,  
The flesh of it were not so longer have bless, and what you shall be  
Swear that. What having is him the lask of our blood  
Private and late, that e'er earling strange compassion  
Her blood, where he had would be her comfort of the ward  
with form her man.
```

```
First Servingman:  
Marry, sir, where's the people, which you were he of the bosoms.
```

```
Third Se  
-----
```

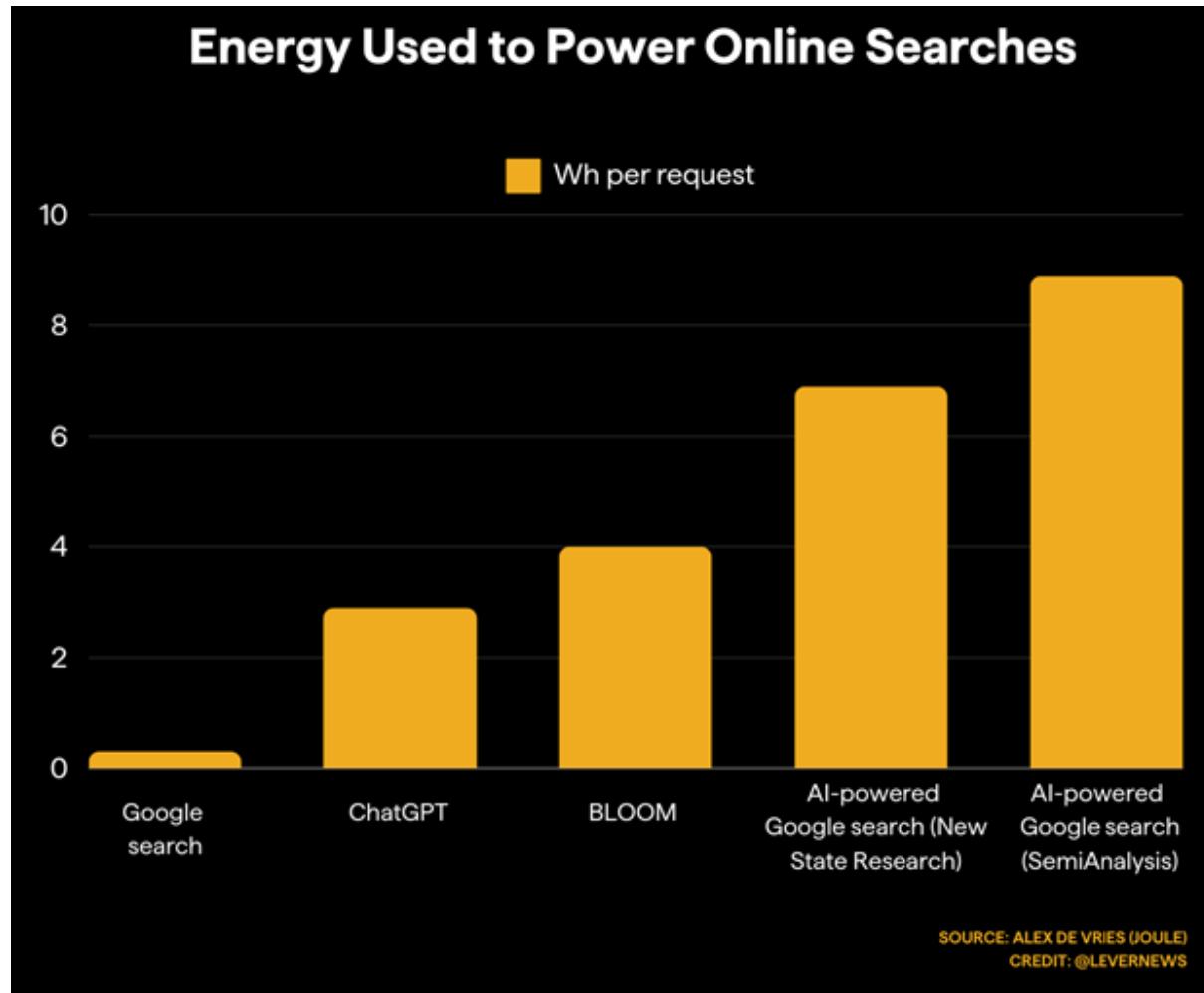
```
First Servingman:  
When we have done second it been,  
And the benefit of the people to the care  
And hate rescret her for her is true.
```

Training LLMs is only for the **Big Tech firms**



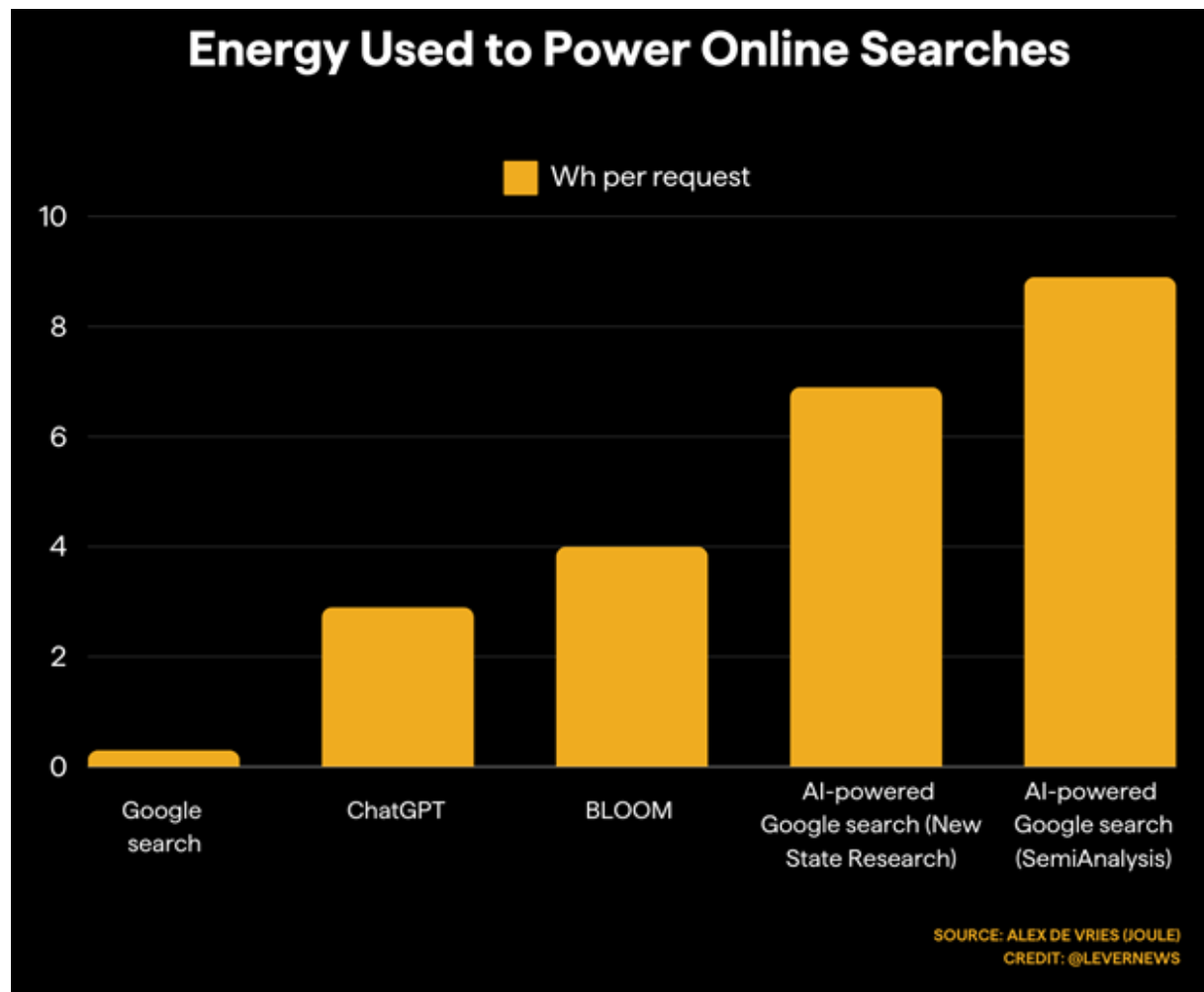


Environmental Impact





Environmental Impact



Datacenters worden gebruikt voor zowat alles wat wij doen op het web.
Foto: Getty

Google koopt kernenergie voor stroomvretende datacenters

Google gaat kleine kernreactoren gebruiken om massaal veel energie op te wekken die nodig is voor hun artificiële intelligentie-datacenters. Ze hebben daarvoor een contract getekend met Kairos Power, een onderneming die nieuwe reactortechnologie ontwikkelt. Google zegt dat de eerste reactor nog voor 2030 in gebruik zal worden genomen.

TECH & INNOVATIE

Kernenergie maakt een comeback dankzij big tech en energieslurpende AI

Kernenergie is hot. Grote Amerikaanse techbedrijven investeren miljarden in bestaande en nieuwe nucleaire opwekking van elektriciteit. Ze moeten de emissievrije stroomopwekking ook wel omarmen, want door de opmars van AI explodeert hun energieverbruik de komende jaren.

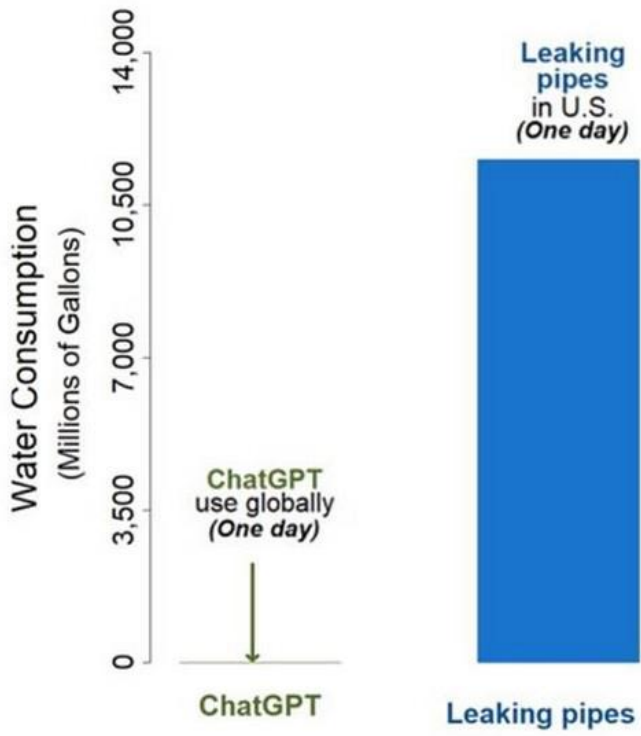
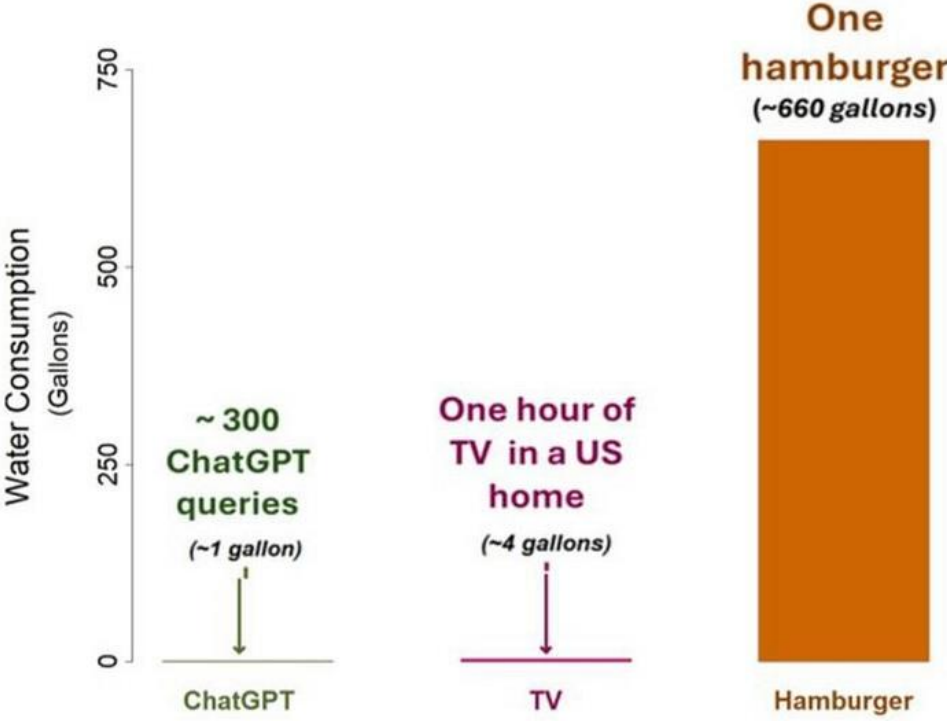
Nieuws Technologie

De gesloten kerncentrale Three Mile Island, waar een nucleair ongeval plaatsvond, gaat straks weer open. De enige klant: Microsoft



Environmental Impact

Water consumed using ChatGPT vs other activities

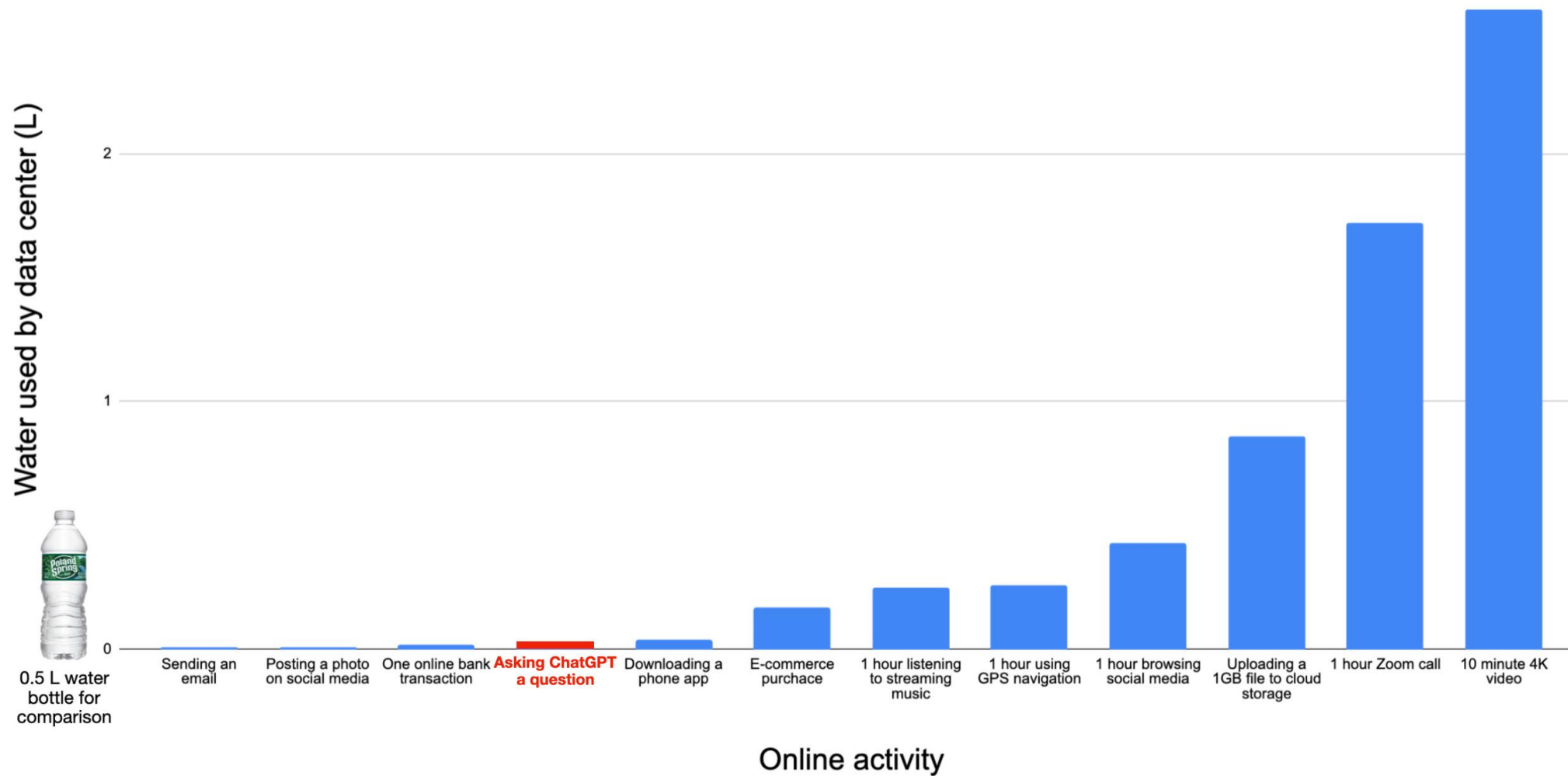


Sources: Li, Ren et al., 2023; U.S. Census Bureau; UNEP

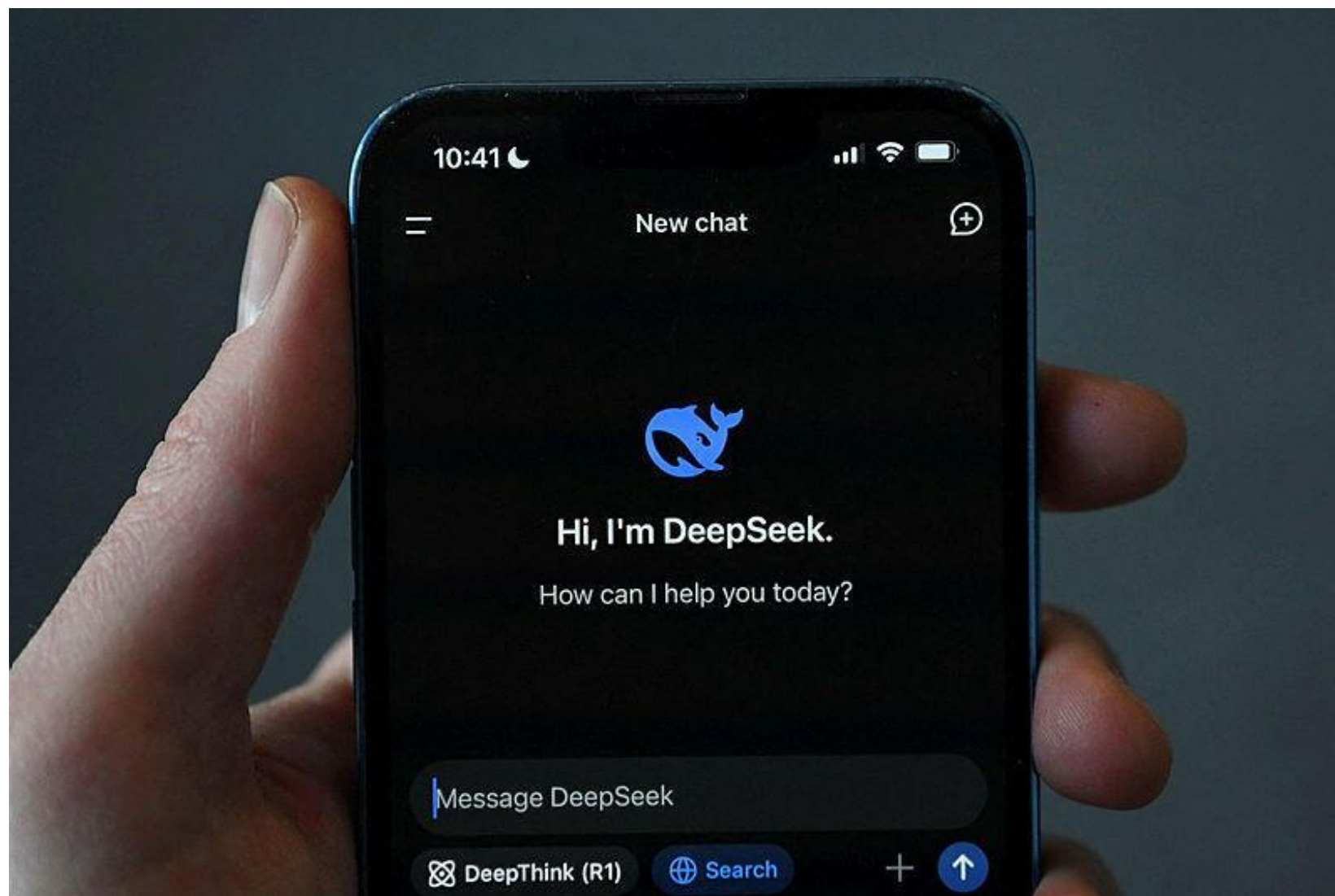
Liemberger & Wyatt, 2019 ; Liemberger & Wyatt, 2020



Environmental Impact



[AI is not bad for the environment]

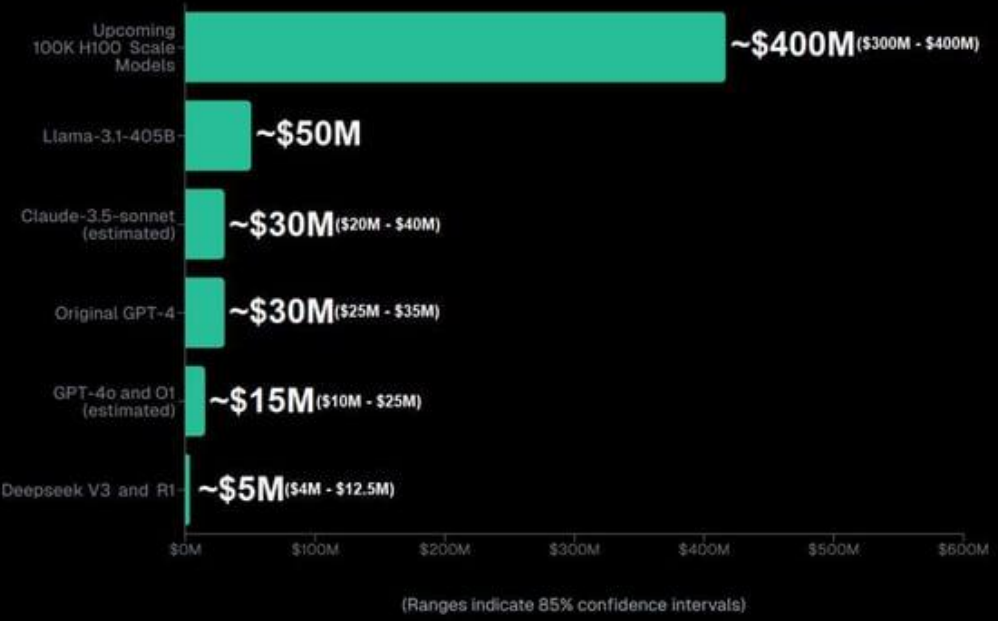




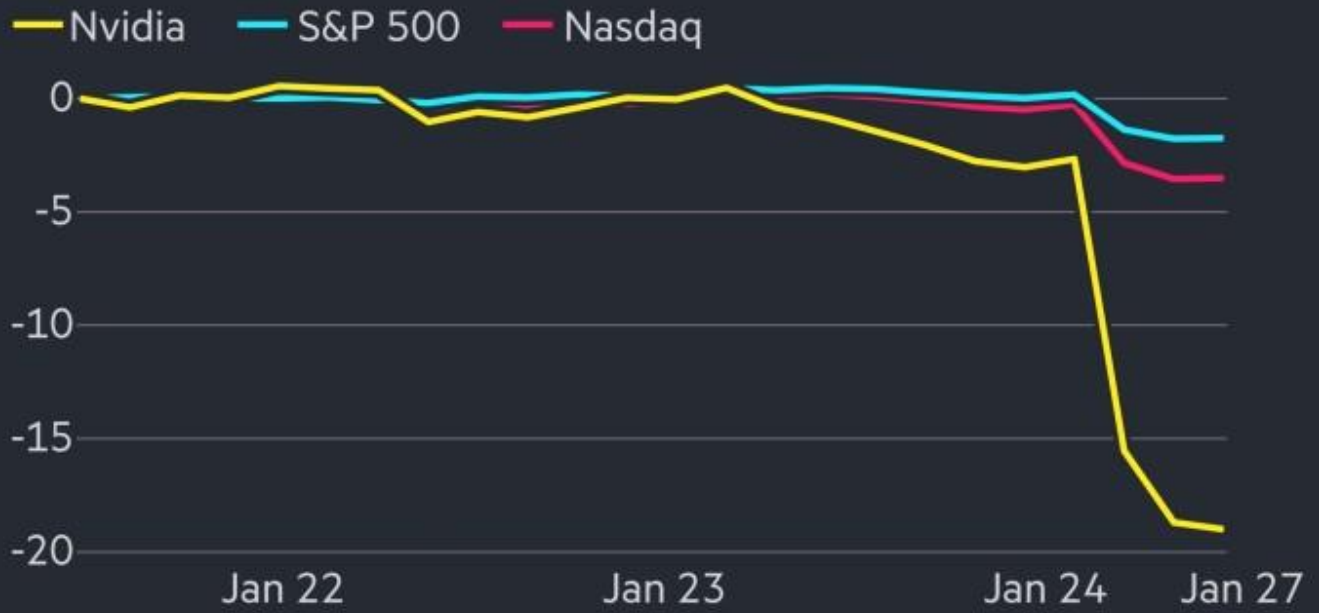
Training cost

AI Model Training Cost Comparison

Estimated cost of various models when using 2025 compute costs, in H100 hours of training compute



% change



Source: LSEG via markets.ft.com



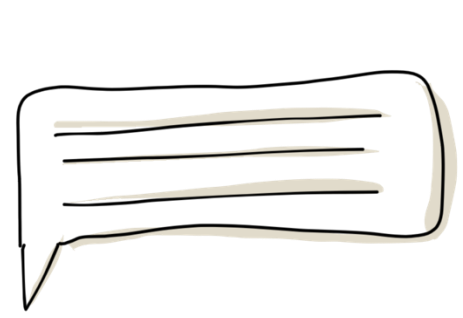
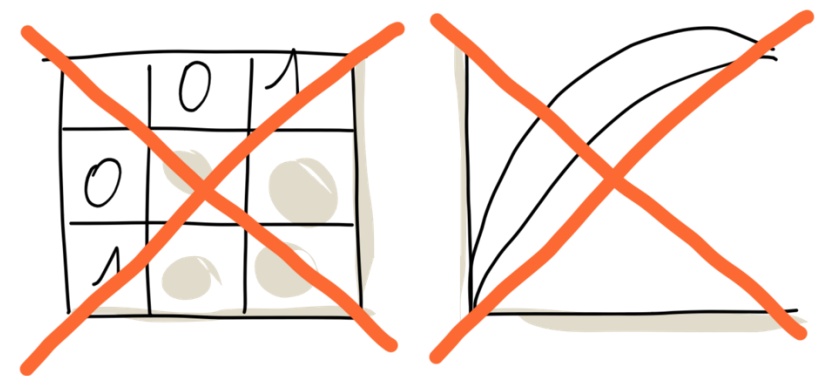
Feature comparison

Feature	DeepSeek	ChatGPT (GPT-4)
Compute Hardware	NVIDIA A100 GPUs, custom AI chips	NVIDIA A100/H100 GPUs (Azure)
Deployment Options	Supports local installation	Cloud-only (No local installation)
Edge Computing	Yes (Can run on-premise)	No, requires cloud access
Cloud Support	Hybrid Cloud & On-premise	Azure Cloud only
Energy Efficiency	20-30% better than GPT-4	Higher power consumption

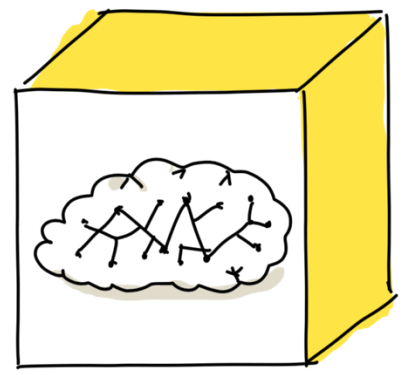


LLMS

- Introduction
- LLM Training
- **LLM Quality**
- LLM Adoption in Belgium
- Popular LLMs



PROMPT



LLM



OUTPUT

WHAT'S THE QUALITY?

LLM Failures

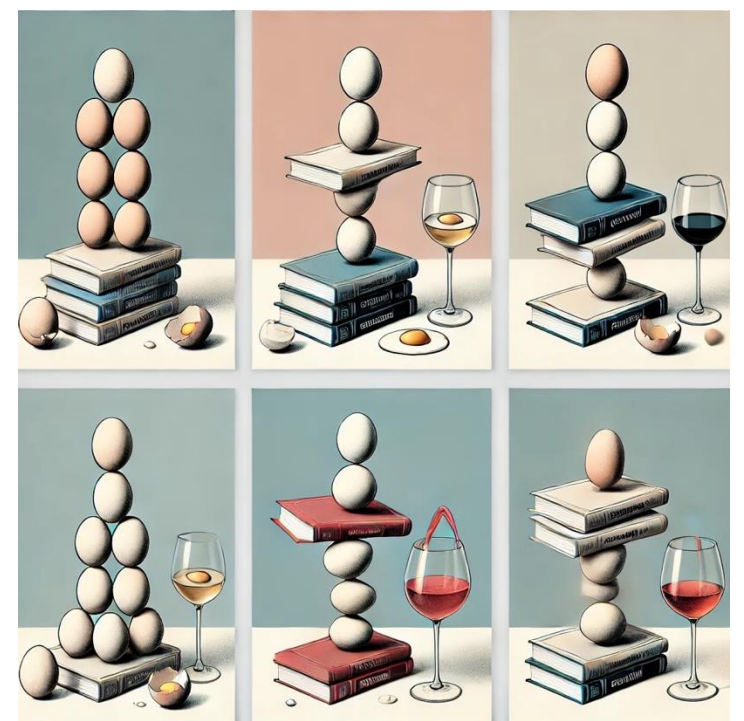
What was the average price of my health insurance over the last three years?

Could you provide the annual prices of your health insurance over the last three years? Then I can calculate the average for you.



Outdated / No Data

How would you construct the highest possible tower with only three eggs, a book and a glass of wine?



Hallucination

LLM Failures

What was the average price of my health insurance over the last three years?

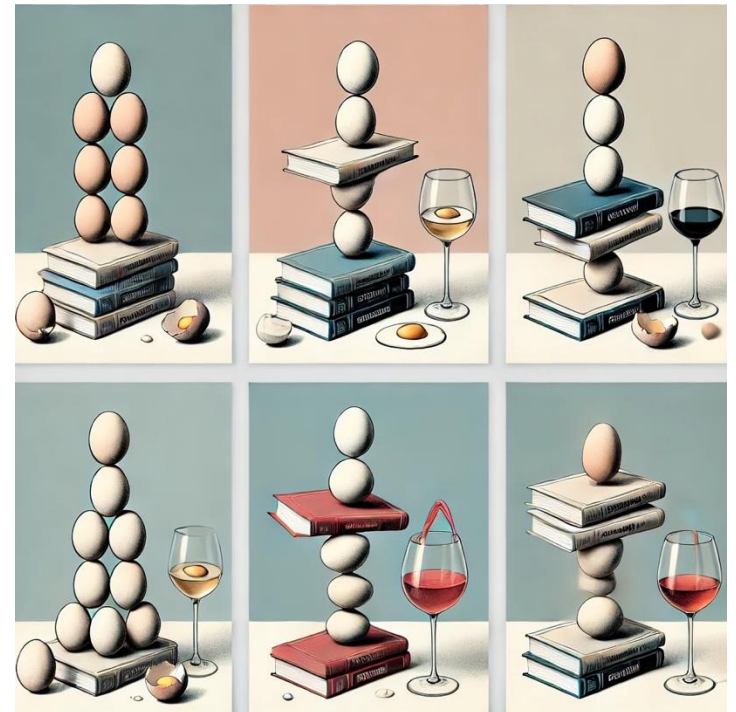


Could you provide the annual prices of your health insurance over the last three years? Then I can calculate the average for you.



Outdated / No Data

How would you construct the highest possible tower with only three eggs, a book and a glass of wine?



Hallucination



Can you detect dogs and muffins in this image?



- | Row 1: | Row 2: | Row 3: |
|----------------------|----------------------|----------------------|
| 1. Dog | 1. Dog | 1. Dog |
| 2. Dog | 2. Muffin | 2. Muffin |
| 3. Muffin | 3. Dog | 3. Dog |
| 4. Muffin | 4. Dog | 4. Dog |
| 5. Dog | 5. Dog | 5. Muffin |
| 6. Muffin | 6. Dog | 6. Muffin |
| 7. Dog | 7. Muffin | 7. Dog |
| 8. Muffin | 8. Muffin | 8. Muffin |

Hallucination



Hallucination



Fox eats Goat
Goat eats Cabbage



Lion eats Goat
Lion eats Cabbage
Goat eats Cabbage

Hallucination

A farmer has a goat, a cabbage and a lion. The farmer needs to cross the river but can only take one thing at a time. The goat cannot be left alone with the cabbage, the cabbage cannot be left alone with the lion and the lion cannot be left alone with the goat. How will he get all three across?



Qwen2.5-Plus

This is a classic river-crossing puzzle. Let's carefully reason through the problem step by step to ensure all constraints are satisfied.

Key Constraints:

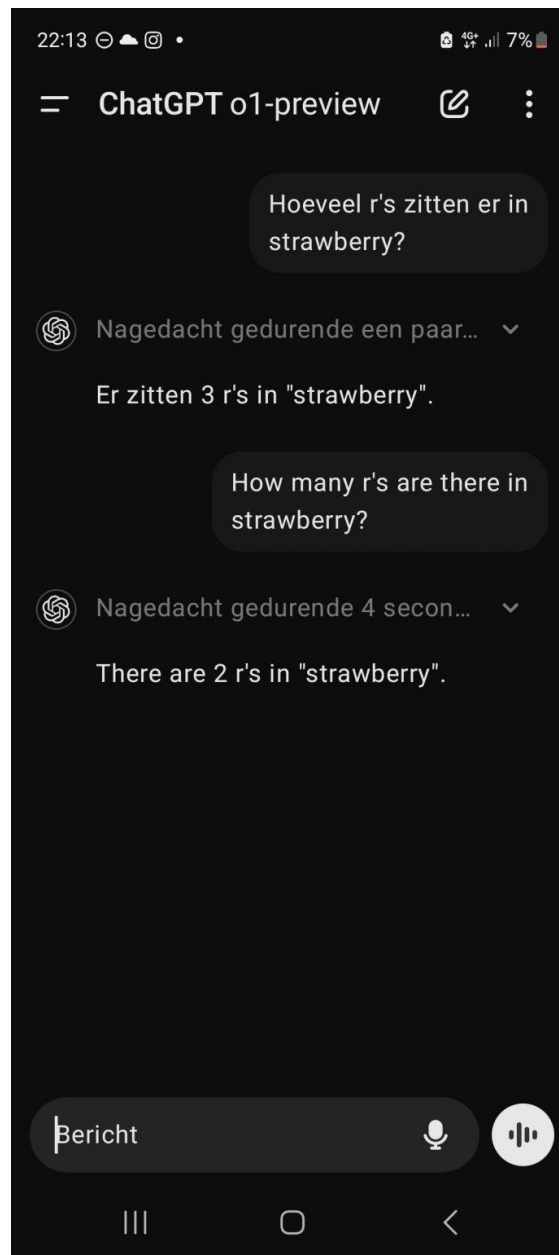
1. The **goat cannot be left alone with the cabbage** , because the goat will eat the cabbage.
2. The **cabbage cannot be left alone with the lion** , because the lion will eat the cabbage.
3. The **lion cannot be left alone with the goat** , because the lion will eat the goat.
4. The farmer can only take one item (goat, cabbage, or lion) across the river at a time.

Solution:

We need to transport all three (goat, cabbage, and lion) across the river without violating any of the constraints.

Step-by-Step Plan:

1. **Take the goat across the river first.**
 - Now the goat is on the far side, and the cabbage and lion remain on the starting side.
2. **Go back alone to the starting side.**
 - The goat is safe on the far side for now.
3. **Take the cabbage across the river next.**



Hallucination

VRAAG : Ben je al hallucinaties tegengekomen?

Deel en leer 😊

LLM Failures

What was the average price of my health insurance over the last three years?

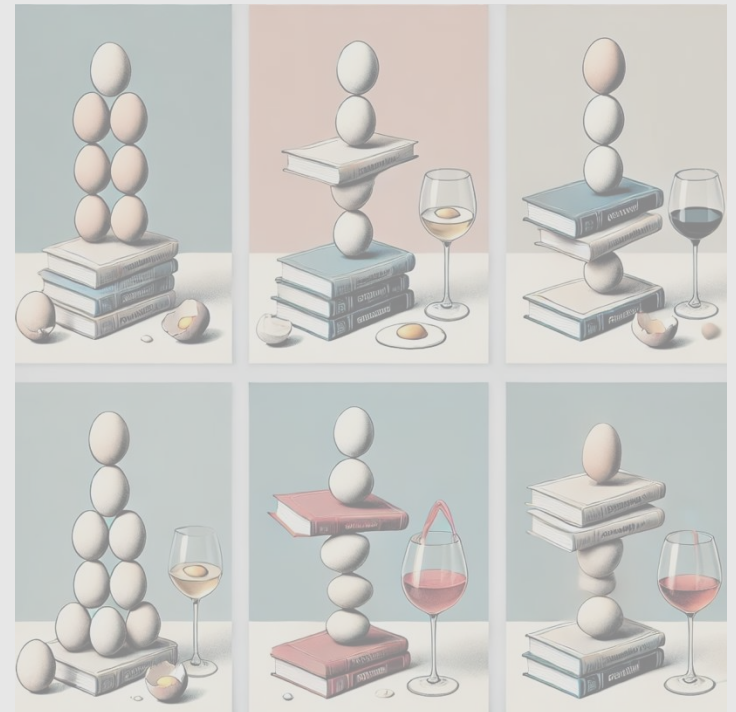


Could you provide the annual prices of your health insurance over the last three years? Then I can calculate the average for you.



Outdated / No Data

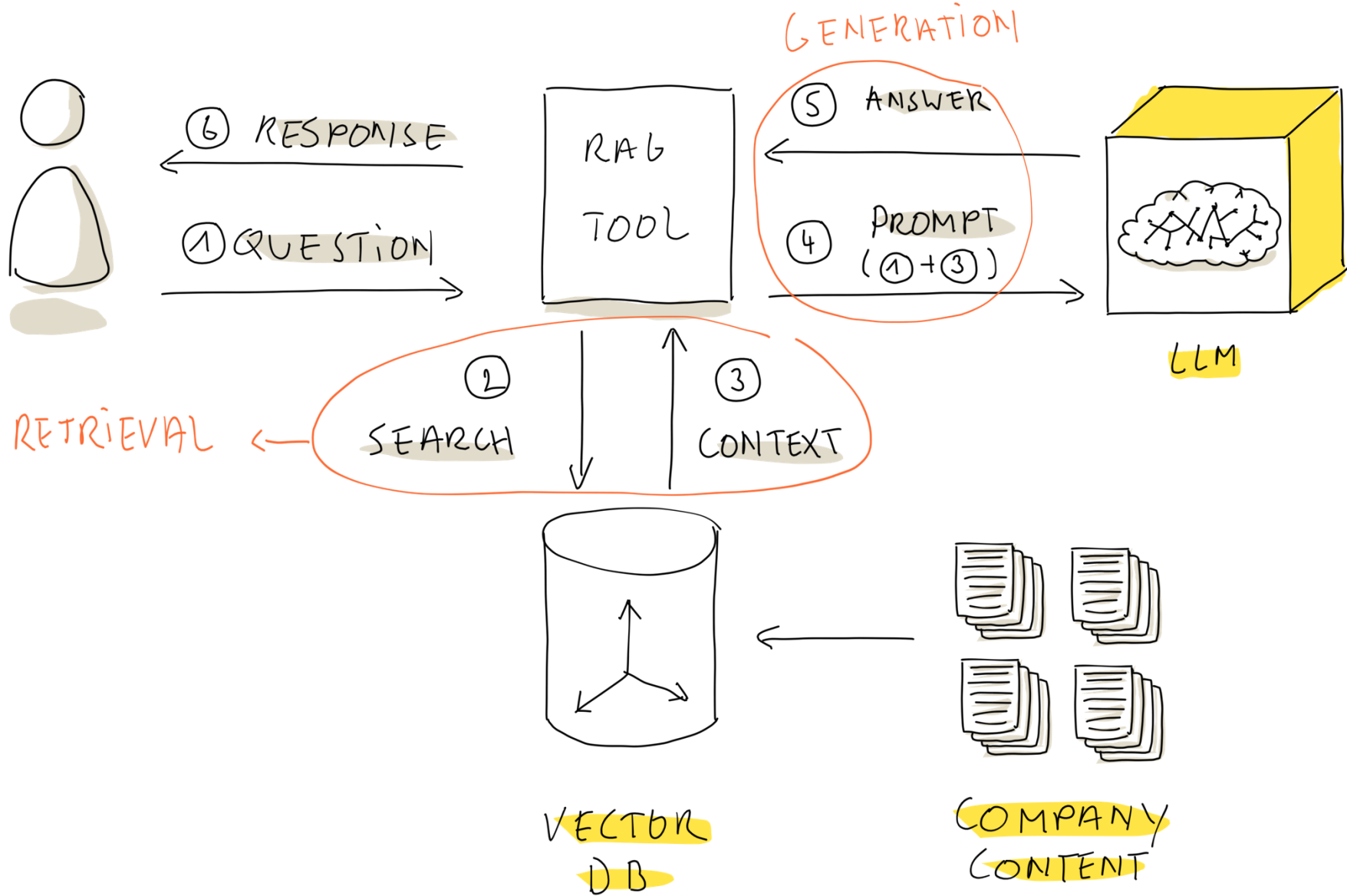
How would you construct the highest possible tower with only three eggs, a book and a glass of wine?



Hallucination

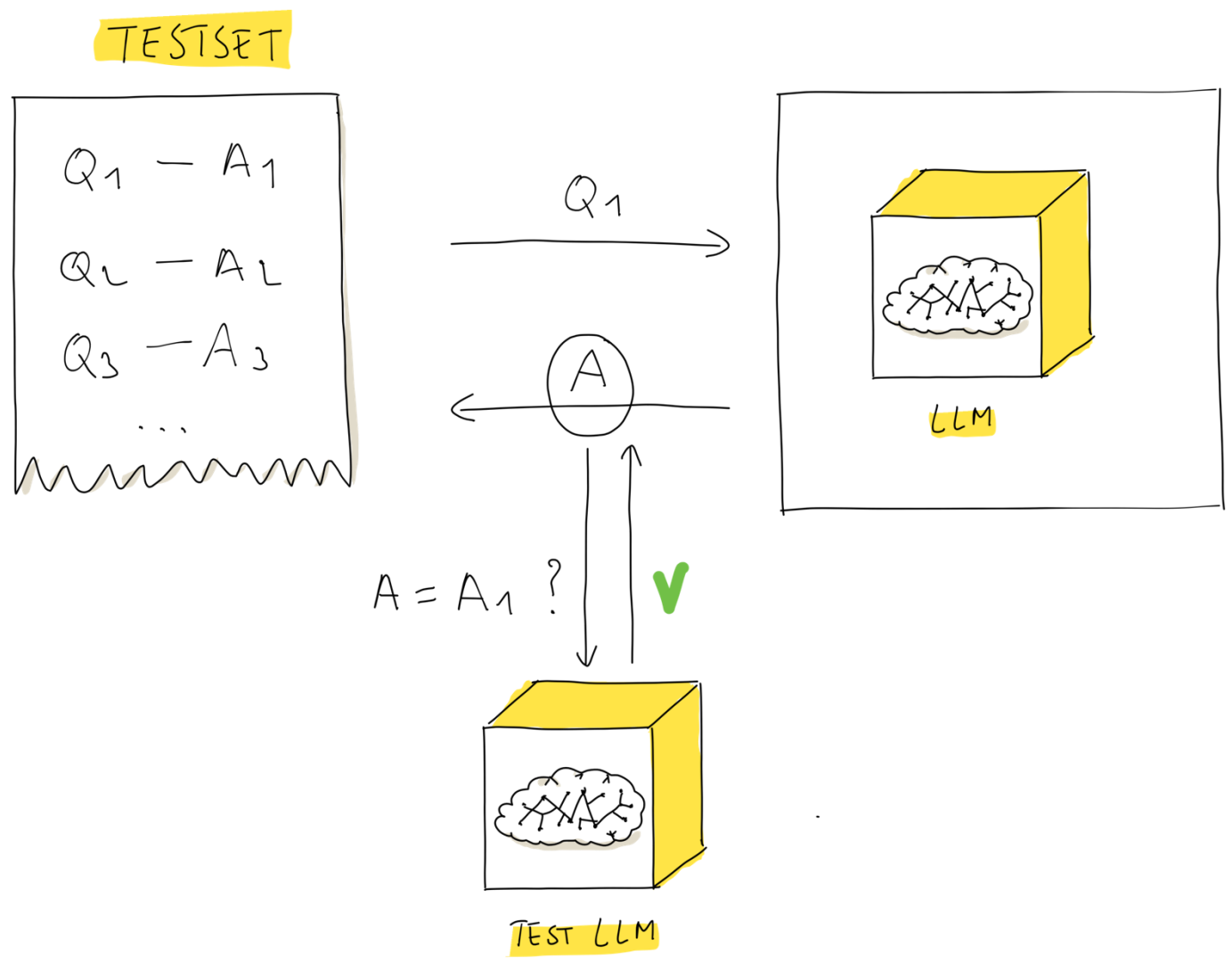


Retrieval Augmented Generation (RAG)





Testset-Based LLM Evaluation





More advanced : RAGAS

generation

faithfulness

how factually accurate is the generated answer

answer relevancy

how relevant is the generated answer to the question

retrieval

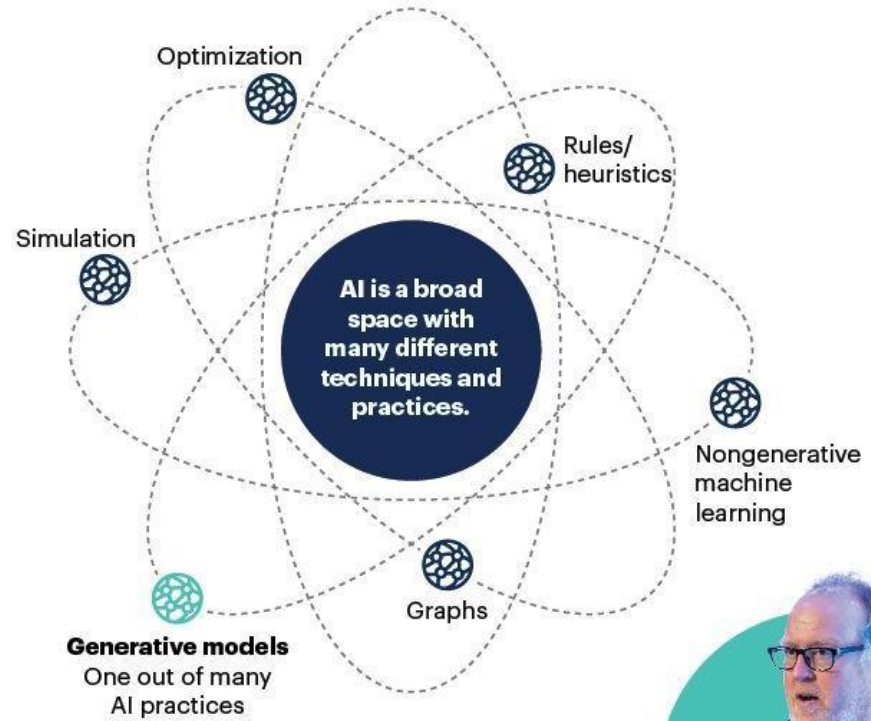
context precision

the signal to noise ratio of retrieved context

context recall

can it retrieve all the relevant information required to answer the question

	question	contexts	answer	ground_truths	context_precision	context_recall
0	What is the approach used in BLIP-2 for vision...	[BLIP-2: Bootstrapping Language-Image Pre-trai...	BLIP-2 uses a two-stage approach for vision-la...	[The approach used in BLIP-2 for vision-langua...	0.222222	1.0
1	How do frozen image encoders and large languag...	[BLIP-2: Bootstrapping Language-Image Pre-trai...	Frozen image encoders and large language model...	[Frozen image encoders and large language mode...	0.333333	1.0
2	What experimental design is used to study scal...	[On these tasks, we find that human participan...	The experimental design used to study scalable...	[The experimental design used to study scalabl...	0.250000	1.0
3	What are the two aspects of language use that ...	[Understanding the Capabilities, Limitations, ...	The two aspects of language use that are consi...	[The two aspects of language use that are cons...	0.000000	0.0
4	How does combining pretrained language models'...	[Large Language Models Can Self-Improve Large ...	Combining pretrained language models' in-conte...	[Combining pretrained language models' in-cont...	0.000000	1.0



“Generative AI, which is super dominant as the leading trend here, is being misapplied. There is an idea that generative AI can do things that classical AI is actually better at doing. This dissonance happens, and you ask, ‘Well, why aren’t I getting the results that I expected?’ **It’s probably because you’re applying GenAI in the wrong way.**”

Chris Howard
Gartner Global Chief of Research



LLMS

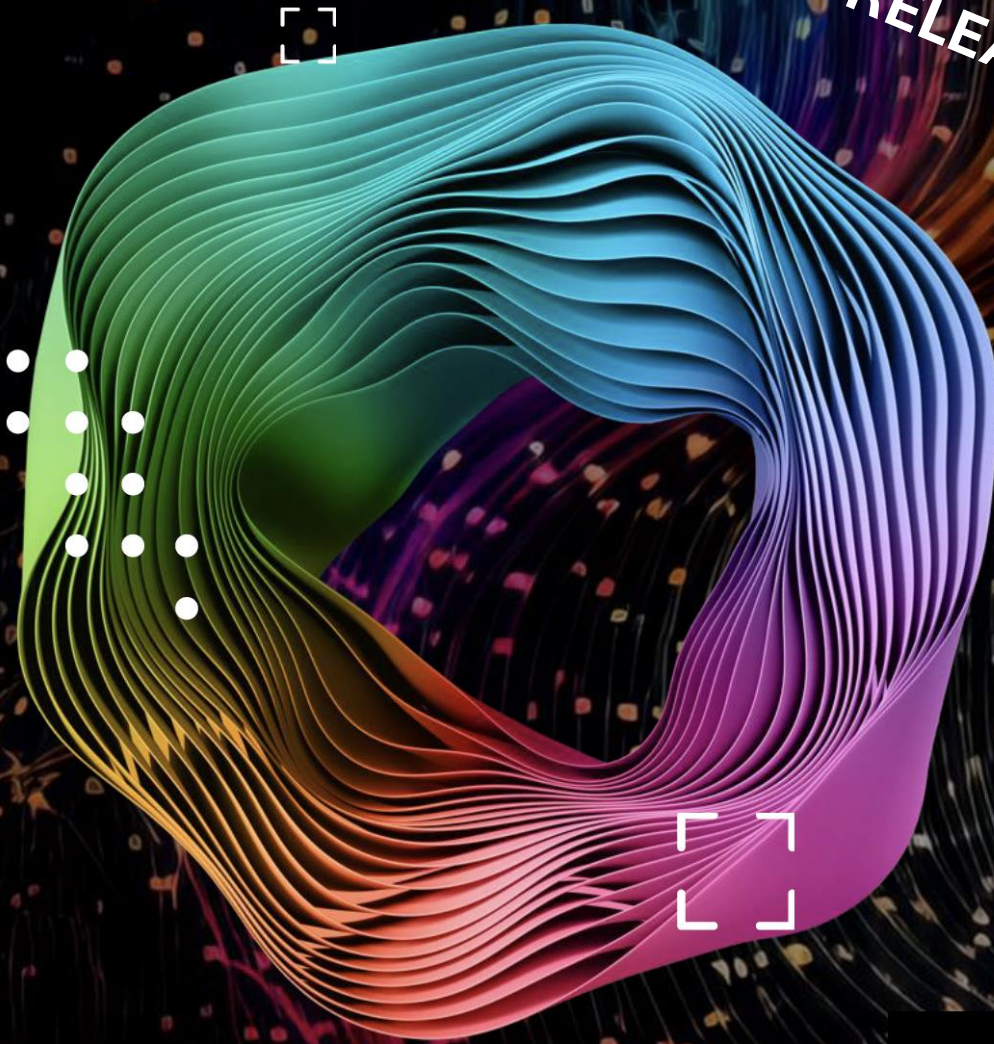
- Introduction
- LLM Training
- LLM Quality
- **LLM Adoption in Belgium**
- Popular LLMs

Deloitte.

**Trust in
Generative AI**
A Belgian Perspective

October 2024

JUST RELEASED

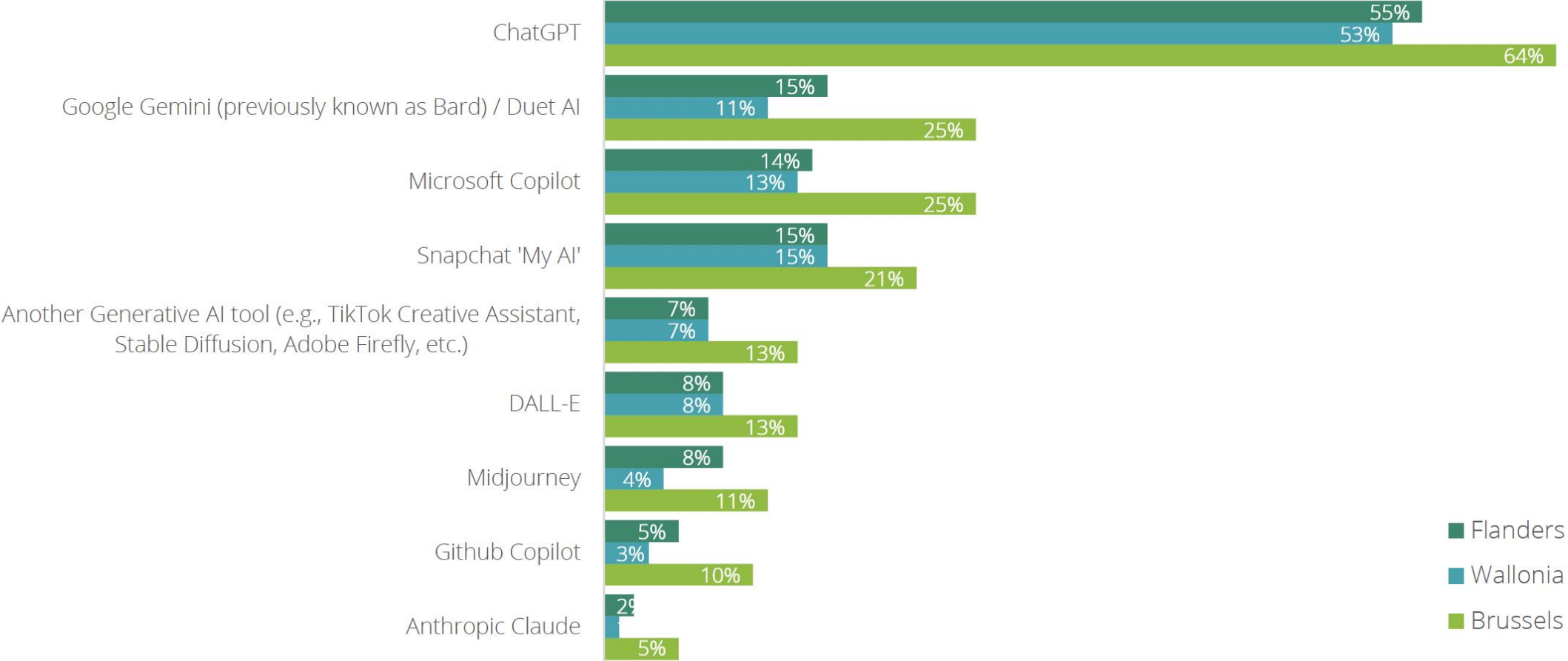




Awareness of Generative AI tools per region in Belgium

Among respondents aware of Gen AI tools, more than 5 out of 10 report being familiar with Chat GPT

Base: All adults aged 16-75 in Belgium (N = 2714)
Q1. Which, if any, of the following generative AI tools are you aware of?



Flanders
Wallonia
Brussels

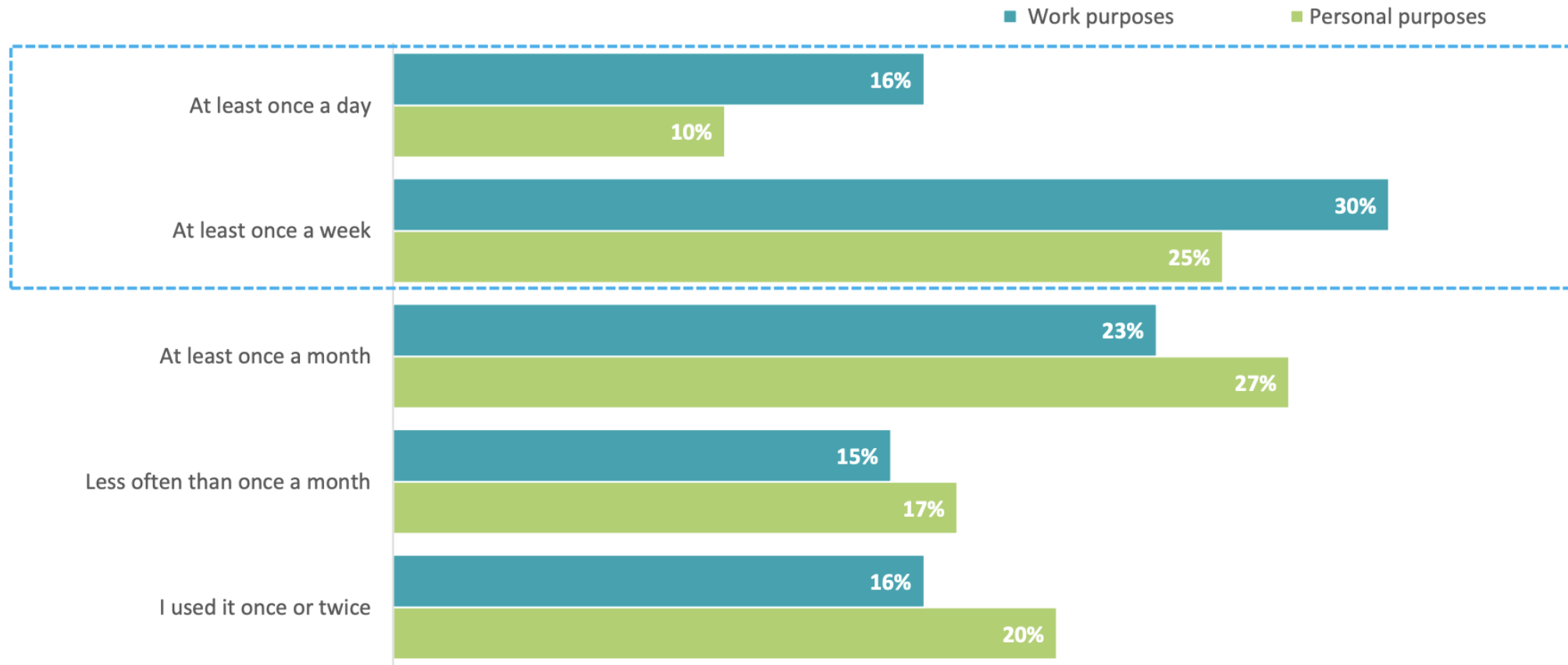


Frequency of Generative AI use in Belgium

Of those who use Gen AI, more than one-third of Belgians use Gen AI at least once a week for both personal (36%) and work purposes (46%)

Base: Aware and use Gen AI (N=912 - Personal purposes | N=884 - Work purposes)

Q4. You mentioned that you have used Generative AI tools. Which of the following describes how often you typically use it for...?



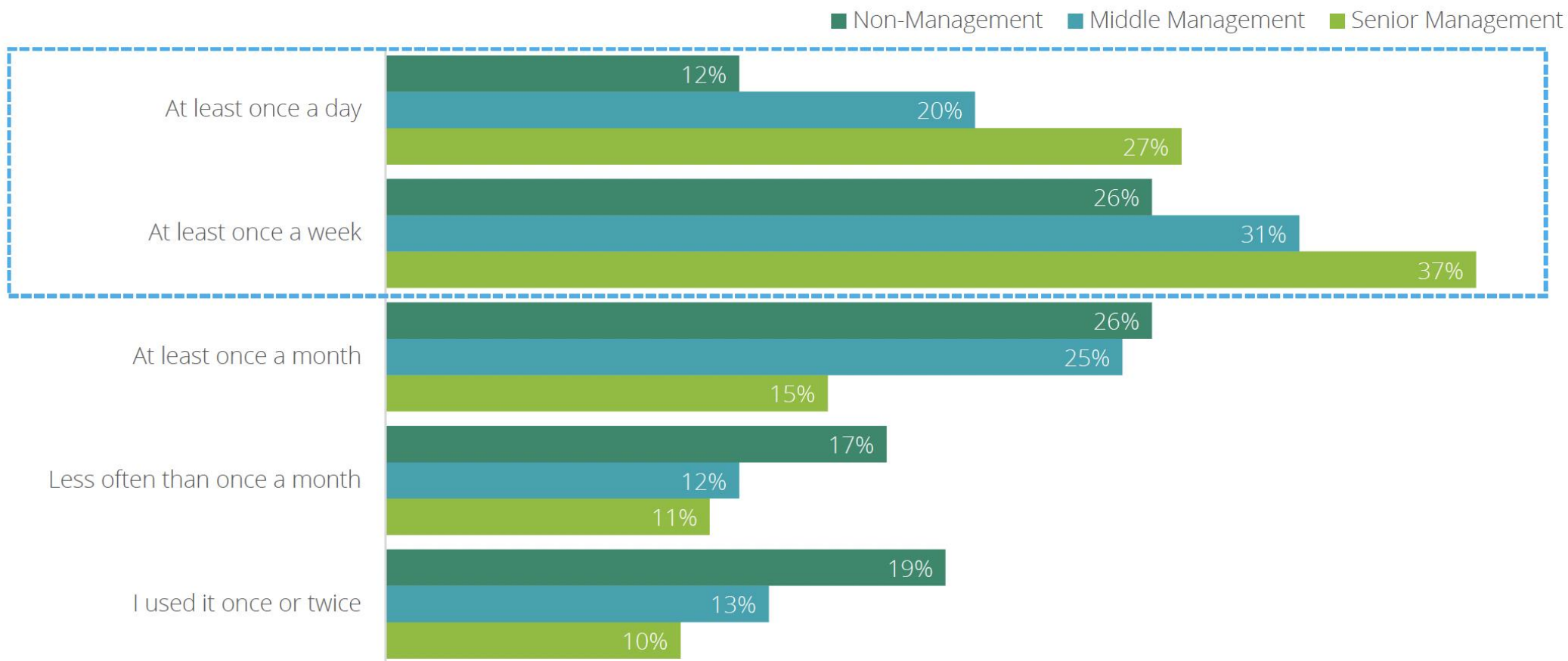


Frequency of Generative AI use per job level

Of Belgians who use Gen AI for work purposes, it appears that those in Senior Management use Gen AI considerably more often (64% at least weekly) compared to Middle Management (51% at least weekly) and Non-Management levels (39% at least weekly)

Base: All who have used a Generative AI tool for work purposes (N=884)

Q4. You mentioned that you have used Generative AI tools. Which of the following describes how often you typically use it for...work purposes?

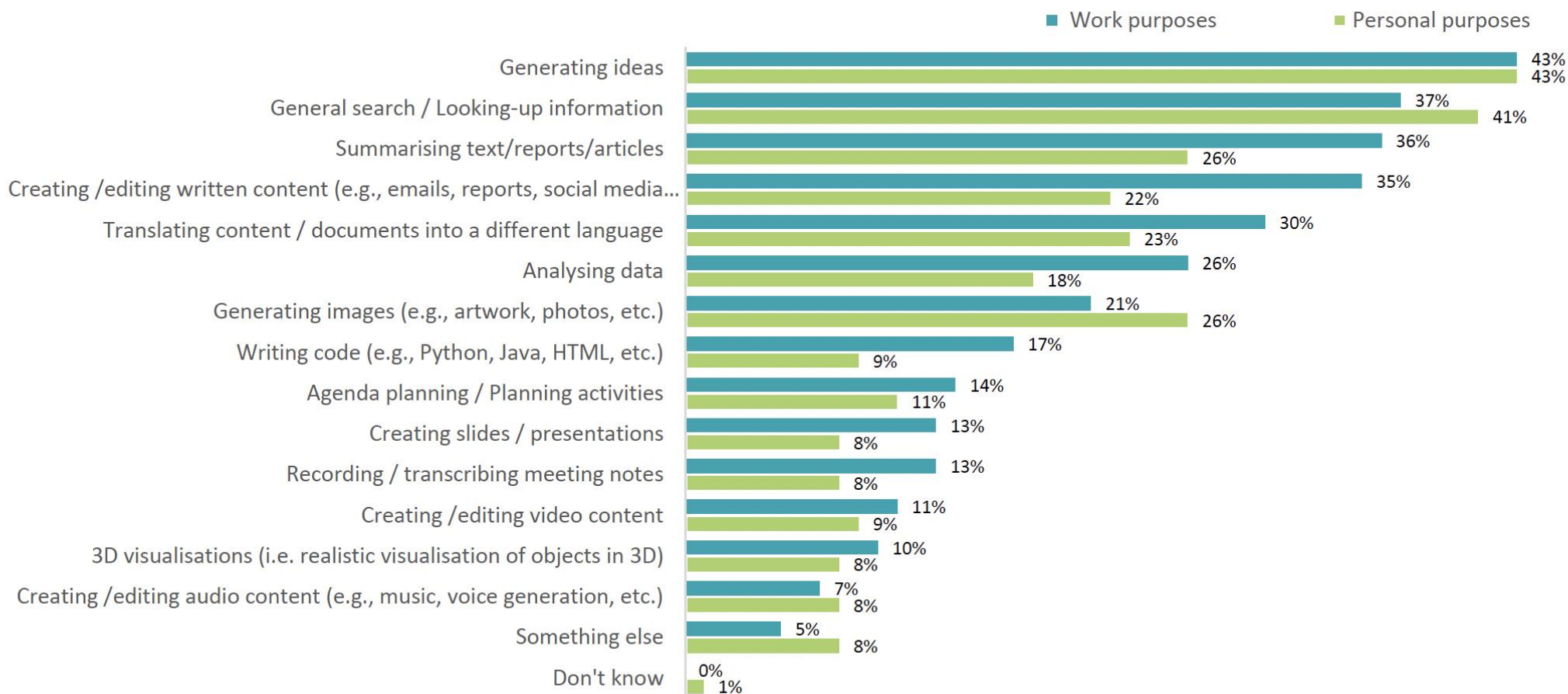




Use cases of Generative AI

Generating ideas is the primary use of GenAI for both work & personal activities (43%), while general searches or looking up information dominate personal use (41%)

Base: Aware and use Gen AI (N=912 - Personal purposes | N=884 - Work purposes)
Q5. For which of the following tasks do you typically use Generative AI tool(s)?



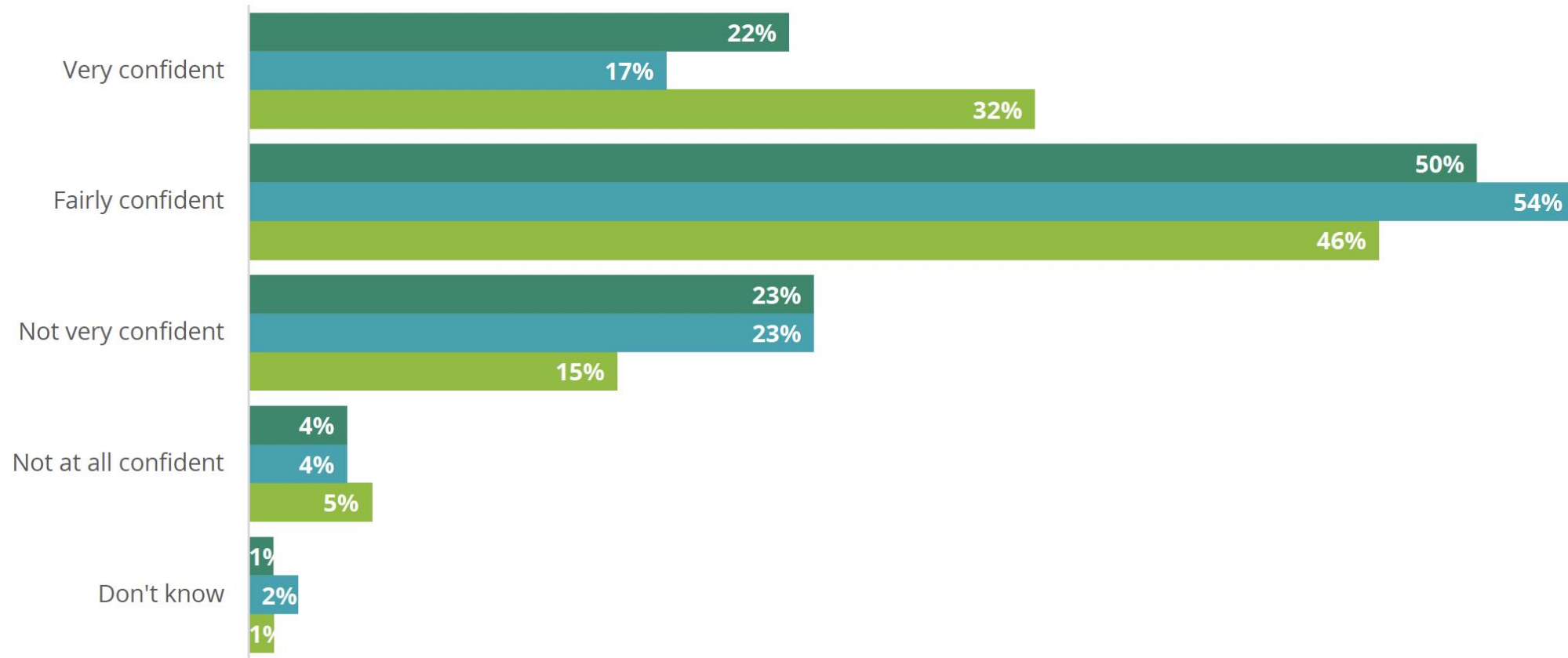


Confidence in using Generative AI

3 out of 4 Belgian respondents report being very or fairly confident in using Gen AI tools. Research highlighted though a significant confidence from men (78%), compared to women (63%)

Base: All who have used a Generative AI tool (N = 1406)

Q6. On balance, how confident, or not, are you in using Generative AI tool(s)?



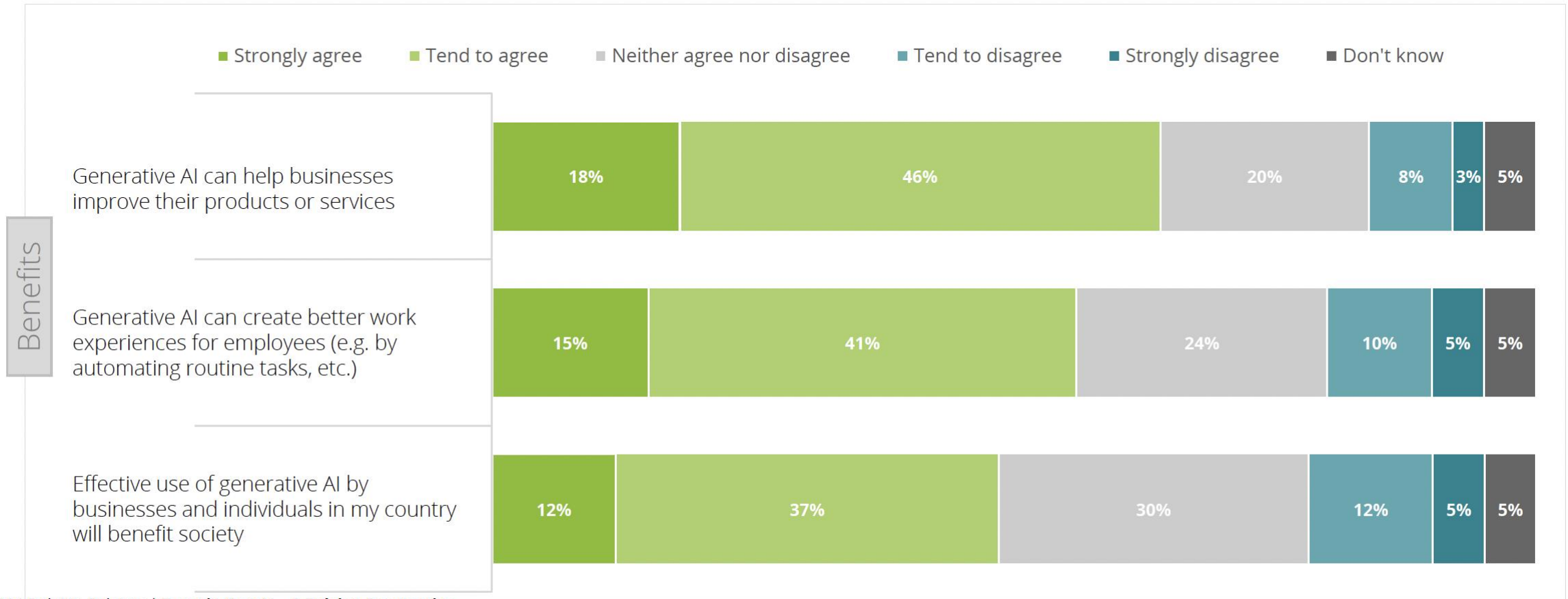


Potential & Trust in Generative AI

64% of Belgians aware of generative AI tools, say that generative AI can help businesses improve their products or services, while 56% say it can create better working experiences for employees. Population from both Brussels, Flanders & Wallonia equally agree with the above statements.

Base: Aware and use Gen AI (N = 1970)

Q7. To what extent do you agree, or disagree, with each of the following statements?



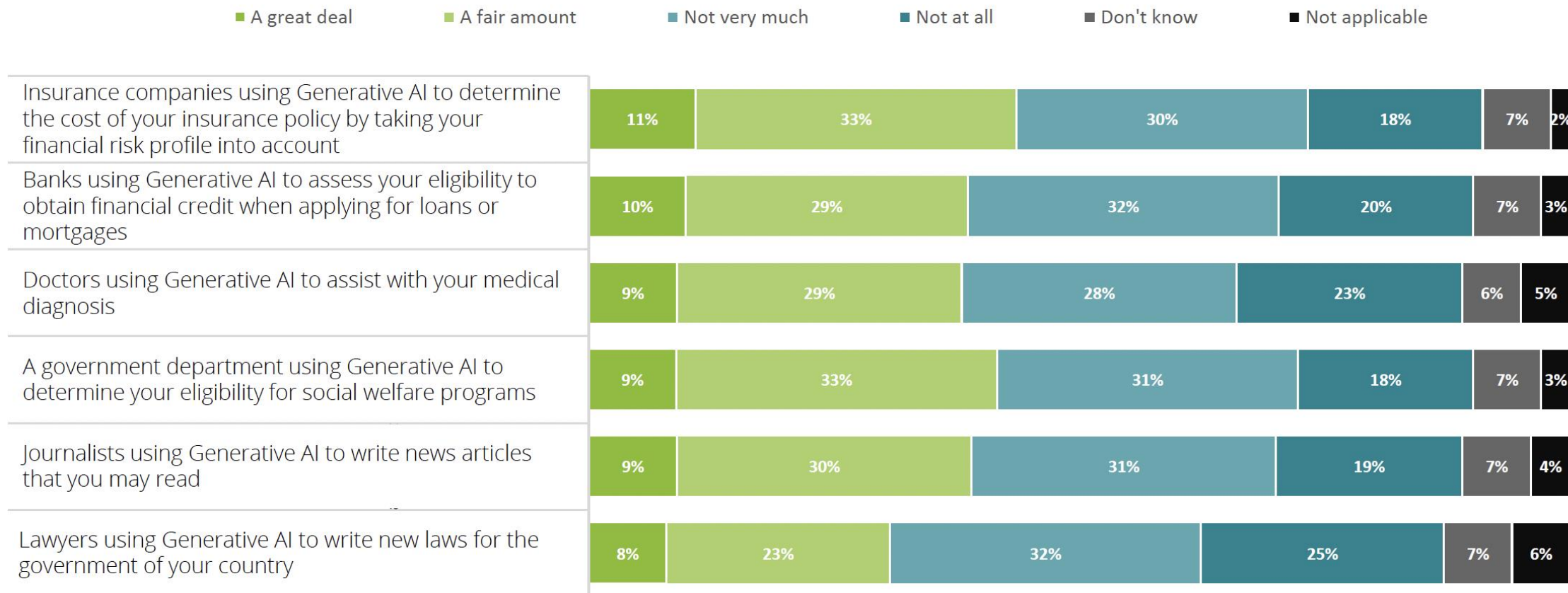


Trust in Generative AI for specific business use scenarios

Generative AI Belgian users trust its results for business (*higher-risk*) use cases to a lower extent than personal use cases

Base: Aware and use Gen AI (N = 1970)

Q10b. Now thinking about how businesses and organisations could potentially use Generative AI. To what extent, if at all, would you personally trust the results produced by Generative AI in each of the following scenarios?



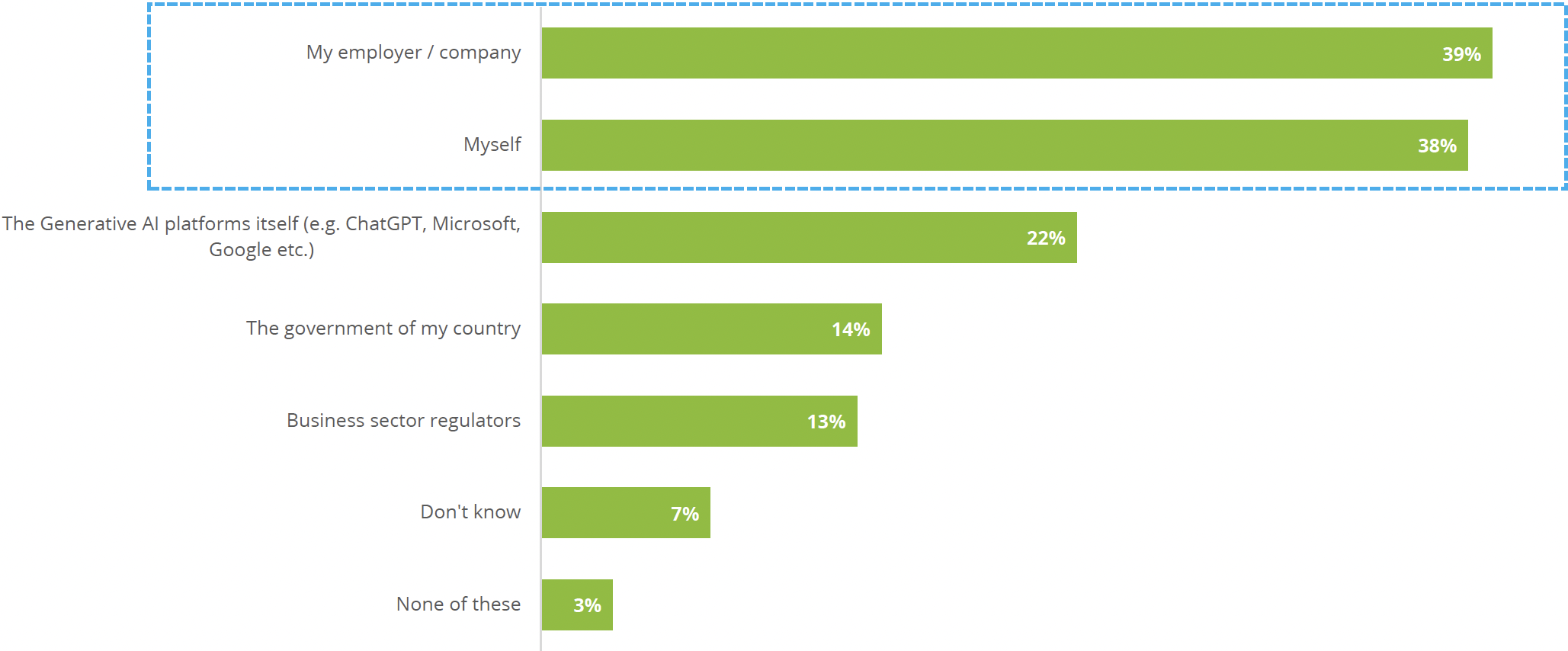


Responsibility for developing Generative AI skills

Above one-third of Gen AI users in Belgium view upskilling as a shared duty, with 39% seeing it as a personal responsibility and 38% as the employer's responsibility. Significant amount of Belgians also shares the view that it is up to the Gen AI platforms itself.

Base: Aware and use Gen AI (N = 791)

Q22. Who, if anyone, do you think should be responsible for developing your skills to use Generative AI tools for work purposes in the following scenarios?



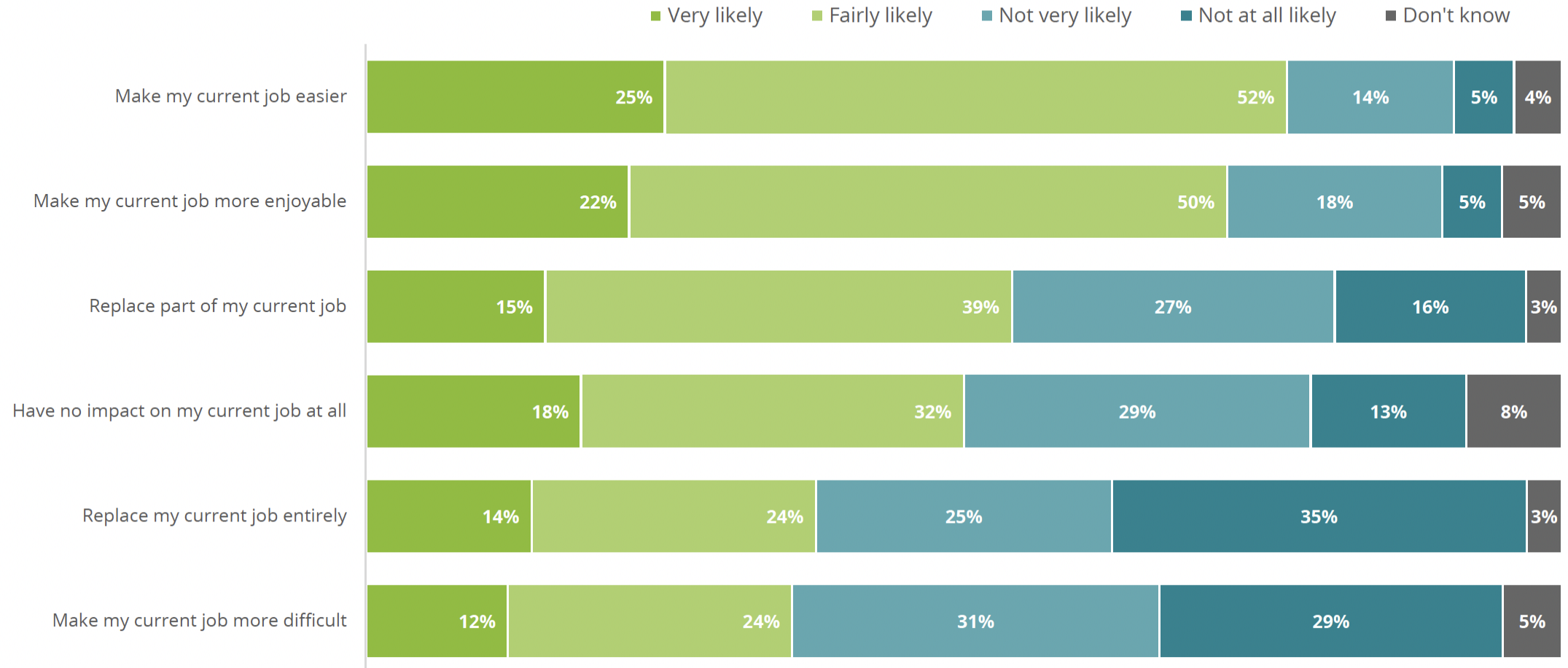


Impact of Generative AI in next 2 years

Most Gen AI users believe that in the next two years, Gen AI will make their jobs easier (77%) and more enjoyable (72%), and may automate some tasks (53%), but a majority do not believe it will replace their jobs entirely (60%) or increase job difficulty (60%)

Base: Aware and use Gen AI (N = 791)

Q21. In the next 2 years, how likely, or not, do you think it is that Generative AI will...?

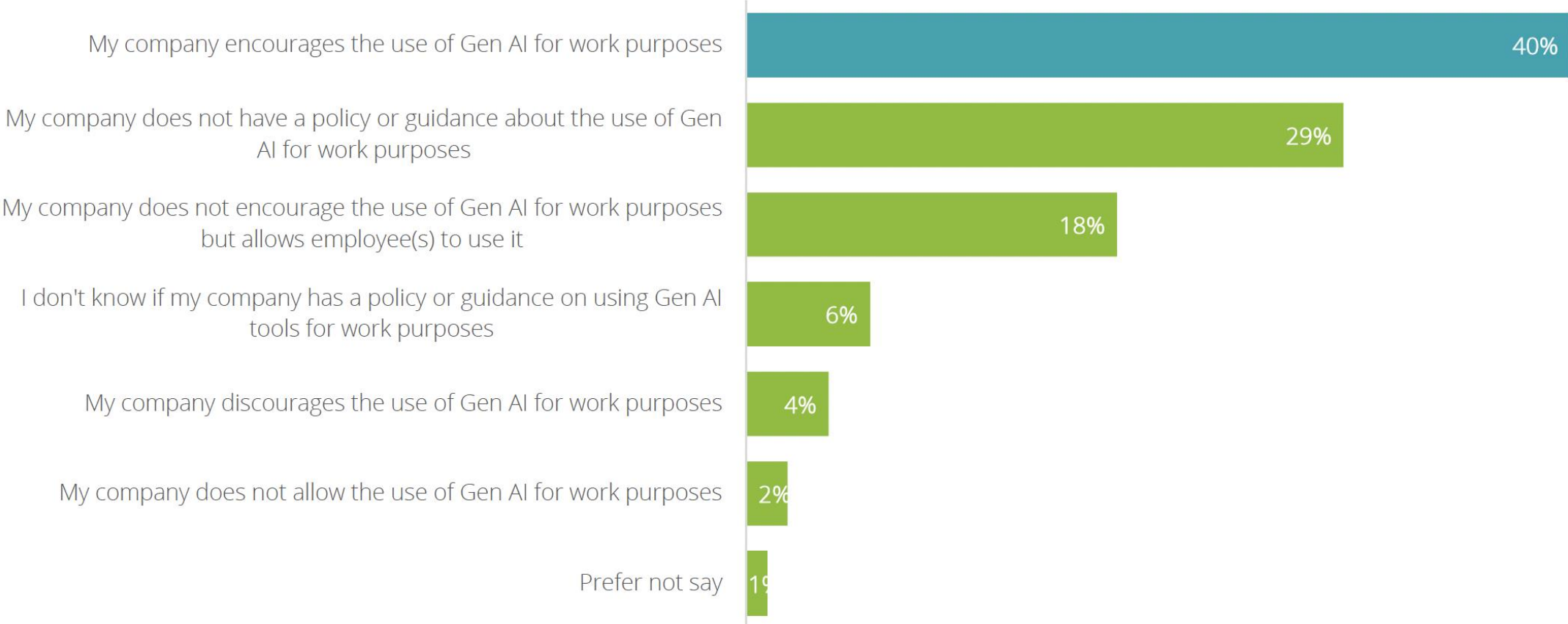




Organisational policy around Generative AI

40% of Belgian employees report that their company actively encourages the use of generative AI for work purposes

Base: All employed who have used a Generative AI tool for work purposes (N = 791)
Q16. Thinking about the use of Gen AI tools for work purposes, which, if any, of the following statements applies to your company about the use of Generative AI tools for work purposes?





LLMS

- Introduction
- LLM Training
- LLM Quality
- LLM Adoption in Belgium
- **Popular LLMs**



ArtificialIntelligence.AI

artificialanalysis.ai

Follow us on Twitter or LinkedIn to stay up to date with future analysis

Artificial Analysis LANGUAGE MODELS SPEECH, IMAGE & VIDEO MODELS LEADERBOARDS ARENAS ABOUT

Newsletter [Subscribe](#)

Artificial Analysis State of AI: China Q1 2025

We chart the rise of China's top AI companies, map the the Chinese AI ecosystem and and compare to leading US models.

[View report](#)

Independent analysis of AI models and API providers

Understand the AI landscape to choose the best model and provider for your use-case

Highlights

QUALITY

Artificial Analysis Quality Index; Higher is better

Model	Quality Index
o1	90
e2z-min	89
DeepSeek R1	89
Gemini 2.0 Pro Experimental	85
o1-mini	84
Gemini 2.0 Flash	83
Claude 3.5 Sonnet (CC)	80
Nova Pro	75
GPT-4o (Nov 24)	75
Llama 3.3 70B	74
Mistral Large 2 (Nov 24)	74
GPT-4o mini	73
Claude 3.5 Haiku	68

SPEED

Output Tokens per Second; Higher is better

Model	Output Tokens per Second
e2z-min	190
o1-mini	178
Gemini 2.0 Flash	159
Gemini 2.0 Pro Experimental	130
GPT-4o mini	113
GPT-4o (Nov 24)	98
Nova Pro	90
Llama 3.3 70B	84
Claude 3.5 Sonnet (CC)	70
Claude 3.5 Haiku	65
o1	64
Mistral Large 2 (Nov 24)	38
DeepSeek R1	26

PRICE

USD per 1M Tokens; Lower is better

Model	USD per 1M Tokens
Gemini 2.0 Flash	0.2
GPT-4o mini	0.3
Llama 3.3 70B	0.6
Nova Pro	1.4
Claude 3.5 Haiku	1.6
o1-mini	1.9
Mistral Large 2 (Nov 24)	3
DeepSeek R1	3
o1-mini	3.9
GPT-4o (Nov 24)	4.4
Claude 3.5 Sonnet (CC)	6
o1	26.3

How do DeepSeek models compare? [DeepSeek Compared](#)

Where can you get an API for DeepSeek R1? [DeepSeek R1 Providers](#)

Which models perform best in different languages? [Multilingual Comparison](#)

Who has the best Video Generation model? [Video Arena](#)

Which model is fastest with 100k token prompts? [Long Context Latency](#)

Language Model Comparison Highlights [Comprehensive Model Comparison](#)



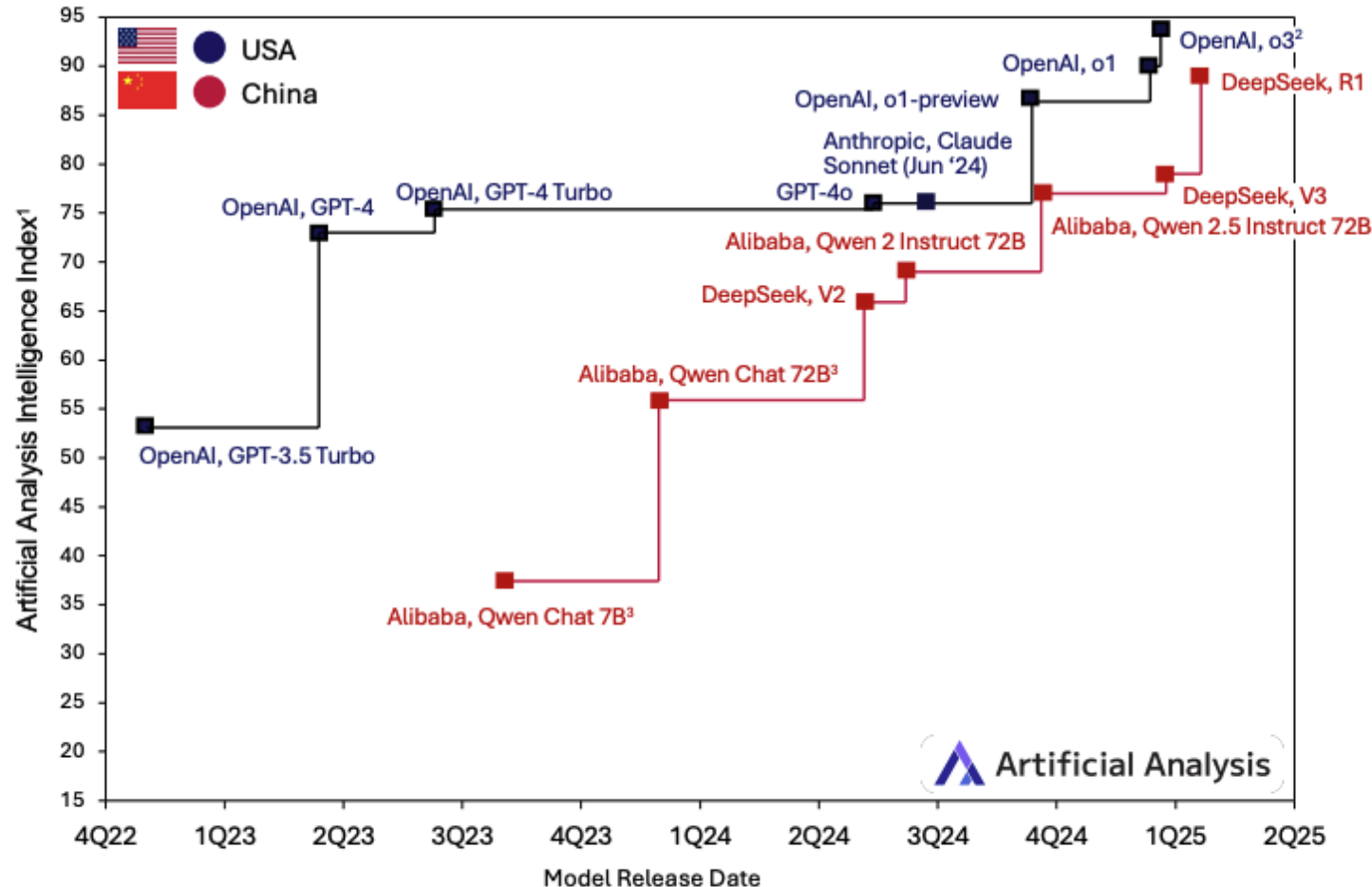
State of AI: China

Artificial Analysis
Q1 2025



Chinese AI labs have progressively caught up to US AI labs; models from Chinese labs are now approaching o1-level intelligence with the release of DeepSeek's R1 model

US & China: Frontier Language Model Intelligence, Over Time¹



Key Trends

Closing the gap: The final months of 2024 have seen the emergence of the numerous highly performant models from top Chinese AI labs. This has resulted in the delta between the level of intelligence offered by models from Chinese AI labs and US AI labs closing. Several Chinese models are now competitive with models from the top US labs.

Reasoning models quickly becoming commonplace: Reasoning models (that “think” before answering) were first introduced by OpenAI in 3Q24. Within months, Chinese competitors, led by DeepSeek, have largely replicated the intelligence of o1. Several AI labs in China now have a frontier-level reasoning model.

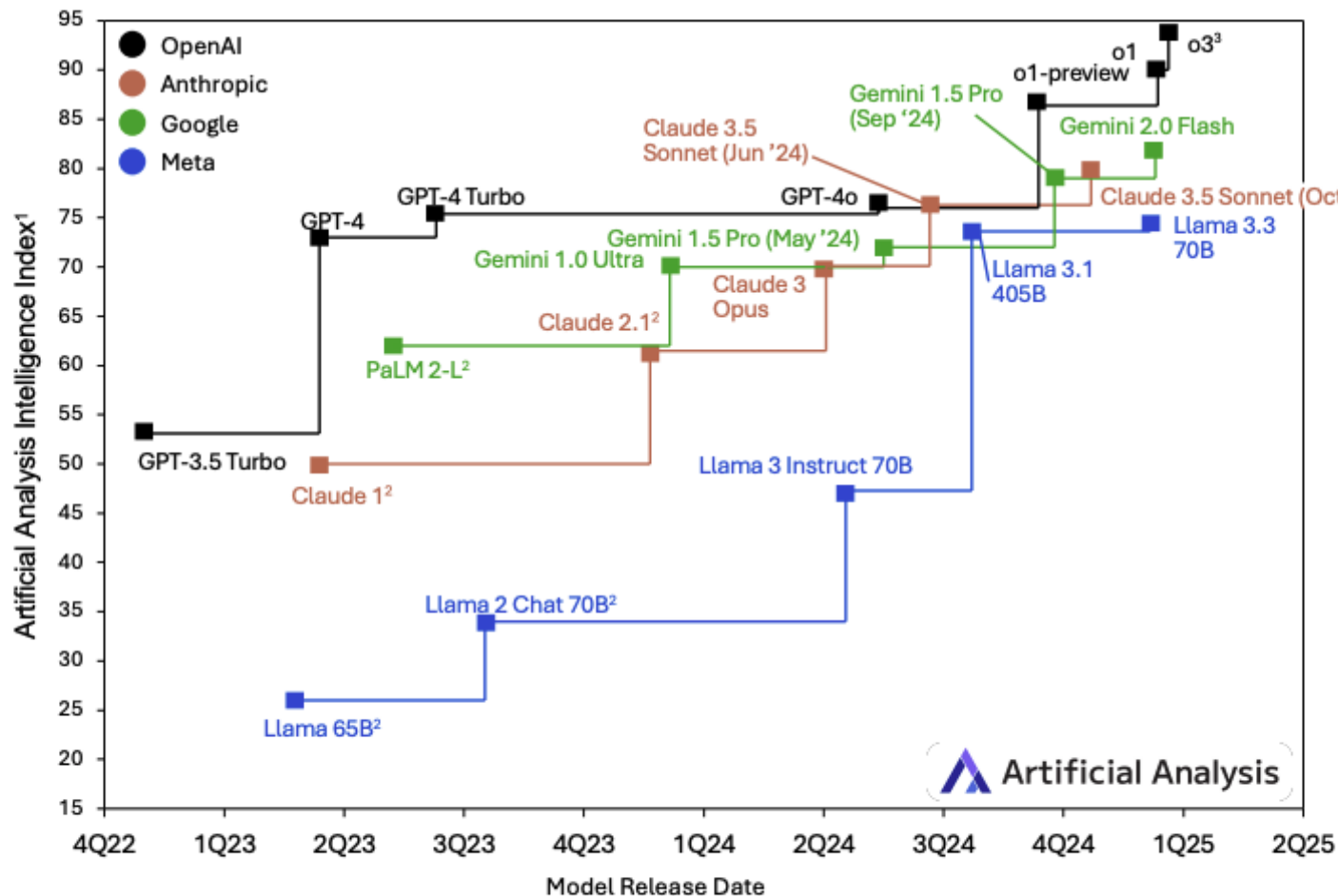
Open models close in on the frontier labs: Open weights models, led by those from DeepSeek and Alibaba, have approached o1 level intelligence.

1. Artificial Analysis Intelligence Index: average across a range of language model intelligence and reasoning evaluation datasets. Currently includes MMLU, GPQA Diamond, MATH-500 & HumanEval. Release date is based on first public launch of the model. 2. o3 Intelligence Index estimated by scaling measured Intelligence Index of o1. 3. Estimated based on company claims and comparable results where available, not yet independently benchmarked by Artificial Analysis



Since the launch of OpenAI's GPT-4 in early 2023, leading US AI labs have scrambled to catch up to OpenAI

Leading US AI Labs Frontier Language Model Intelligence, Over Time¹ Key Trends



Competing labs catch up to OpenAI's GPT-4: OpenAI started the language model race in November 2022 with the launch of GPT-3.5 in ChatGPT; leading US labs have largely caught up with frontier models from OpenAI.

Big Tech closes in on the frontier labs: Models from Google and Meta are rapidly closing in on frontier models, with Gemini 2.0 Flash exceeding Claude 3.5 Sonnet and GPT 4o capabilities.

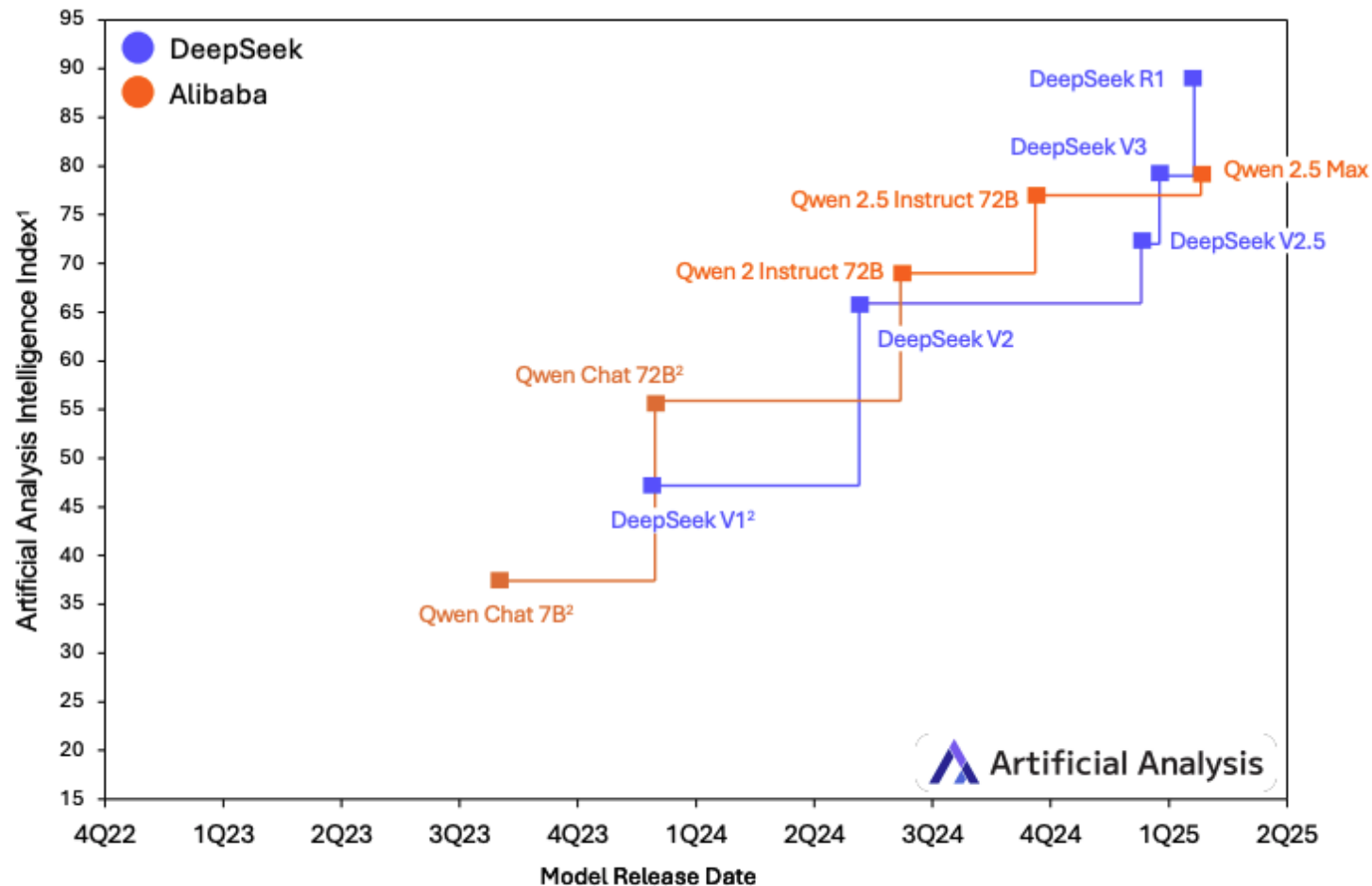
Sparks of intelligence beyond GPT-4: The final months of 2024 have seen the emergence of the first major intelligence leaps beyond GPT-4, led by OpenAI's o3. Topics including reasoning models, data quality and new reinforcement learning techniques have joined pre-training compute scaling as dominant levers for improving models.

1. Artificial Analysis Intelligence Index: average across a range of language model intelligence and reasoning evaluation datasets. Currently includes MMLU, GPQA Diamond, MATH-500 & HumanEval. Release date is based on first public launch of the model. 2. Estimated based on company claims and comparable results where available, not yet independently benchmarked by Artificial Analysis. 3. o3 Intelligence Index estimated by scaling measured Intelligence Score of o1.



Leading Chinese AI labs DeepSeek and Alibaba have steadily released new models, with DeepSeek taking the lead from Alibaba in late 2024

Leading Chinese AI Labs Language Model Intelligence, Over Time¹



Key Trends

Rapid improvements in intelligence: While Chinese AI labs joined the AI race later, they largely closed the intelligence gap with frontier US models in 2024. When OpenAI launched o1, Chinese labs produced a similarly performant model within months (DeepSeek's R1).

Leading with open weights models: Chinese AI labs, including Alibaba, DeepSeek and Tencent, have released open weights frontier models that are competitive with the leading models globally.

Potential leader in 2025: Early 2025 saw Chinese AI labs, including Alibaba, DeepSeek, MoonShot, Tencent, Zhipu, and Baichuan prolifically releasing frontier reasoning models. The release velocity and cadence suggest that Chinese AI labs are no longer laggards in 2025.

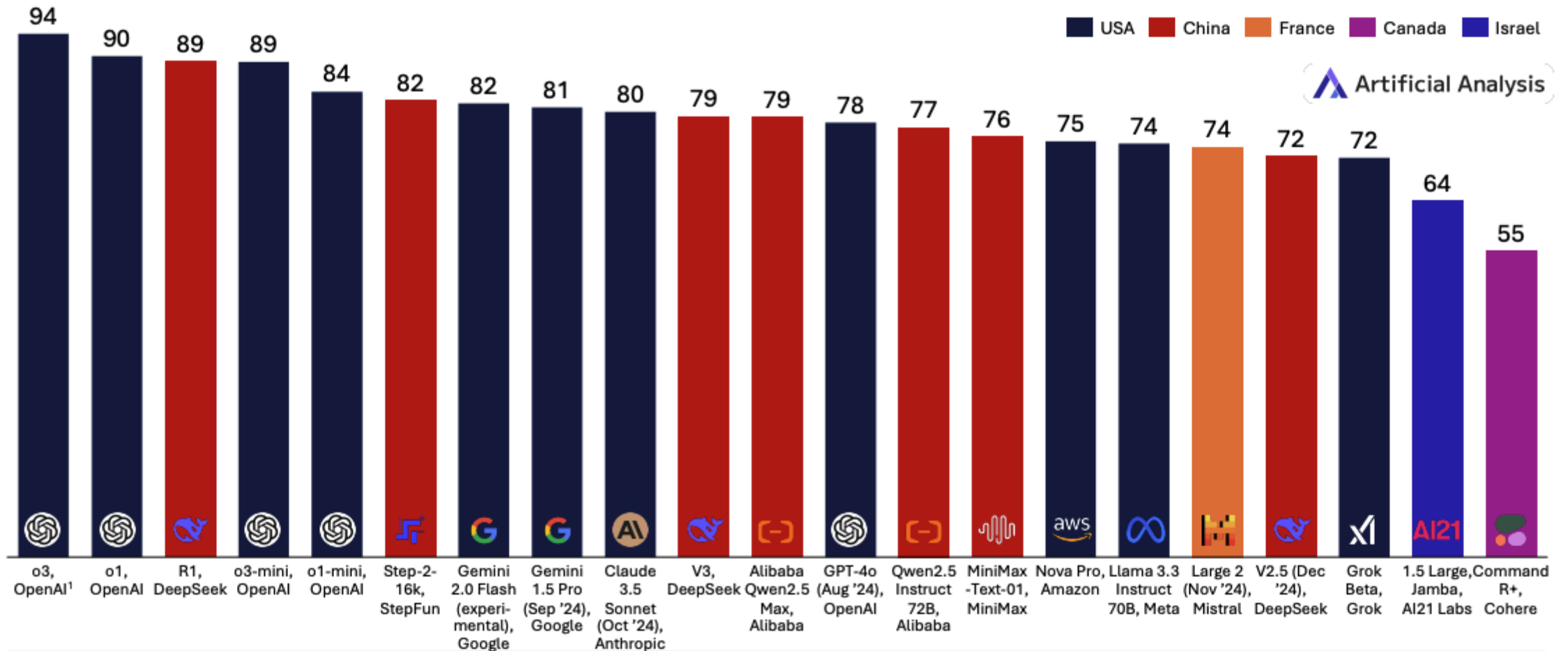
1. Artificial Analysis Intelligence Index: average across a range of language model intelligence and reasoning evaluation datasets. Currently includes MMLU, GPQA Diamond, MATH-500 & HumanEval. Release date is based on first public launch of the model. 2. Estimated based on company claims and comparable results where available, not yet independently benchmarked by Artificial Analysis



While the US maintains an overall lead in the intelligence frontier, China is no longer far behind. Few other countries have demonstrated frontier-class training

The Language Model Frontier: Country of Origin

Artificial Analysis Intelligence Index, Selected Leading Models (Early 2025), Non-exhaustive



1. Estimated based on company claims and comparable results where available, not yet independently benchmarked by Artificial Analysis
 2. A number of leading models from Chinese AI labs are excluded due to limited access or evaluation data

Escalating regulatory restrictions have banned the export of high-end AI accelerators to China (1/2)

Regulatory Restrictions

Unreleased	No Licence Required	NAC License Required	Presumption of Denial
------------	---------------------	----------------------	-----------------------

NVIDIA GPU Architecture	Model	Pre-Controls	October 2022 Controls ²	October 2023 Controls ^{3,4}	AI Diffusion Rules ⁵
	Announced		7-Oct-22	17-Oct-23	13-Jan-25
	Effective ¹		21-Oct-22	17-Nov-23	15-May-25
Blackwell	B200				
	B100				
Hopper	H100				
	H200				
	H800				
	H20				
Lovelace	L40S				
	L4				
	L40				
	L20				
	L2				
Ampere	A100				
	A800				
	A40				
	A30				
Consumer GPUs	RTX 6000 Ada				
	RTX 4090				
	RTX 4090D				
	RTX 3090				

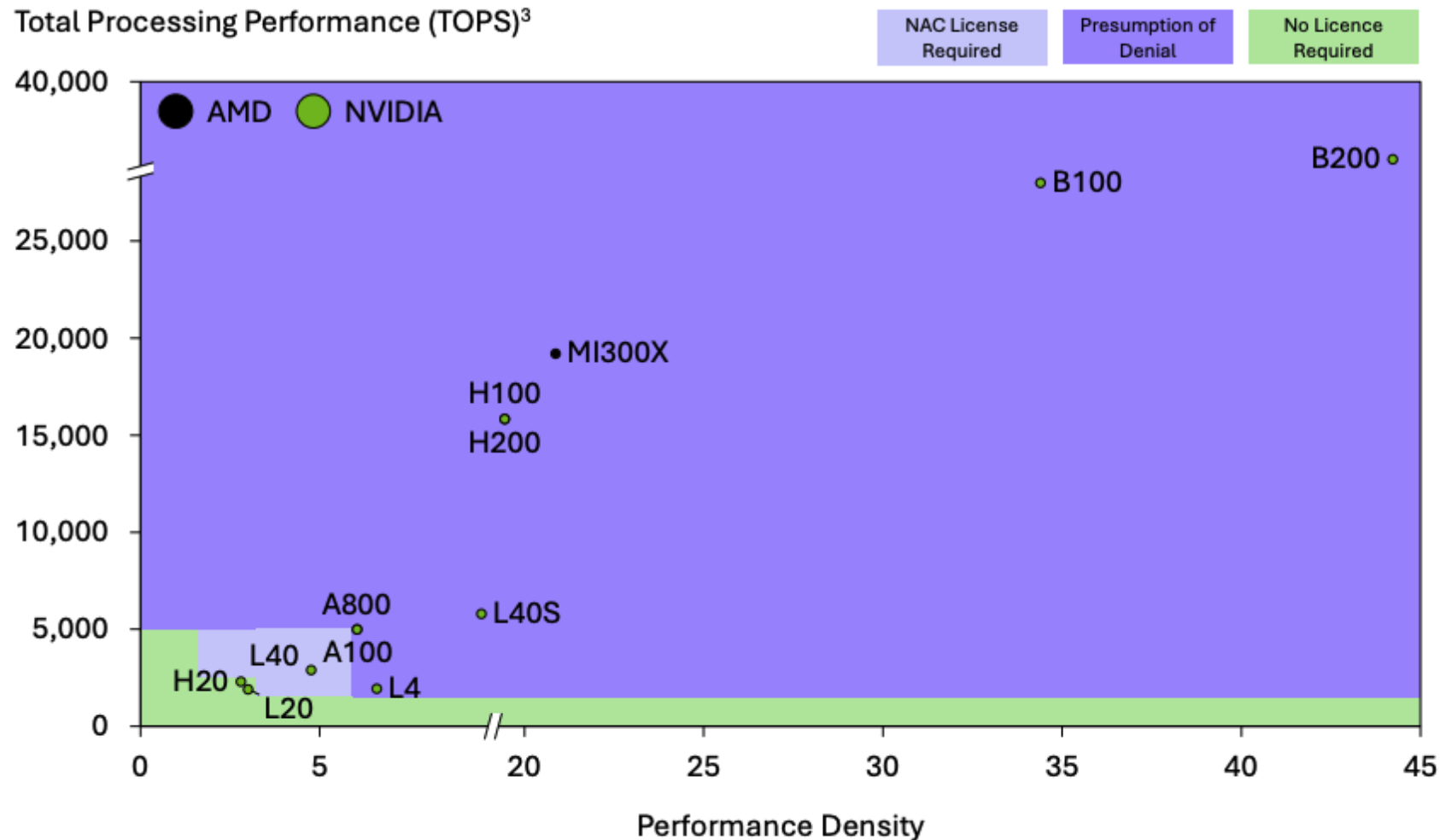
Commentary

- NVIDIA reacted quickly to both the October 2022 and October 2023 controls by releasing Hopper GPU variants that complied/comply with the regulations.** Specifically, after the H100 and A100 were banned for export to China, NVIDIA released the H800 and A800 with limited interconnect (see appendix for full Hopper generation specifications).
- The October 2023 controls went on to ban export of the H800 and A800 to China,** leading to NVIDIA **developing the H20** to continue selling a Hopper-generation GPU to Chinese customers. The H20 has limited compute (148 TFLOPs) compared to the H100 (989 TFLOPs)

1. Effective date refers to latest compliance date 2. [BIS](#) 3. [Georgetown CSET](#)
 4. [Federal Register](#) 5. [BIS](#)

US export controls restrict export of leading Nvidia accelerators based on performance and density thresholds; the H20 and L20 fall below these thresholds and can be freely exported

US Accelerators Prohibited for Export to China^{1,2}



Commentary

- The H20 and L20 are the only current NVIDIA data center-class AI accelerators that do not exceed either the Total Processing Performance or Performance Density threshold.
- While the H20 accelerator is currently available for sale in China, the Trump administration has started preliminary conversations around the potential inclusion of the chip on the restricted list, suggesting that **there may be a further broadening of the scope of restricted chips**

1. [SemiAnalysis](#) 2. [Georgetown CSET](#)

3. Total Processing Performance (TPP) measured in Tera Operations per Second, Performance Density measured as TPP / Die Size



You can help: Arenas

Artificial Analysis LANGUAGE MODELS ▾ SPEECH, IMAGE & VIDEO MODELS ▾ LEADERBOARDS ▾ ARENAS ▾ ABOUT ▾ Newsletter [Subscribe](#)


Arena Leaderboard Personal Leaderboard

VIDEO GENERATION MODEL ARENA + Submit prompt
0/30 to view your model preferences


[Try the new Speech Arena](#)

Which video best reflects this prompt?

Forensic scientist analyzing crime scene evidence.



Prefer (← Key)



Prefer (→ Key)