Emerging Trends in AI

Cory Hymel

VP Research / Crowdbotics

Short History of AI

Symbolic Era	Connectionist Era	Generative Era
FIRST ERA (1956-1990)	SECOND ERA (1990 – 2010)	THIRD ERA (2010 – PRESENT)

1956: The Dartmouth Conference - birthplace of AI research

1967: The first computer program capable of natural language understanding

1974: The MYCIN expert system for medical diagnosis is developed.

1981: The first autonomous mobile robot, the Stanford Cart

1997: Deep Blue defeats world chess champion Garry Kasparov.

2006: Amazon introduces its recommendation on system

2012: Google develops the deep learning neural network

2015: Google's DeepMind defeats AlphaGo

2022: OpenAI releases ChatGPT

Types of AI

AI

Machine Learning

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning
- Deep Learning
- Transfer Learning

Natural Language Processing

- Text Generation
- Sentiment Analysis
- Machine Translation
- Speech Recognition
- Named Entity Recog.

Computer Vision

- Image Recognition
- Object Detection
- Image Generation
- Facial Recognition
- Scene Understanding

Robotics

- Autonomous Systems
- Robot Perception
- Robot Planning and | Control
- Human-Robot Interaction

AI Ethics & Policy

- Fairness and Bias in AI
- Ethical AI
- AI Governance

Expert Systems

- Rule-Based Systems
- Knowledge Representation
- Inference Engines

Knowledge & Reasoning (KRR)

- Ontologies
- Logic and Automated Reasoning
- Semantic Web

Reinforcement Learning

- Markov Decision
- Processes
- Multi-Agent Systems
- Policy Optimization

Cognitive Computing

- Simulation of Human Thought Processes
- Decision Making
- Problem Solving

Evolutionary Computation

- Genetic Algorithms
- Genetic Programming
- Swarm Intelligence

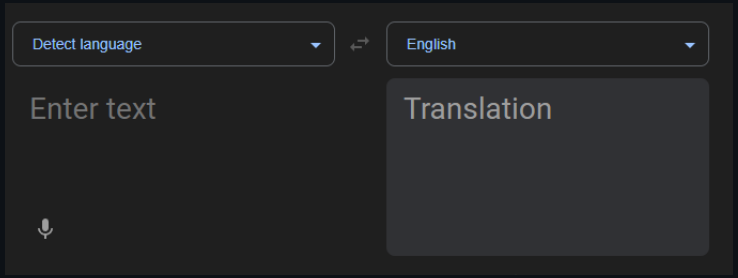
Real Examples

Natural Language Processing

Machine Learning

Computer Vision

Google Translate



amazon

TESLA

AI in Industry

airbnb

Cloud Pricing Optimization

Airbnb is one company using AI to optimize pricing on AWS, utilizing AI to manage capacity, to build custom cost and usage data tools, and to optimize storage and computing capacity.

Dropbox is another company that is using AI to optimize cloud costs and operational expenditures, reducing its dependency on AWS and saving nearly $75 million in the process.

ESTĒE LAUDER

Voice assistance and chatbots

Estée Lauder, the company has released a voice-enabled makeup assistant designed to assist visually impaired people with applying makeup.

Meanwhile, companies such as Pentagon Credit Union (PenFed) are using chatbots and conversational AI to help customers get answers to common questions faster, reducing the load on customer service reps.

Predictive maintenance

At GE, AI is leveraged regularly for predictive maintenance, analyzing data directly from aircraft engines to identify any problems, needed maintenance, and to ensure the overall safety of aircrafts.

The District of Columbia Water and Sewer Authority is using predictive maintenance to identify potential water main breaks and to monitor performance of collection systems.

Unilever

Cust. Service operations

McKinsey reports that around 67% of millennials “expect real-time customer service,” and 75% of customers expect a “consistent cross-channel service experience.”

Unilever, which is leveraging GPT API to create AI tools to minimize food waste and auto-generate product listings, is also using the API to create a platform that filters emails sent to customer service, sorting spam from legitimate messages, and scaling those up to customer service agents.

Generative AI

Generative AI refers to algorithms that create new content, such as text, images, or audio, similar to the data they were trained on. Unlike traditional AI, which focuses on analyzing data and making predictions, generative AI produces novel data that mimics the training data.

G **Generative:** Refers to the model's ability to generate text.

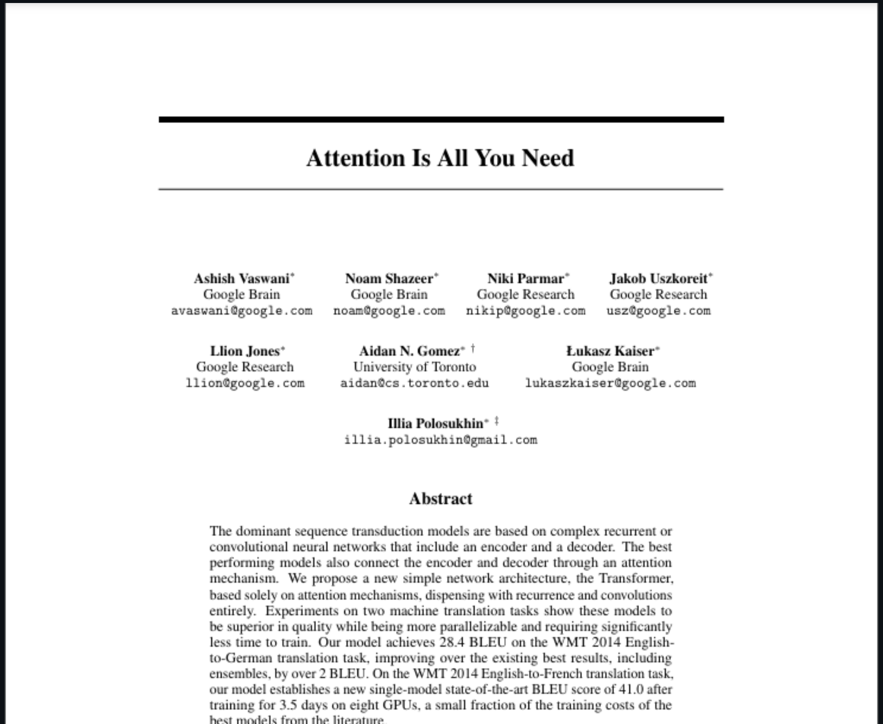
P **Pre-trained:** Indicates that the model is initially trained on a large corpus of text data.

T **Transformer:** The underlying architecture of the model.

Tale of 2 Papers

Google Brain

2017



Attention Is All You Need

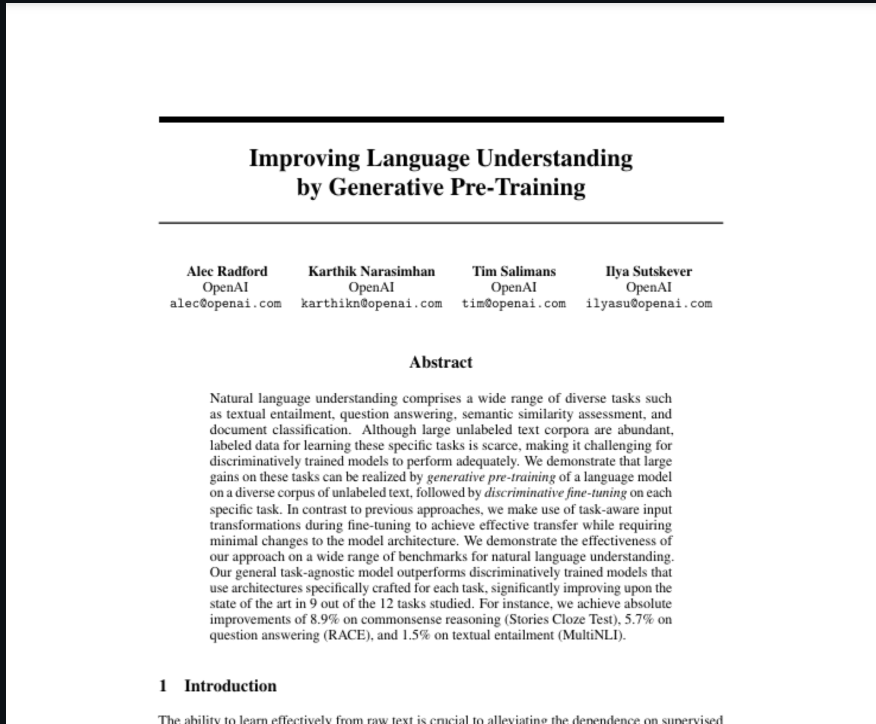
Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* 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 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

OpenAI

2018



Improving Language Understanding by Generative Pre-Training

Alec Radford OpenAI alec@openai.com
Karthik Narasimhan OpenAI karthikn@openai.com
Tim Salimans OpenAI tim@openai.com
Ilya Sutskever OpenAI ilyasu@openai.com

Abstract

Natural language understanding comprises a wide range of diverse tasks such as textual entailment, question answering, semantic similarity assessment, and document classification. Although large unlabeled text corpora are abundant, labeled data for learning these specific tasks is scarce, making it challenging for discriminatively trained models to perform adequately. We demonstrate that large gains on these tasks can be realized by *generative pre-training* of a language model on a diverse corpus of unlabeled text, followed by *discriminative fine-tuning* on each specific task. In contrast to previous approaches, we make use of task-aware input transformations during fine-tuning to achieve effective transfer while requiring minimal changes to the model architecture. We demonstrate the effectiveness of our approach on a wide range of benchmarks for natural language understanding. Our general task-agnostic model outperforms discriminatively trained models that use architectures specifically crafted for each task, significantly improving upon the state of the art in 9 out of the 12 tasks studied. For instance, we achieve absolute improvements of 8.9% on commonsense reasoning (Stories Cloze Test), 5.7% on question answering (RACE), and 1.5% on textual entailment (MultiNLI).

1 Introduction

The ability to learn effectively from raw text is crucial to alleviating the dependence on supervised

So how does it work?

LLMs are basically autocomplete on steroids. Their job is to predict the next work in a sequence.

`[ The cat likes to sleep in the _______ ]`

Problem Statement:

What word comes next?

So how does it work?

LLMs are basically autocomplete on steroids. Their job is to predict the next work in a sequence.

`[ The cat likes to sleep in the ________ ]`

Word	Probability
ability	0.002
bag	0.071
box	**0.085**
….	….
zebra	0.0001

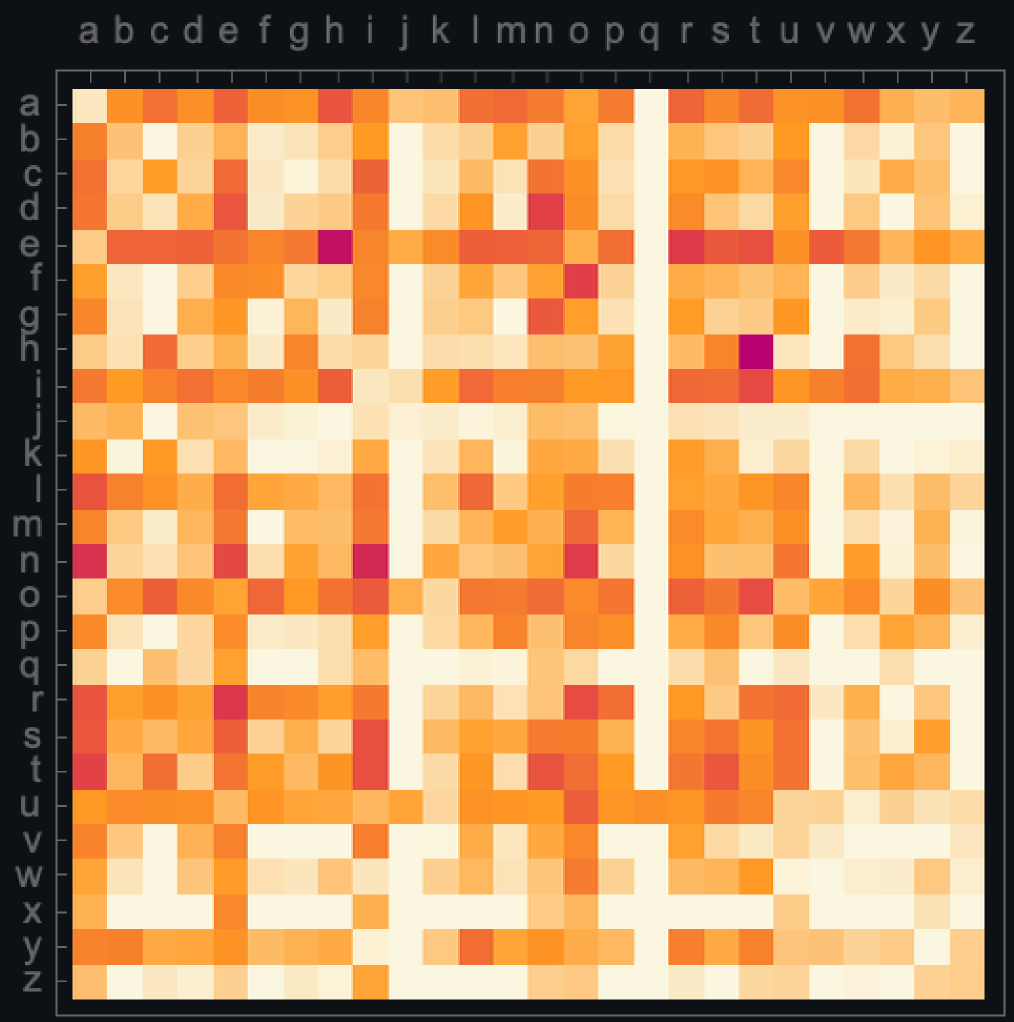
We do this over and over

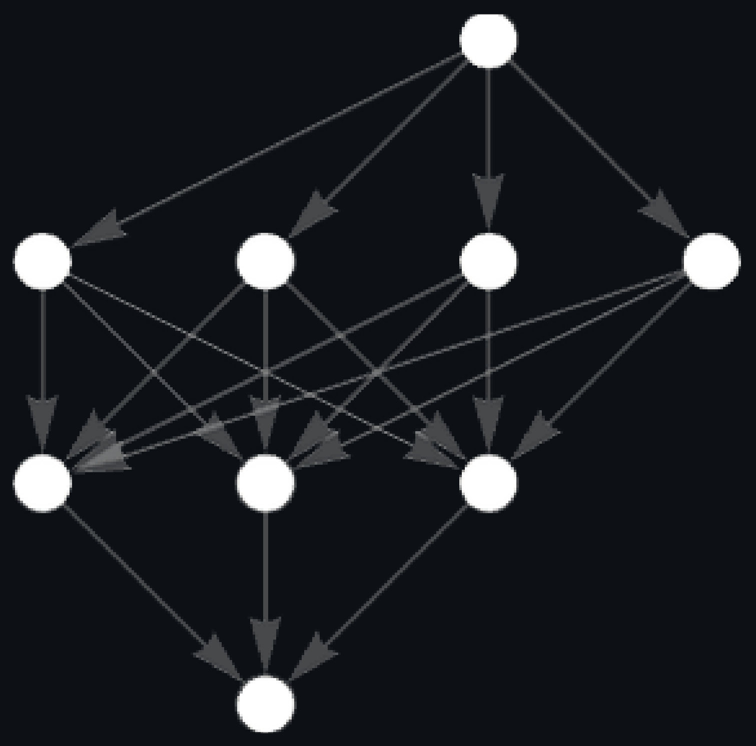
```
[ The cat likes to sleep in the ]
[ The cat likes to sleep in the ]
[ The cat likes to sleep in the ]
[ The cat likes to sleep in the ]
[ The cat likes to sleep in the ]
```

But what does it look like?

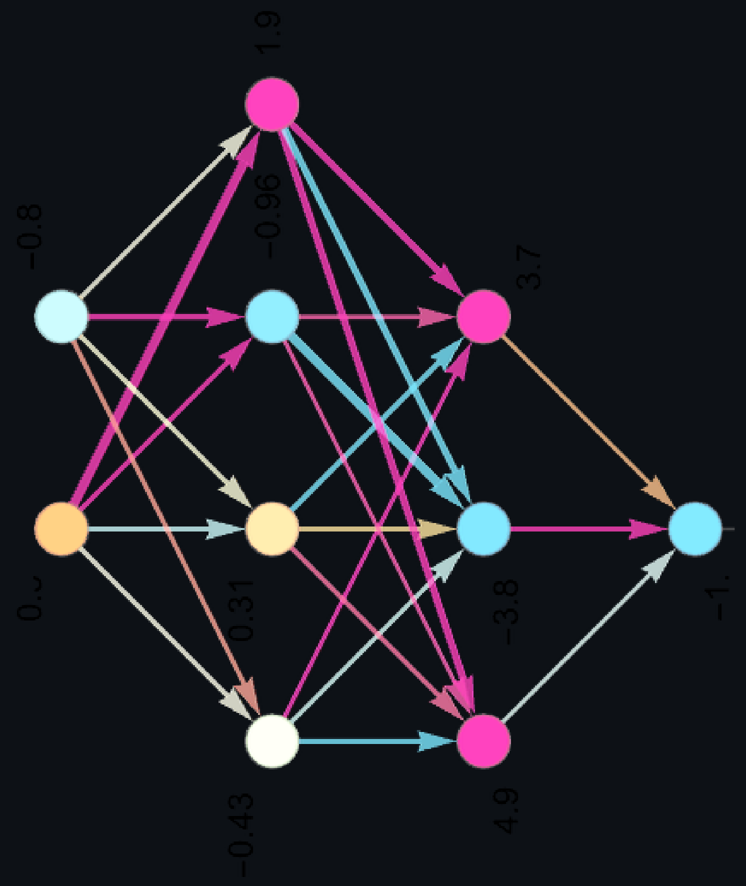
Weights and probabilities

Here's a plot that shows the probabilities of pairs of letters ("2-grams") in typical English text. The possible first letters are shown across the page, the second letters down the page

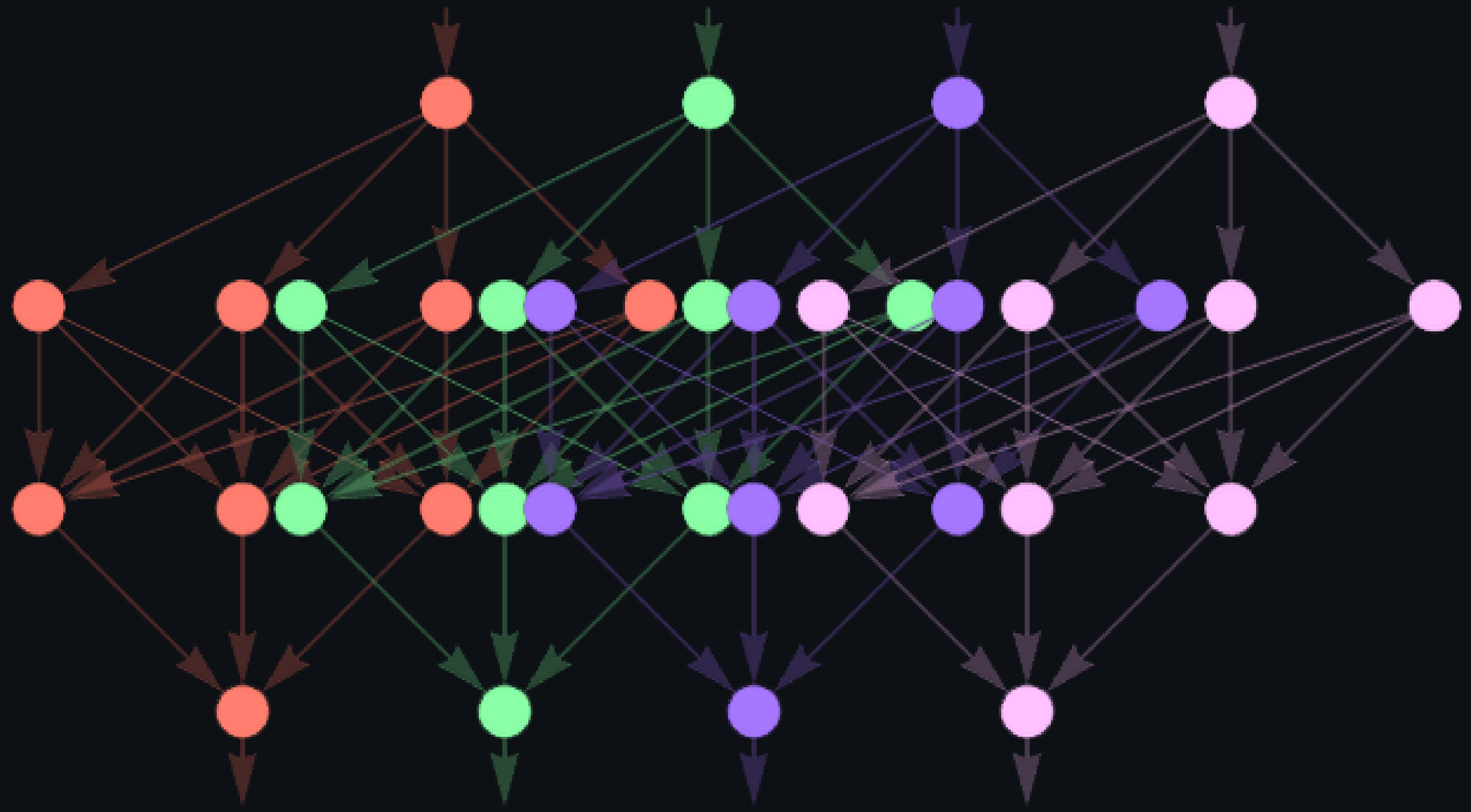




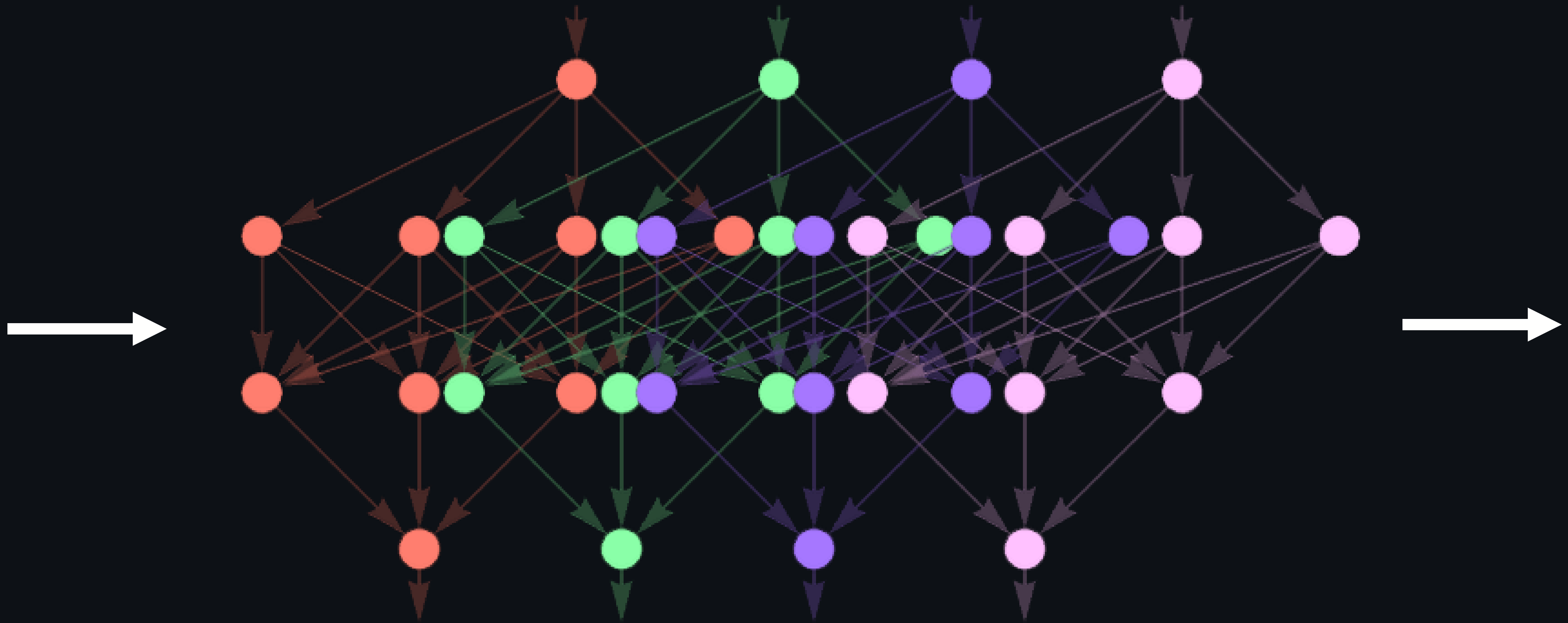
cat

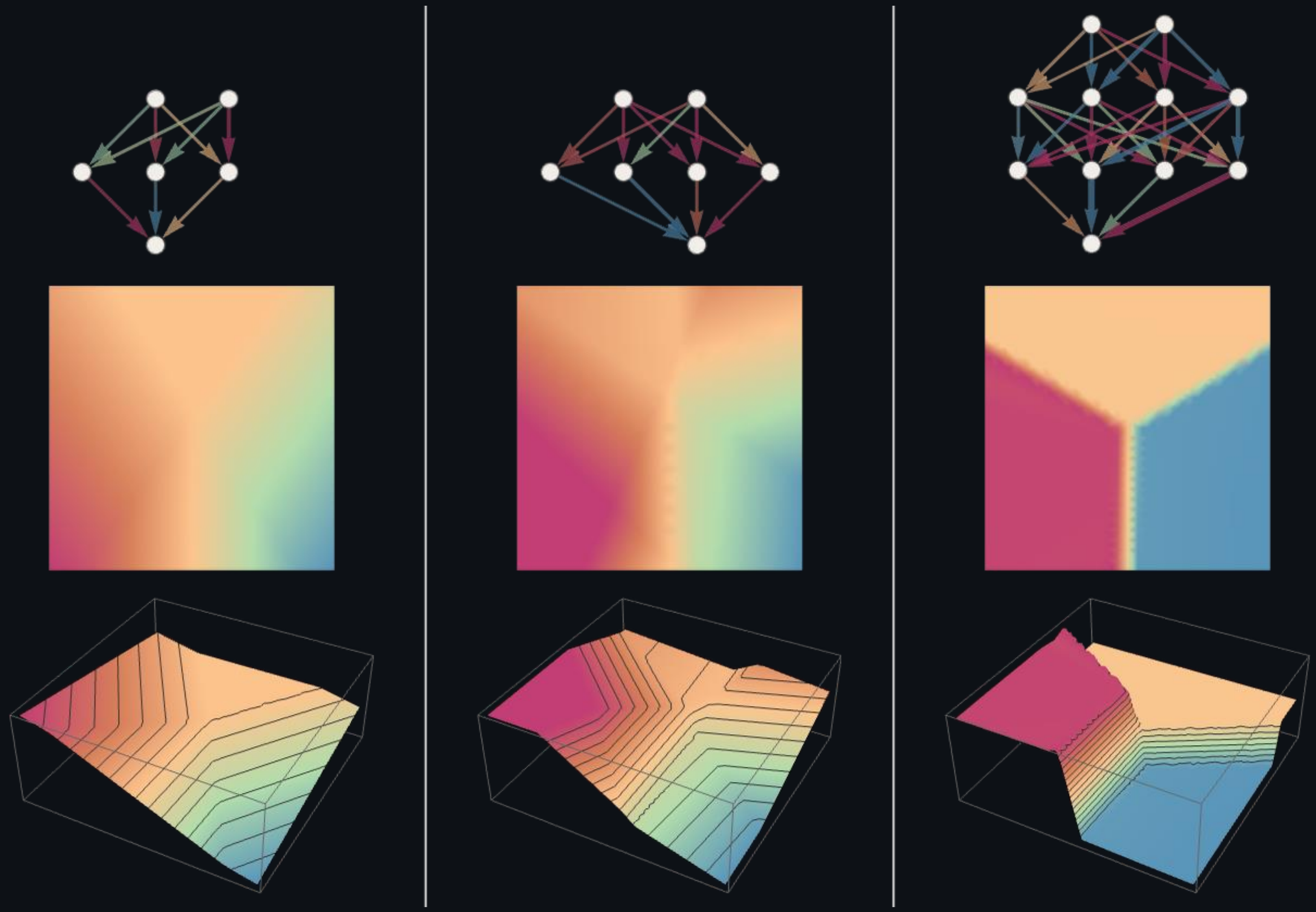


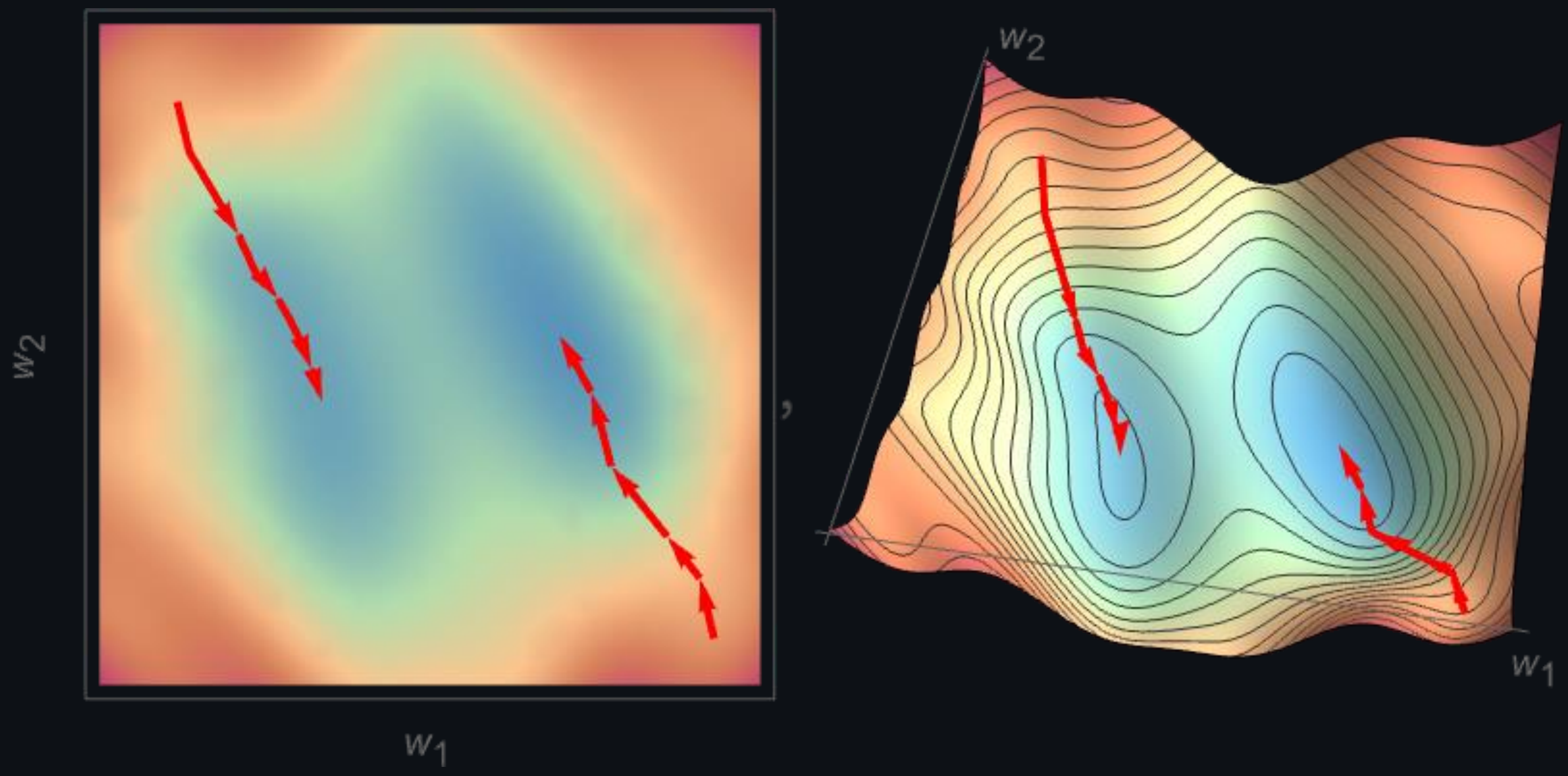
likes



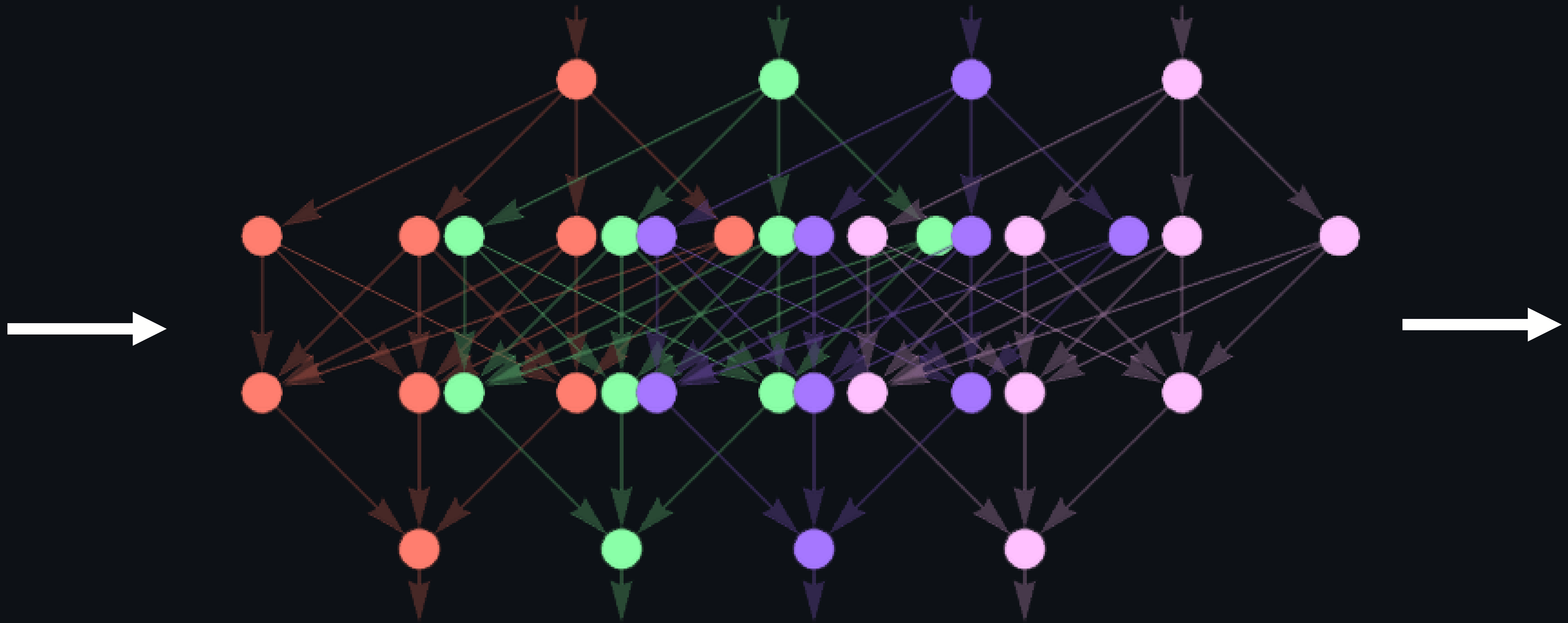
GPT3 has 400 (core) layers with millions of nodes and 175b connections

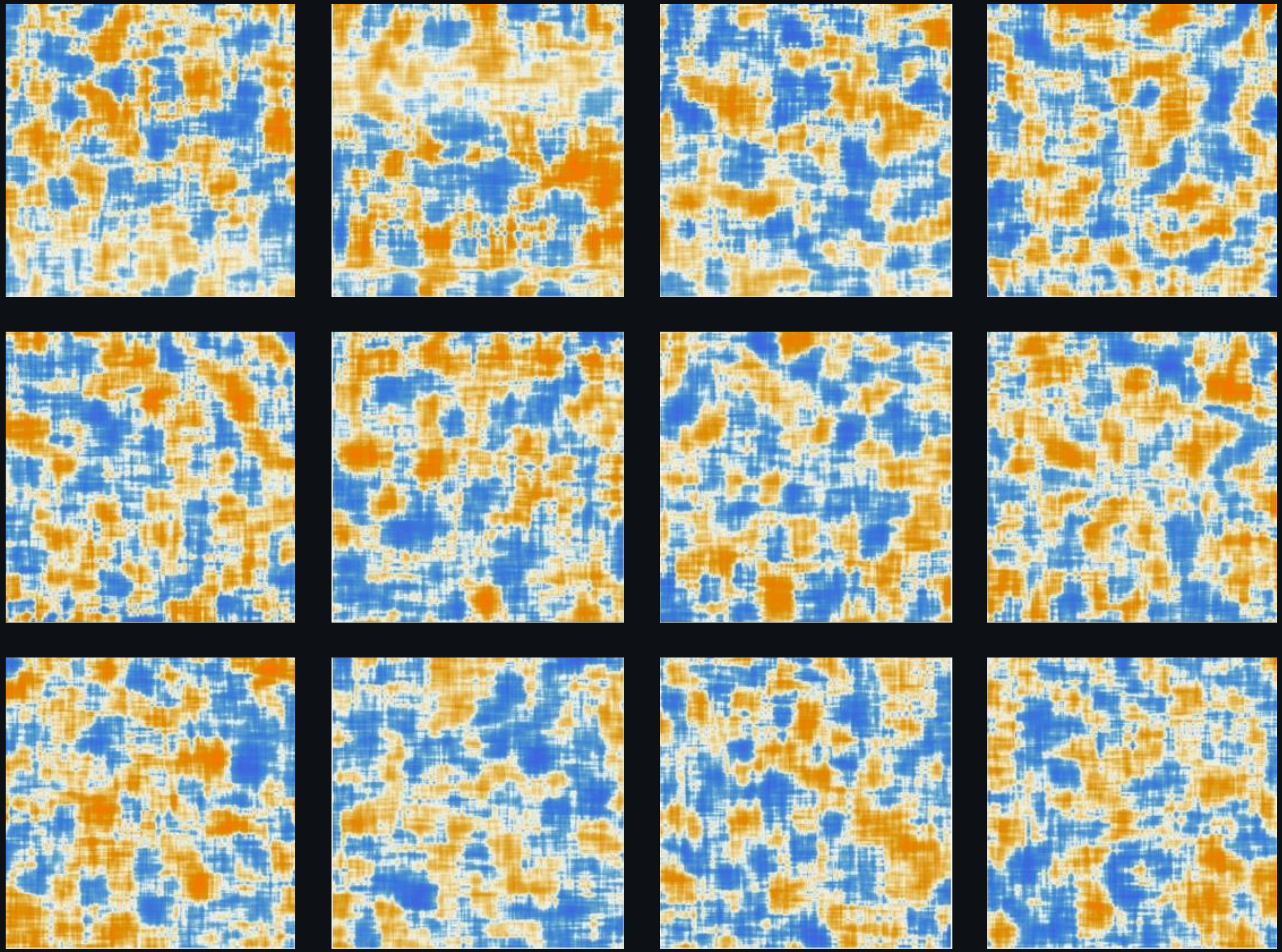






w2
w1
w2
w1





How was ChatGPT made?

Stage 1: Pretraining

1. Get about 45TB of text (few hundred billion words)
2. Get a cluster of 25,000 GPUs
3. Compress into a neural network
4. Pay $150m and wait 90 days
5. Obtain base model

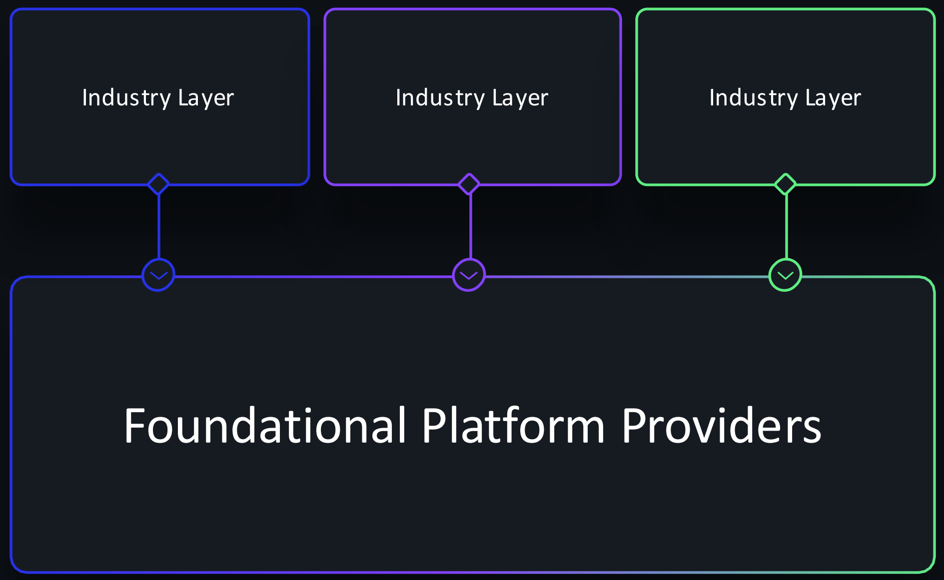
Stage 2: Fine Tune

1. Write labeling instructions
2. Hire thousands of people to collect 100k+ quality tags
3. Finetune the base model
4. Pay $300m for steps 1-3
5. Run lots of evaluations
6. Deploy
7. Profit

AI as an Opportunity

The AI boom is really similar to the mobile revolution or the cloud transformation cycle. You have a few, large core providers that then offer a platform to allow others to build on.

You've seen this playbook before…



How business are looking at AI

Enhancing Customer Experience

Using AI to create new product lines, enhancing existing ones, or create more competitive advantage.

Example Applications: chatbots, recommendation systems, and predictive analytics.

Operational Efficiency

Find ways to increase efficiency from large businesses processes to individual efficiency gains.

Large Example Applications: supply chain management, fraud detection, and maintenance schedules.

Individual Examples Applications: coding assistant, documentation writing, expense tracking.

Data-Drive Decision Making

AI can give new insights into large amounts of data making strategic decisions faster and easier to make.

Example Applications: market analysis, risk management, financial forecasting, trend finding

How business are looking at AI

INCREASE REVENUE

Enhancing Customer Experience

Using AI to create new product lines, enhancing existing ones, or create more competitive advantage.

Example Applications: chatbots, recommendation systems, and predictive analytics.

REDUCE COST

Operational Efficiency

Find ways to increase efficiency from large businesses processes to individual efficiency gains.

Large Example Applications: supply chain management, fraud detection, and maintenance schedules.

Individual Examples Applications: coding assistant, documentation writing, expense tracking.

GAIN INTELLIGENCE

Data-Drive Decision Making

AI can give new insights into large amounts of data making strategic decisions faster and easier to make.

Example Applications: market analysis, risk management, financial forecasting, trend finding

Things to know

Bias and Fairness
LLMs can inherit biases present in the training data, leading to biased or unfair outputs. This can perpetuate stereotypes and discrimination in applications like hiring, lending, and law enforcement.

Misinformation
LLMs will CONFIDENTLY lie. Make sure to fact check and not blindly copy and paste what it gives you.

Privacy and Data Security
Don't put anything in there you don't want to potentially be shared with others. LLMs can inadvertently expose sensitive information present in the training data.

Ethical and Legal Issues
Generating content can raise ethical questions, such as authorship and intellectual property rights.

Security Vulnerabilities
This is especially important if you're building on top of existing LLMs. They can be susceptible to adversarial attacks, where malicious inputs are designed to cause the model to produce harmful or misleading outputs.

Dependence on LLMs
Over-reliance on LLMs can lead to a lack of critical thinking and human oversight. Automated decision-making systems need to be complemented with human judgment to mitigate risks and ensure accountability.

Keep an eye on (things I find interesting in order)

- Agents
- Multi Model Systems
- Digital Twins
- Healthcare & Life Sciences

AI Agents

While there isn't a widely accepted definition for LLM-powered agents, they can be described as a system that can use an LLM to reason through a problem, create a plan to solve the problem, and execute the plan with the help of a set of tools.

In short, agents are a **system with complex reasoning capabilities, memory, and the means to execute tasks.**

Exciting Idea: Agent Swarms – orchestration of multiple AI agents to complete a complex task.

ChatDev: Communicative Agents for Software Development

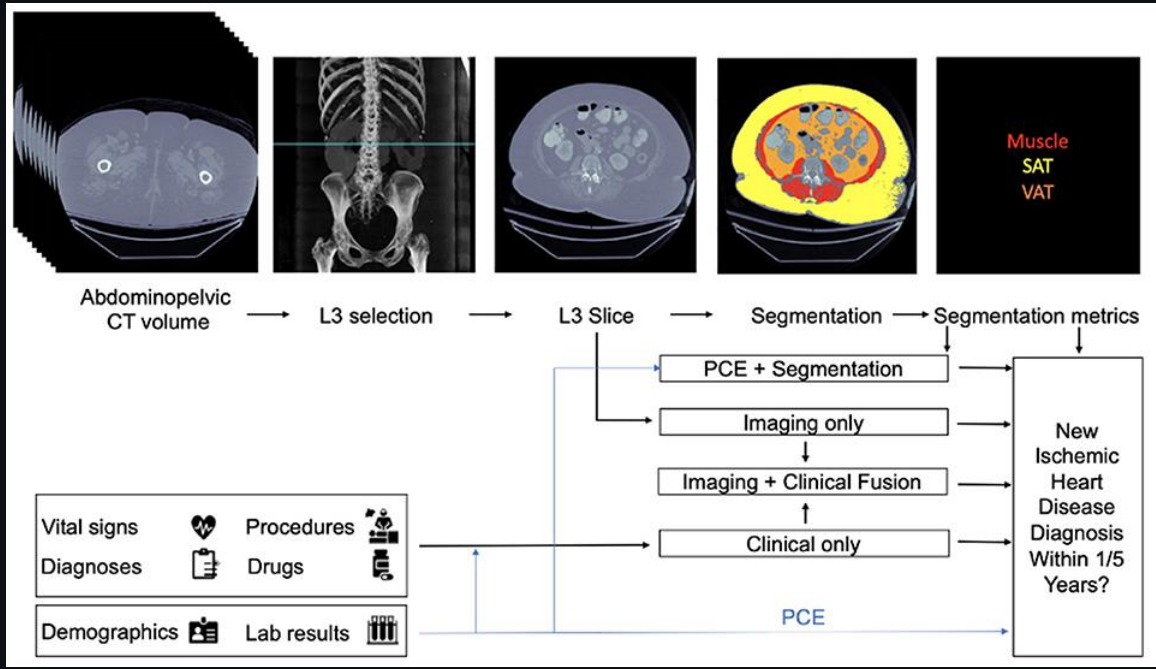
Multi Model

With regard to artificial intelligence, modality refers to data types. Data modalities include — but are not limited to — text, images, audio, and video.

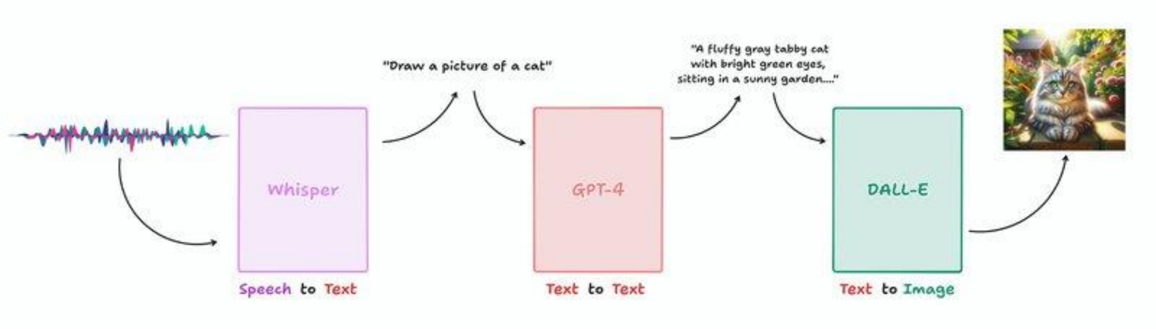
Multimodal AI is a systems that **can process multiple data inputs** to produce more accurate, sophisticated outputs than unimodal systems.

Why is it cool?

Enables the ability recognize patterns and connections between different types of data inputs, the output is more accurate, natural, intuitive, and informative.



Architecture of multi-modal data fusion combining Imaging and clinical data



Digital Twins

Not the metaverse.

A digital twin is a virtual replica of a physical object, person, or process that can be used to simulate its behavior to better understand how it works in real life.

Digital twins are linked to real data sources from the environment, which means that the twin updates in real time to reflect the original version.

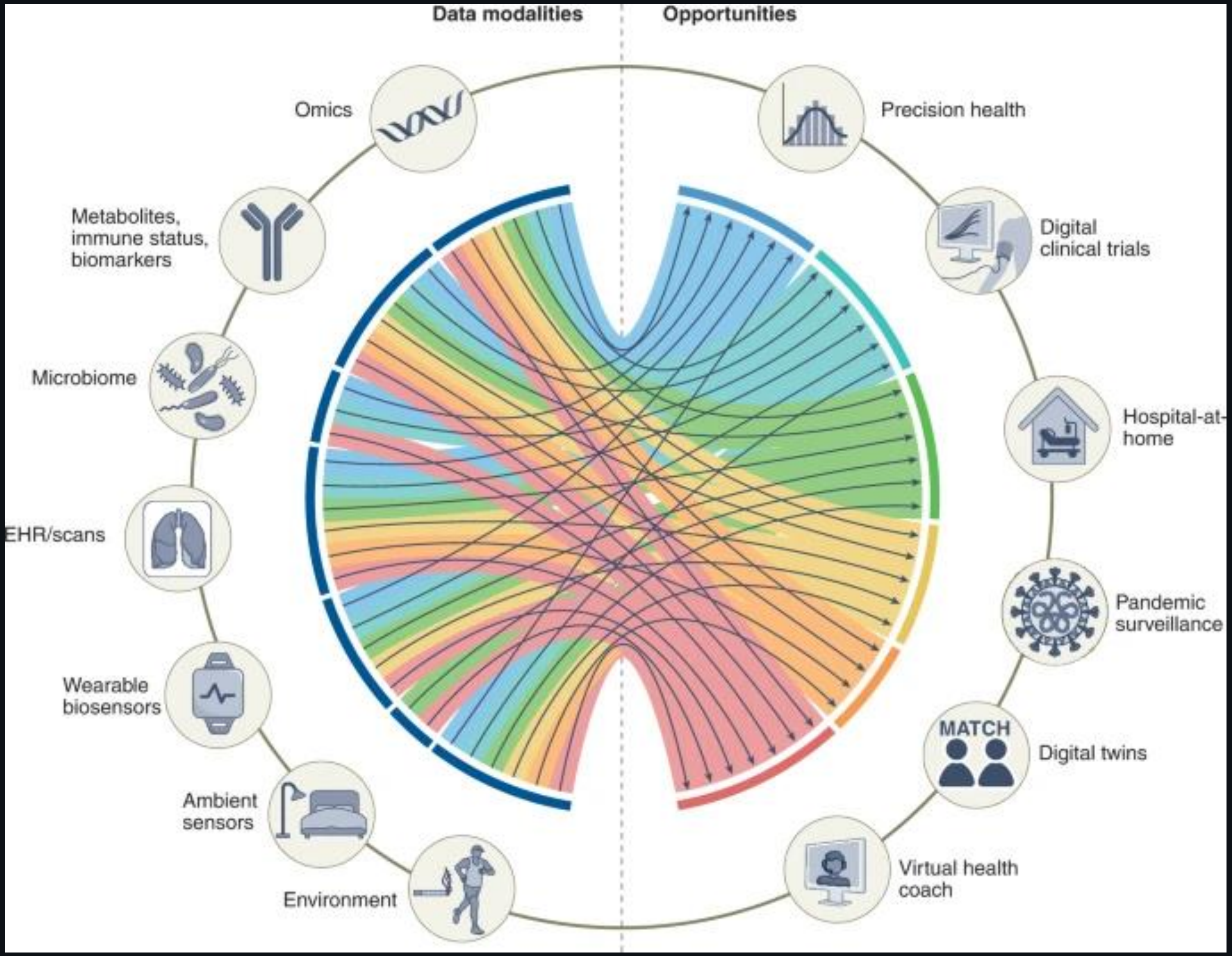
Why cool?
Lowers the barrier to research, innovation and has the opportunity to advance everything for climate change to healthcare.

Healthcare & Life Sciences

Improving humanities overall quality of life should always be the goal of any new technology.

AI has a significant opportunity to revolutionize healthcare by improving patient outcomes, enhancing operational efficiency, advancing medical research, and driving health equity.

In the life sciences, AI is changing the game for drug discovery, minimizing bias in clinical trials, and identifying diseases in earlier stages



Acosta, J.N., Falcone, G.J., Rajpurkar, P. et al. **Multimodal biomedical AI**. Nat Med 28, 1773–1784 (2022). https://doi.org/10.1038/s41591-022-01981-2

Thank you!