AI-AT-YOUR-SERVICE

Omar Mohout Partner, Head of Digital

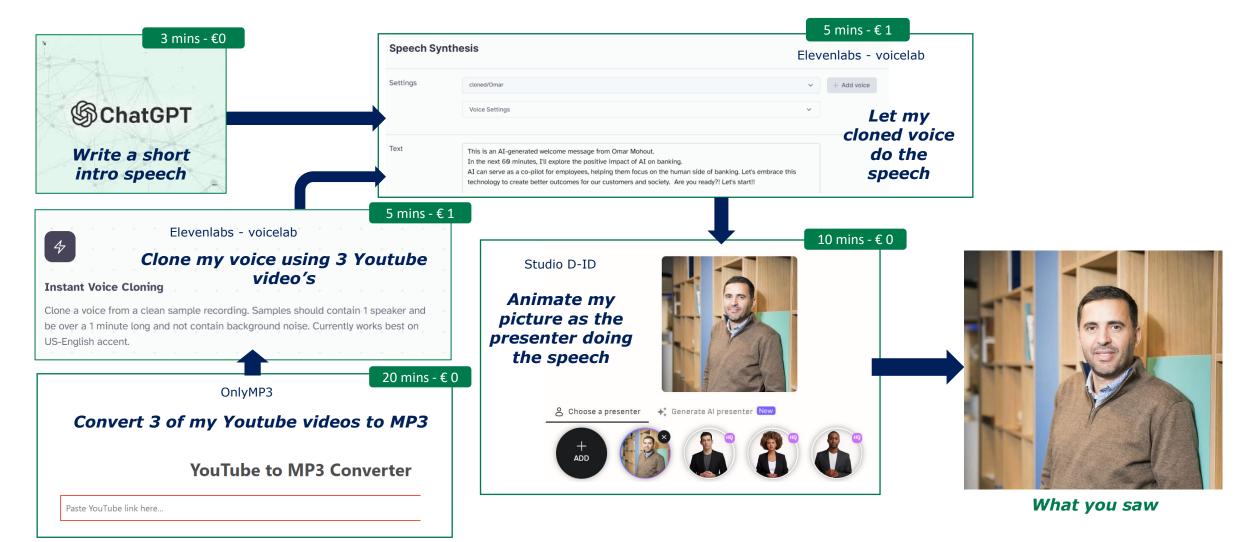








What did you just experience?

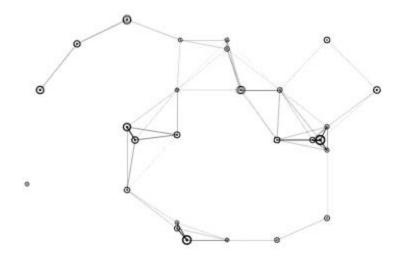


A short history of AI



Deep Blue/Watson (IBM) and AlphaGo (Google DeepMind) showcased the power of AI by defeating human world champions





Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar* Google Research nikip@google.com Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

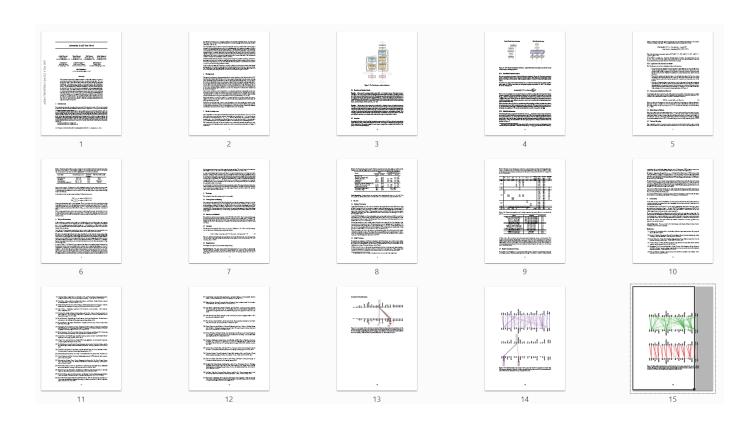
Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

1 Introduction

Recurrent neural networks, long short-term memory [13] and gated recurrent [7] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and

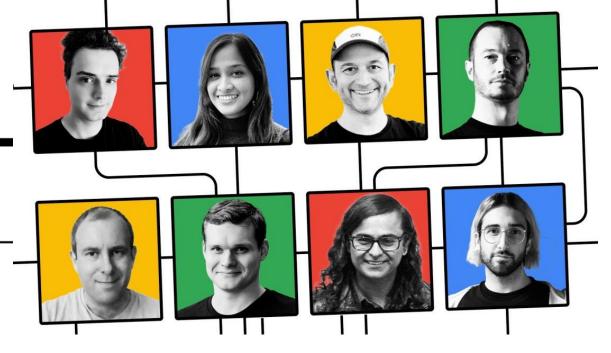
31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.



^{*}Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

[†]Work performed while at Google Brain.

[‡]Work performed while at Google Research.



Attention Is All You Need

Ashish Vaswani*

Google Brain avaswani@google.com

Adept Al Labs, \$415M

Llion Jones*

Google Research llion@google.com

Sakana Al

Noam Shazeer*

Google Brain noam@google.com

Character AI, \$200M

Aidan N. Gomez* †

University of Toronto aidan@cs.toronto.edu

Cohere AI, \$400M

Illia Polosukhin* ‡

illia.polosukhin@gmail.com NEAR Protocol, \$375M

Niki Parmar*

Google Research nikip@google.com

Adept Al Labs, \$415M

Jakob Uszkoreit* Google Research

usz@google.com

Inceptive Labs, \$20M

Łukasz Kaiser*

Google Brain

lukaszkaiser@google.com

OpenAl





Text-to-image revolution

Dall-E, Mid-Journey, Stable Diffusion...

Sam Altman @ @sama · Apr 6, 2022

DALL·E 2 is here! It can generate images from text, like "teddy bears

It's so fun, and sometimes beautiful.

openai.com/dall-e-2/



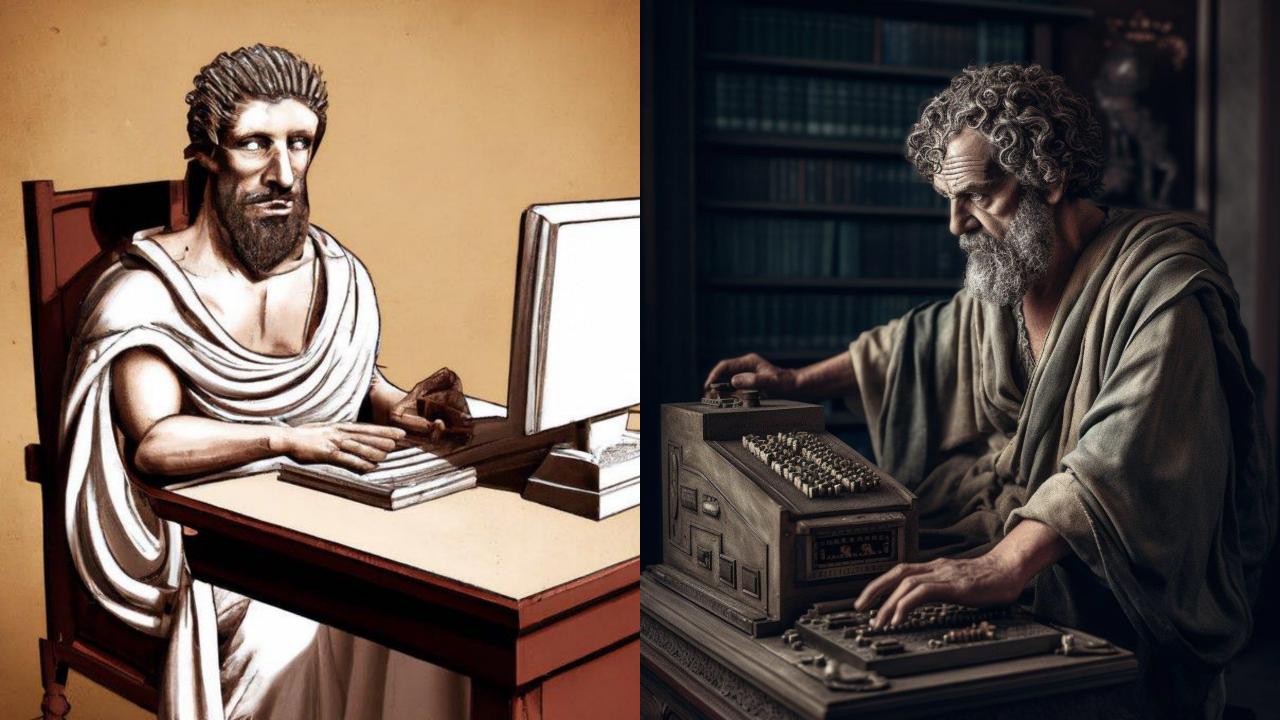




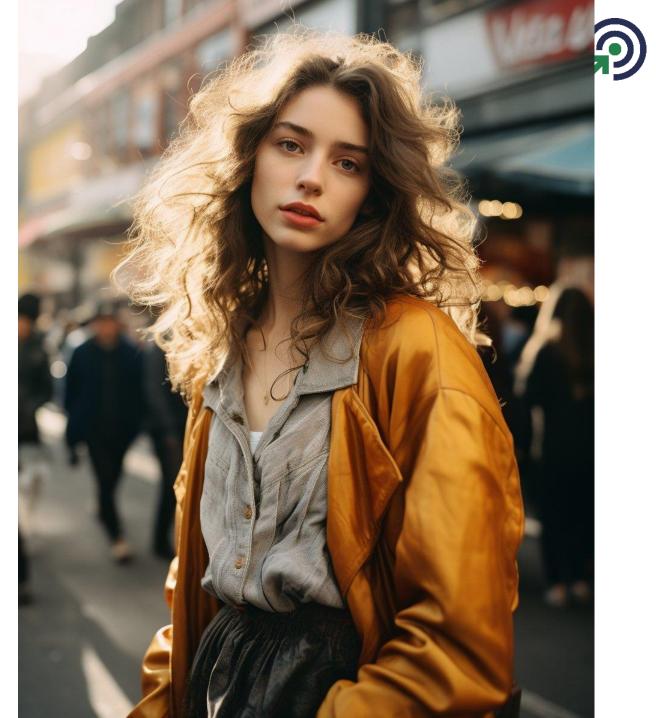








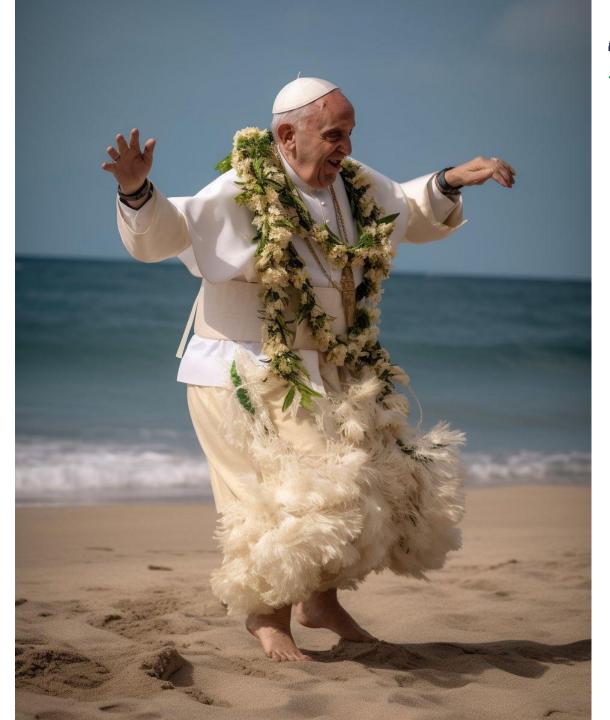






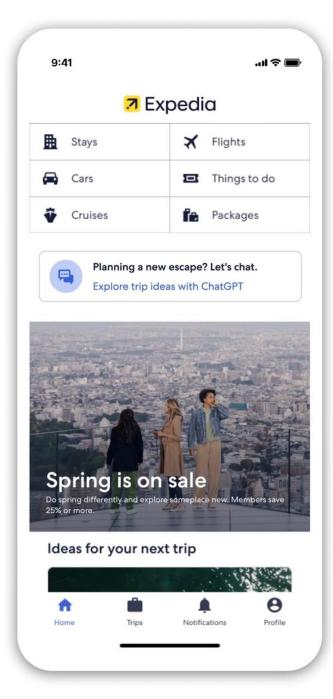
Pope Francis dancing Hula at the beach wearing a grass skirt and flower lei

Canon EOS R camera with a 50mm f/ 1. 8 lens, f/ 2. 2 aperture, shutter speed 1/ 200s, ISO 100 and natural light, Full Body, Hyper Realistic Photography





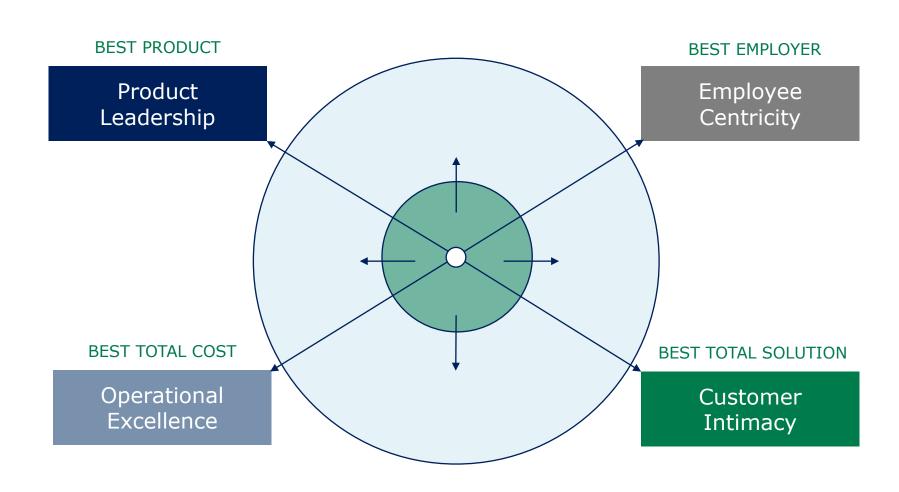
ChatGPT - plugins





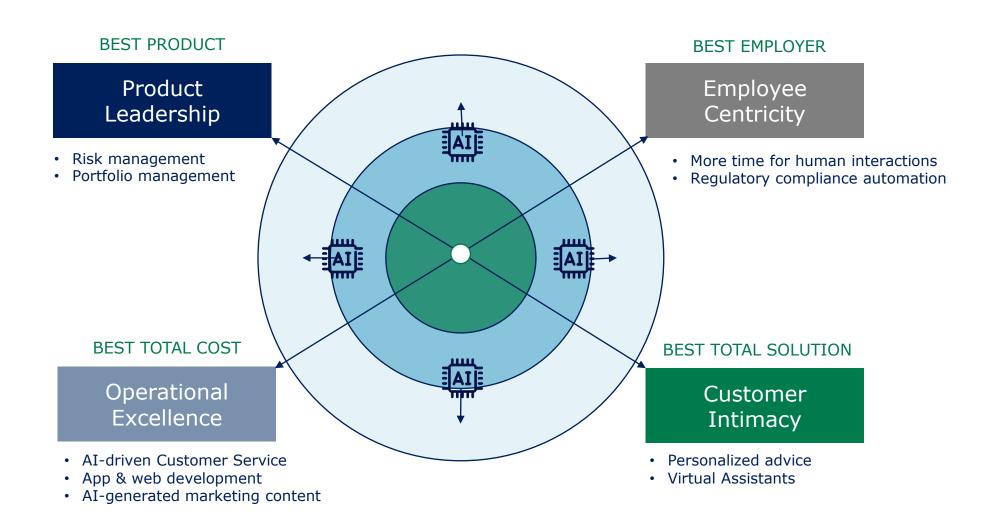


Businesses have seen increased expectations on standards, forcing them to play on more fields if they want to stay competitive





AI will enable companies to perform on all these 4 fields





Everyone overestimates how much AI will replace humans, and underestimates how much AI will empower humans



No one in this room
will be replaced by AI
but by someone
who is better with AI

What other companies are doing



A chatbot powered by OpenAI is being tested by Morgan Stanley's 16,000 financial advisors



FinChat.io v1.1

+ New Chat

Coverage

Chats

Q Search Coverage

This list shows the 750+ companies you can currently ask FinChat about.



- Apple Inc.
- Microsoft Corporation
- G Alphabet Inc.
- a Amazon.com, Inc.
- BH Berkshire Hathaway Inc.
- NVIDIA Corporation

Sign Up

Welcome to FinChat

Finchat.io is the ChatGPT for Finance. Using AI, FinChat generates answers to questions about public companies and investors. To generate accurate answers, FinChat provides reasoning, sources and data.

Summarize Microsoft's Q2 2023 quarterly report with a focus on the cloud business. Include transcript





DEMO

Embracing AI: A shared roadmap for Success





- 1. Strategize and Understand Business Objectives
- 2. Conduct a Data Audit
- 3. Choose the Right Tools and Platforms
- 4. Create a Cross-Functional Team
- 5. Adopt a Test-and-Learn Approach



- 6. Always put humans first
- 7. Playing = learning
- 8. Trust, but verify
- 9. You're the pilot, AI is the co-pilot
- 10. Figure out the impact on your customers



Learn to use AI and find the right tools to set yourself up for success

- Basic introduction to A.I. elementsofai.com
- Learn prompting <u>learnprompting.org</u>
- Largest AI tools directory: <u>www.futurepedia.io</u>
- Non-tech intro to A.I. course <u>www.coursera.org/learn/ai-for-everyone</u>
- Extensive introduction to A.I. www.udacity.com/course/intro-to-artificial-intelligence--cs271
- Machine Learning <u>www.coursera.org/learn/machine-learning</u>
- Python <u>www.youtube.com/watch?v= uQrJ0TkZlc&t=2278s</u>

29

More questions on how to make AI work for you?

Omar Mohout

omar.mohout@novareperta.com +32 474 711 368







Nova Reperta in a nutshell

Who?

A core team of over 60 FTEs, supported by a wide network of 100+ global functional/industry experts and a smart ecosystem

Nova Reperta is a management consulting boutique focusing on 'Strategy Deployment' Nova Reperta is a management consulting boutique focusing on 'Strategy Deployment'

= from 'Strategy Validation' to 'Strategy Execution'.

What?

We architect and help execute company-wide transformations to make them future proof

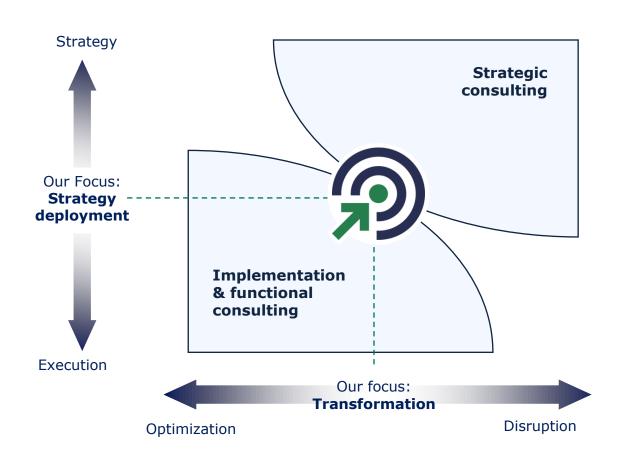
= profitable digital happy clients & employees with a positive contribution to soci

= profitable, digital, happy clients & employees,... with a positive contribution to society

Years of **expertise** among our ranks, yet with **different DNA** than our competitors...
...complemented by a **smart ecosystem**, to respond to our clients' challenges – end-to-end.



We have a distinctive positioning in the consulting market





Top & Down – A hands-on approach "from the trenches to the boardroom", connecting top to bottom



Hard & Soft – We provide skills applicable to all functions, and covering 'hard' and 'soft' elements of change management → € impact + capability building



Frontside vs. Backside – We have an eye for connecting front-stage (Customer-Centric functions) and back-stage (Operations)

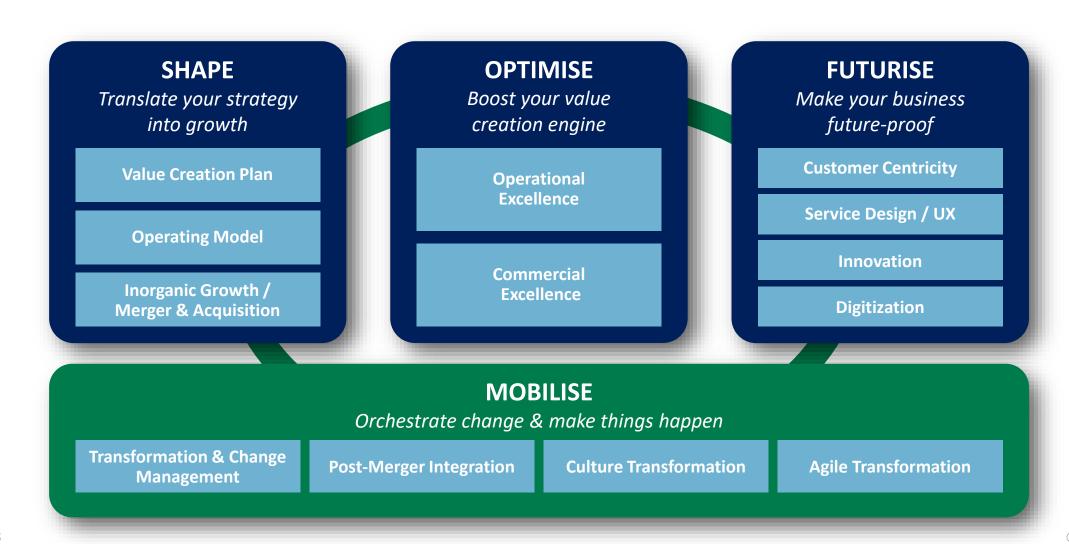


Digital savvy – Our consultants are capable of fully exploiting digital as the key enabler for your organization's growth

© Nova Reperta



We propose a range of services enabling your strategic deployment from A to Z – from validation to execution ...





5 core values we live by while helping companies



1 ENTREPRENEURIAL

2 PRAGMATIC

#WeAreNovaReperta

3 INCLUSIVE

5 ALWAYS EXPLORING

4 TRUSTWORTHY



Nova Reperta is led by a group of partners with diverse backgrounds supported by a well-balanced team with a high degree of seniority



François Barbellion



François Delfosse



Filip Leflot



Nick Dieltiens



Eljakim Caus



Omar Mohout

- 20+ years of experience in driving corporate change programs and making change stick.
- Founding Partner of Nova Reperta
- He has specific experience in leading large-scale transformations, across a variety of industries and functions, in both blue- and white-collar environments.
- Francois is an engineer with an MBA from London Business School.

- 20+ years of experience in consulting, business and Lean transformations
- Founding Partner of Nova Reperta
- Key areas of expertise include designing, delivering and steering Transformation Programs (500-2000 FTEs) across various industries and functions
- Ir Micro-Electronics, Université Catholique de Louvain.

- 20+ years of experience in consulting, private equity, finance, M&A and General Management
- Founding partner of Nova Reperta, co-lead of 2 investment funds
- Key areas of expertise include strategy design & implementation, financial and operational advisory, coaching of top management
- Previously, Partner at private equity-fund Bencis Capital and COO of AS Adventure Retail Group

- 20+ years of experience in business transformations, lean management, customer design & CX, change and program management
- Founding Partner of Nova Reperta, co-lead of the Nova Reperta Customer Centricity practice
- MBA Vlerick, General Management & master Handelswetenschappen

- 20+ years of experience in strategy consulting, merger management and large-scale lean transformations.
- Co-owner of an eco-friendly venture
- Partner of Nova Reperta, colead of Value Creation Plan practice
- Eljakim holds a master in civil engineering from TU Delft

- 8 years of experience in digital transformation and deployment of go-to-market strategies for high growth companies
- Relevant experience in developing and executing buy, build and ally strategies for +100 software companies to seek to grow internationally
- Recognized hands-on expert and thought leader in the Belgian scaleup ecosystem with a broad network in the European tech scene

Large Language Models in finance

Laurent Sorber

Financieel Forum Vlaams-Brabant Leuven, 20 November 2023



Outline

- Who are InvestSuite & Radix
- What are Large Language Models
- The magic of LLMs
- Dispelling the magic of LLMs
- Applications of LLMs
- What does the future hold



Outline

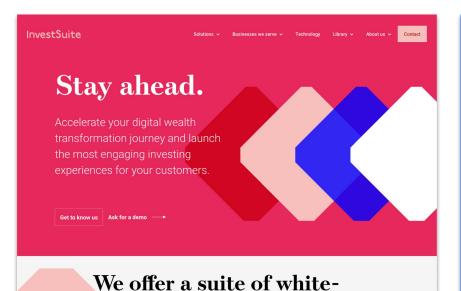
- Who are InvestSuite & Radix
- What are Large Language Models
- The magic of LLMs
- Dispelling the magic of LLMs
- Applications of LLMs
- What does the future hold





Radix

A suite of investtech products with which we want to make investing more accessible to a wider audience **Bespoke AI solutions** with which we want to superpower people for clients across industries and AI domains



label InvestTech solutions.







A suite of investtech products with which we want to make investing more accessible to a wider audience

Founded 2018

Offices in Leuven, CH, UK

50+ colleagues

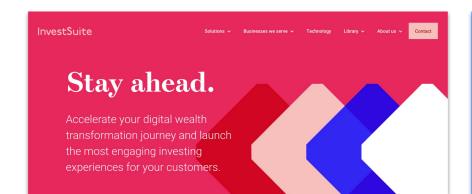
Sales presence in BE, DE, NL, USA **Bespoke AI solutions** with which we want to superpower people for clients across industries and AI domains

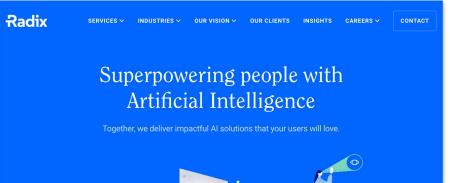
Founded 2018

Offices in Brussels & Ghent

50+ colleagues

Active in BE, NL, DE, IT, Nordics, USA









Portfolio Construction



Investment
Insights & Reporting



Robo Advisor Self Investor

Portfolio Optimizer

StoryTeller

Insight API

Radix









Al solutions

- Al modelling
- Software engineering
- Cloud infrastructure
- MLOps
- Data annotation

Knowledge transfer

- Training
- Coaching
- Team extensions

Al strategy

- Discover the best opportunities in Al
- Value & effort estimation of AI opportunities

Research

- Develop new AI models
- Explore new applications
- Validate research results

Radix



Natural Language Processing

- Text generation
- Question answering
- Summarization
- Named entity recognition
- Classification
- Duplicate detection
- Information retrieval

Classification

- Regression
- Anomaly detection
- Explainable prediction

Tabular Data 📊

- Confidence intervals
- Matrix imputation
- Feature engineering
- Data visualization

Computer Vision ®

- Image detection
- Object detection
- · Semantic segmentation
- Instance segmentation
- Real-time detection
- Hyperspectral imaging
- Super-resolution
- Optical Character Recognition

Planning 17

- Scheduling optimization
- Assignment optimization
- Inventory optimization
- Routing optimization

Time Series M

- Forecasting
- Anomaly detection
- Classification
- Confidence intervals

Speech 🥕

- Transcription
- Speech recognition
- Voice cloning
- Speaker identification

Outline

- Who are InvestSuite & Radix
- What are Large Language Models
- The magic of LLMs
- Dispelling the magic of LLMs
- Applications of LLMs
- What does the future hold





What is an LLM?

The best thing about AI is its ability to

learn	4.5%
predict	3.5%
make	3.2%
understand	3.1%
do	2.9%

What is an LLM?

The best thing about AI is its ability to create

The best thing about AI is its ability to create worlds

The best thing about AI is its ability to create worlds that

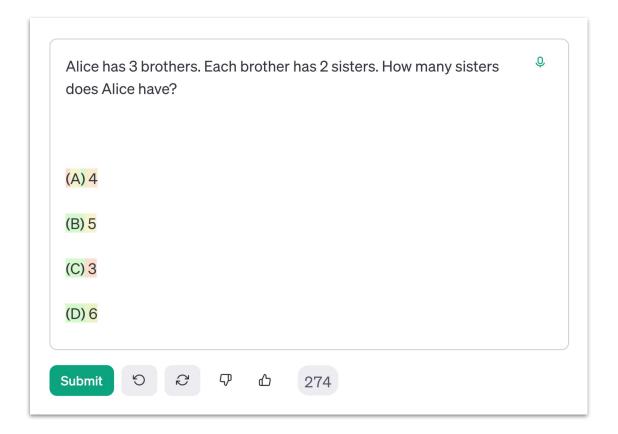
The best thing about AI is its ability to create worlds that are

The best thing about AI is its ability to create worlds that are both

The best thing about AI is its ability to create worlds that are both exciting

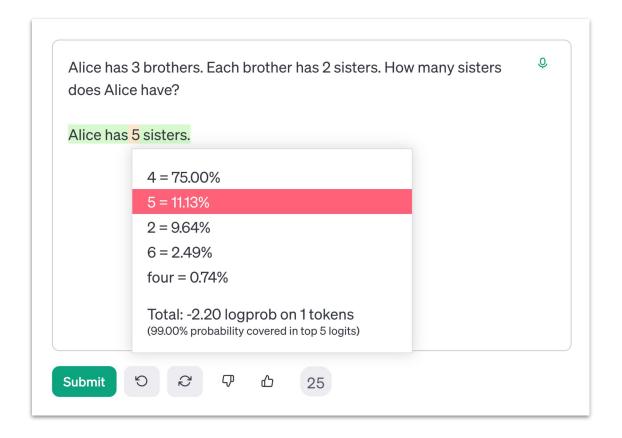
Base LLM:

Completes documents



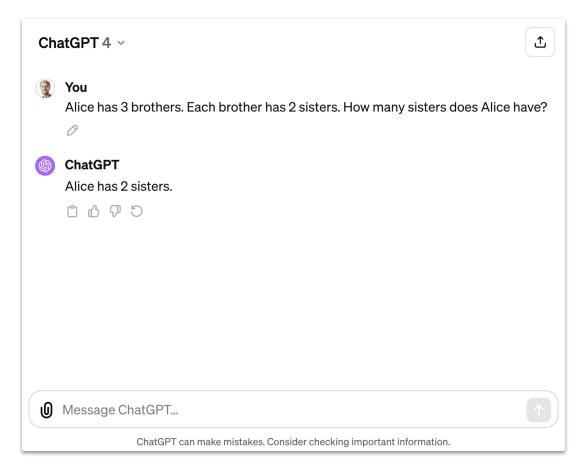
Instruction tuned LLM:

Answers questions





A web app on top of GPT-4



Outline

- Who are InvestSuite & Radix
- What are Large Language Models
- The magic of LLMs
- Dispelling the magic of LLMs
- Applications of LLMs
- What does the future hold



Demo

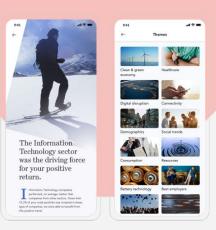




Portfolio Construction



Investment
Insights & Reporting



Robo Advisor Self Investor

Portfolio Optimizer

StoryTeller

Insight API





You



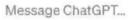
Here are some stock prices. What are the annualised return and risk of these stocks? Show me with a graph.





Laurent Sorber







The keys to success



Outline

- Who are InvestSuite & Radix
- What are Large Language Models
- The magic of LLMs
- Dispelling the magic of LLMs
- Applications of LLMs
- What does the future hold



All LLMs are Transformers

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer' Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu Łukasz Kaiser* Google Brain

lukaszkaiser@google.com

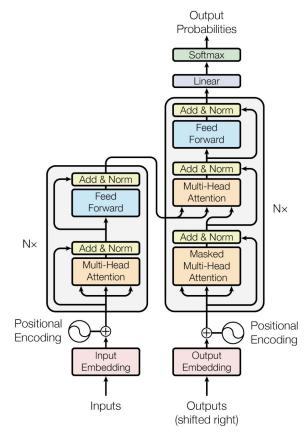
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 Englishto-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Work performed while at Google Research.





^{*}Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

Work performed while at Google Brain.

Input text

Tokenize

LLM

Output probabilities Sample

Weights & Biases is the [1135, 2337, 1222, 8436, 1386, 318, 262]



$$\begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 0 \\ 2 & 3 & 4 \end{bmatrix} \times \begin{bmatrix} 2 & 5 \\ 6 & 7 \\ 1 & 8 \end{bmatrix}$$

***	0.001 0.001 0.001		
We			
ights			
&			
the			
machine	0.15		
learning	0.001		
platform	0.2		

machine

¹ LLM image source: https://wandb.ai/darek/llmapps/reports/A-Gentle-Introduction-to-LLM-APIs--Vmlldzo0NjM0MTMz

² GPU image source: https://www.nvidia.com/en-us/data-center/h100/

³ Matrix multiplication source: https://math.stackexchange.com/questions/3715598/when-should-hadamard-matrix-multiplication-be-used-vs-other-methods

What do LLMs imitate?

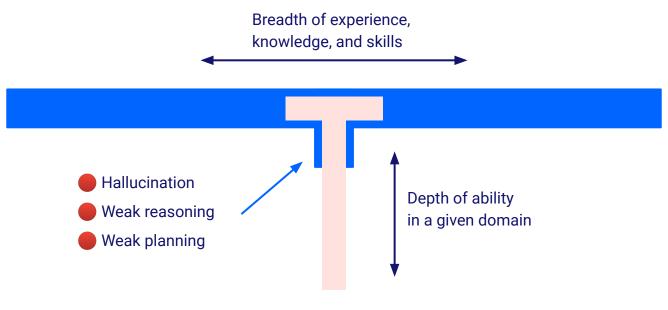
Cognitive ability

X Consciousness

X Sentience

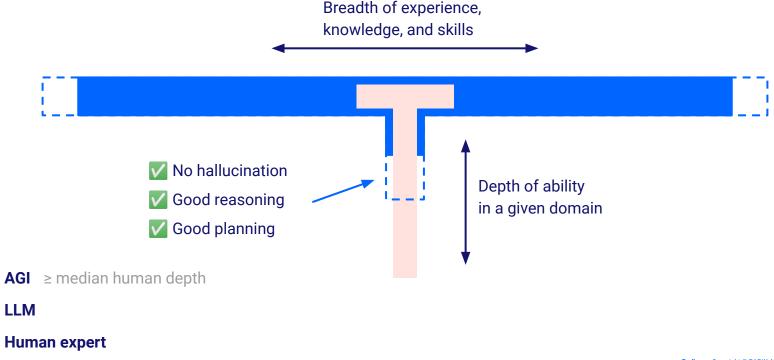


The T-shape of an LLM

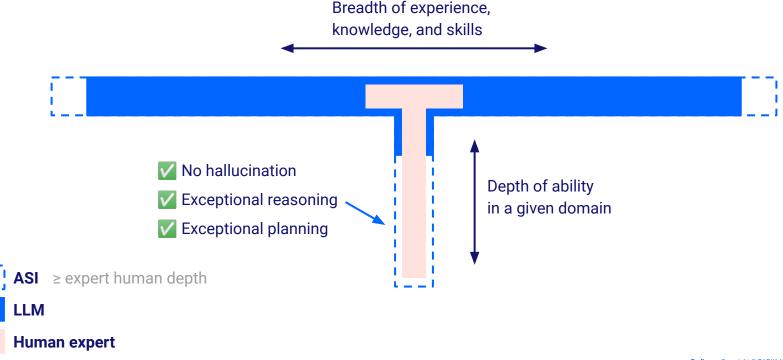




The T-shape of an LLM



The T-shape of an LLM

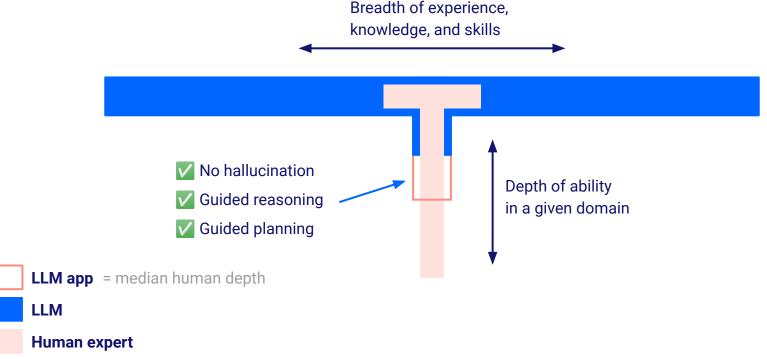


Outline

- Who are InvestSuite & Radix
- What are Large Language Models
- The magic of LLMs
- Dispelling the magic of LLMs
- Applications of LLMs
- What does the future hold



The T-shape of an LLM app

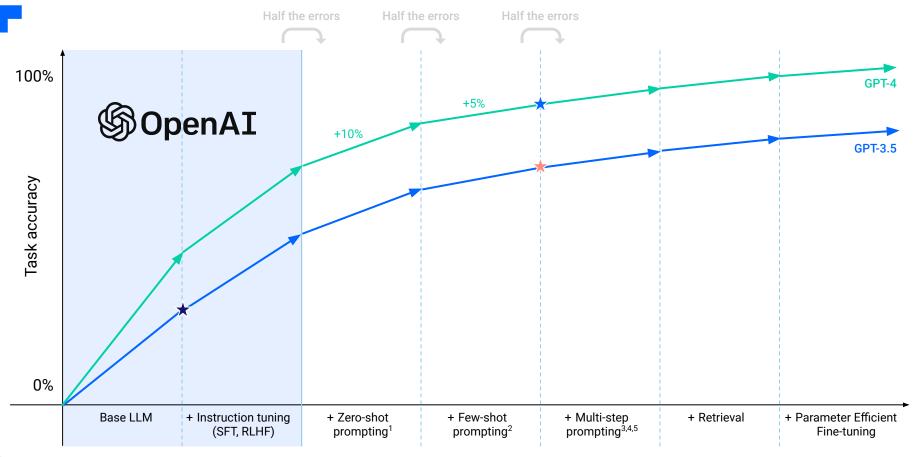


Anatomy of an LLM app

LLM application Software tools **Audio & Video** Speech, images, video Python, applications LLM Context **Files Network** window Confluence, Sharepoint Databases, internet

LLM app: Portfolio Assistant

LLM application Software tools Audio & Video Python, applications LLM Context **Files** Network window Confluence, Sharepoint



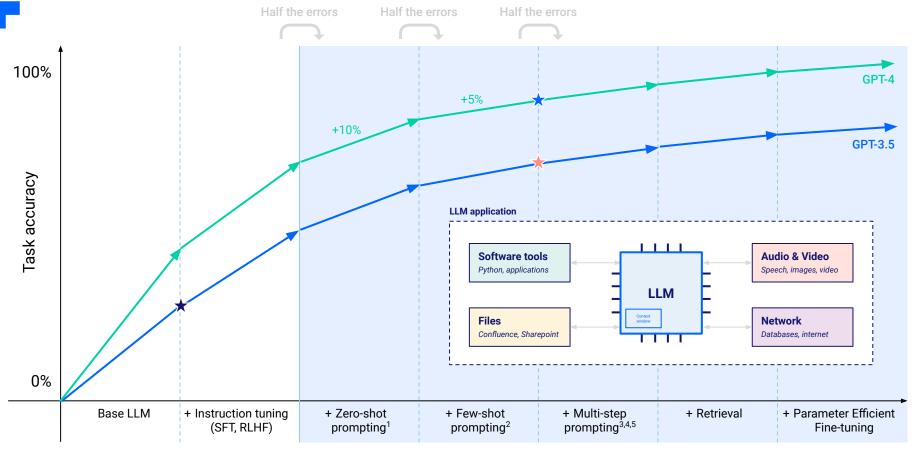
^{*} GPT-4 achieves a few-shot chain-of-thought performance on MMLU of 86%: https://arxiv.org/abs/2305.17306

^{*} GPT-3.5 achieves a few-shot chain-of-thought performance on MMLU of 67%; https://arxiv.org/abs/2305.17306
* GPT-3 achieves a few-shot performance on MMLU of 44%; https://arxiv.org/pdf/2009.03300.pdf

¹ An automatically discovered chain-of-thought prompt generalizes to novel models and datasets; https://arxiv.org/pdf/2305.02897.pdf

² Language Models are Few-Shot Learners: https://arxiv.org/pdf/2005.14165.pdf

^{3.45} Self-consistency: https://arxiv.org/pdf/2203.11171.pdf, Self-reflection: https://evjang.com/2023/03/26/self-reflection.html, Tree of Thoughts: https://arxiv.org/pdf/2305.10601.pdf



^{*} GPT-4 achieves a few-shot chain-of-thought performance on MMLU of 86%: https://arxiv.org/abs/2305.17306
* GPT-3.5 achieves a few-shot chain-of-thought performance on MMLU of 67%: https://arxiv.org/abs/2305.17306
* GPT-3 achieves a few-shot performance on MMLU of 44%: https://arxiv.org/pdf/2009.03300.pdf

¹ An automatically discovered chain-of-thought prompt generalizes to novel models and datasets; https://arxiv.org/pdf/2305.02897.pdf

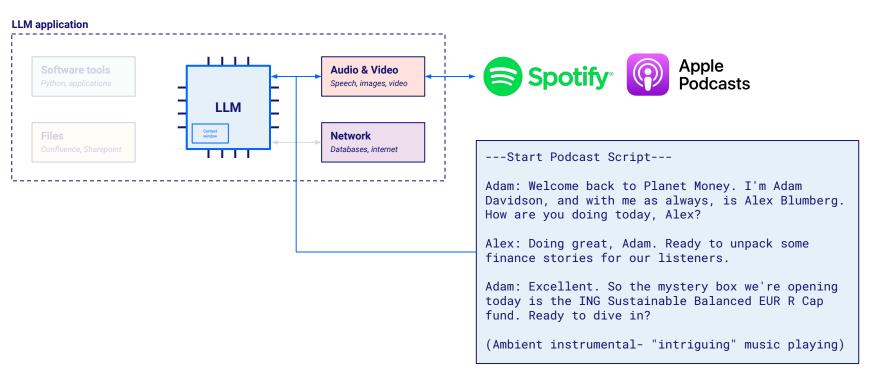
² Language Models are Few-Shot Learners: https://arxiv.org/pdf/2005.14165.pdf

^{3.45} Self-consistency: https://arxiv.org/pdf/2203.11171.pdf, Self-reflection: https://evjang.com/2023/03/26/self-reflection.html, Tree of Thoughts: https://arxiv.org/pdf/2305.10601.pdf

LLM app: StoryTeller podcast

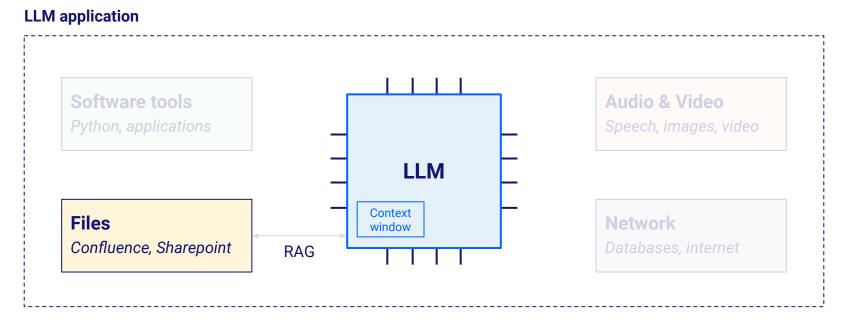
LLM application Software tools **Audio & Video** Speech, images, video LLM Context **Files Network** window Databases, internet

LLM app: StoryTeller podcast

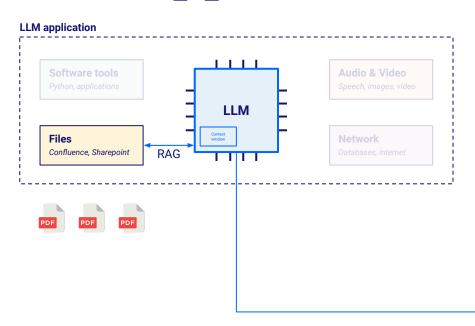


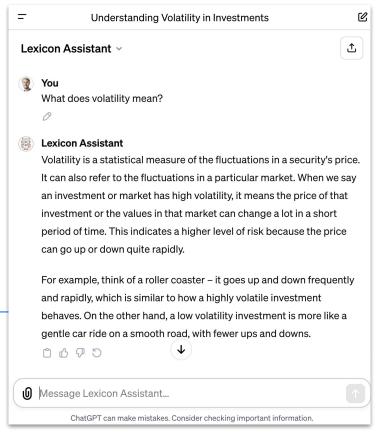


LLM app: Lexicon Assistant



LLM app: Lexicon Assistant





Outline

- Who are InvestSuite & Radix
- What are Large Language Models
- The magic of LLMs
- Dispelling the magic of LLMs
- Applications of LLMs
- What does the future hold











Talk to Pi, your personal AI

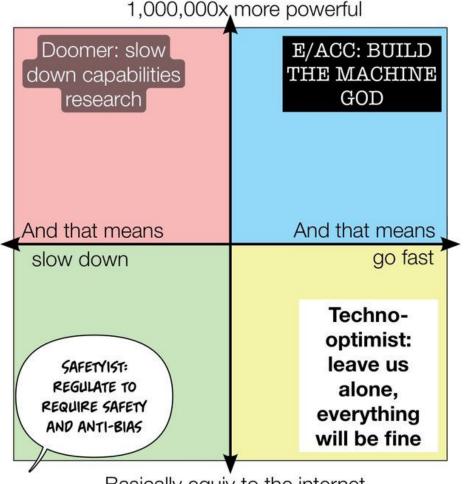
heypi.com

Personal Computer

Personal Intelligence



AI political compass



Basically equiv to the internet

What are the H100s working on?









Hallucination

Alignment, Interpretability

Reasoning & planning

Reasoning & planning





cohere





Multi-modality



Retrieval

Reasoning & planning

'Truth'

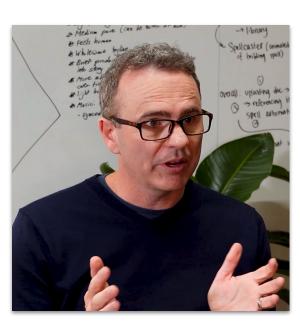
AGI, when?



"Less than 10 years" 1 Sam Altman (OpenAl)



"Within 18 months"²
Mustafa Suleyman (Inflection AI)



"As early as 2025, expected 2028"³ Shane Legg (Google DeepMind)

¹ Sam Altman (OpenAI) interview, 21 October 2023: https://www.youtube.com/watch?v=byYlC2caqLw

² Mustafa Suleyman (Inflection AI) interview, 14 November 2023: https://www.youtube.com/watch?v=VUOPClgi-7U

³ Shane Legg (Google DeepMind) interview, 26 October 2023: https://www.youtube.com/watch?v=Kc1atfJkiJU

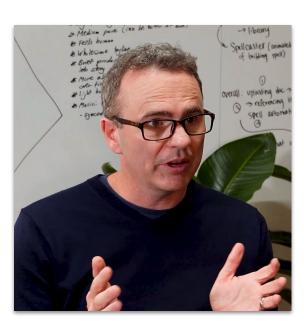
AGI, when?



"Less than 10 years" 1 Sam Altman (OpenAl)



"Within 18 months"²
Mustafa Suleyman (Inflection AI)



"As early as 2025, expected 2028"³ Shane Legg (Google DeepMind)

¹ Sam Altman (OpenAI) interview, 21 October 2023: https://www.youtube.com/watch?v=byYlC2caqLw

² Mustafa Suleyman (Inflection AI) interview, 14 November 2023: https://www.youtube.com/watch?v=VUOPClgj-7U

³ Shane Legg (Google DeepMind) interview, 26 October 2023: https://www.youtube.com/watch?v=Kc1atfJkiJU

Thank you!

MM InvestSuite

Radix

