Introduction of Generative AI and Large Langue Models

Prof. Ching-Yung Lin

October 25th, 2024

Overview of Large Language Models

THE EVOLUTION OF NATURAL LANGUAGE PROCESSING

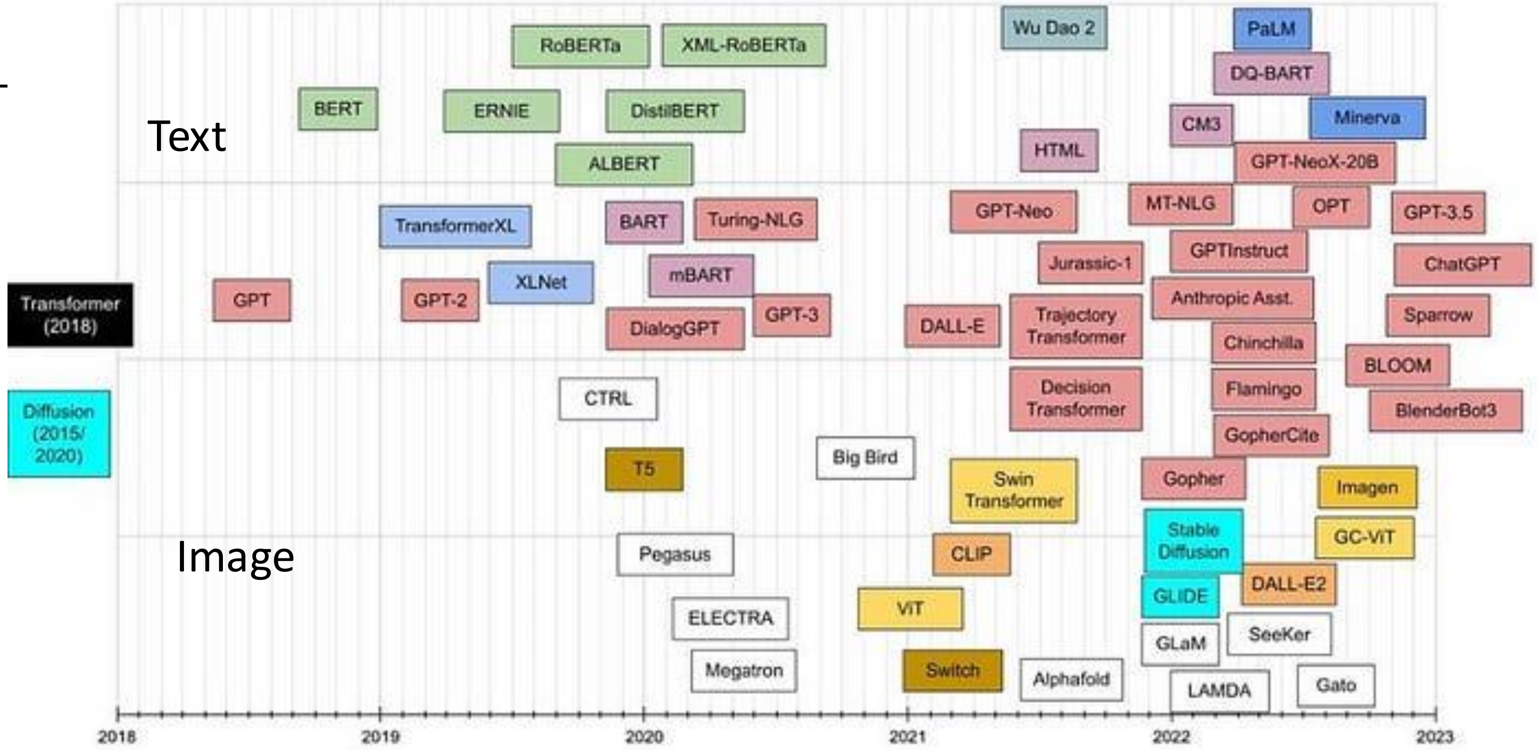
The Evolution of LLMs

1. In 2017, Google released the "Transformer Model", ~~which can be used in question-answering systems,~~ reading comprehension, sentiment analysis, instant translation of text or speech, and more

2. In 2018, OpenAI proposed "GPT" and Google proposed the "BERT" model, widely used in search engines, speech recognition, machine translation, question-answering systems, and more.

3. From 2018 to 2022, most of the research focused on BERT-related algorithms, when GPT performance was inferior to BERT

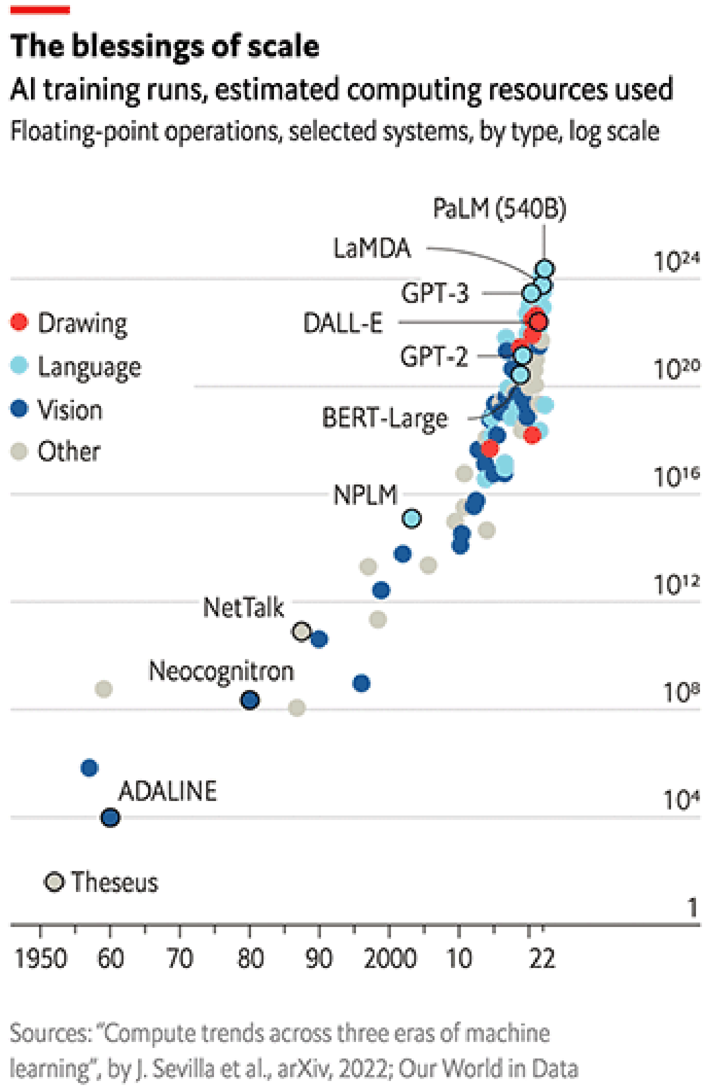
4. In 2023, ChatGPT (GPT3.5) was proposed by OpenAI, which significantly improves NLU's ability to understand most texts and surpasses humans in some area

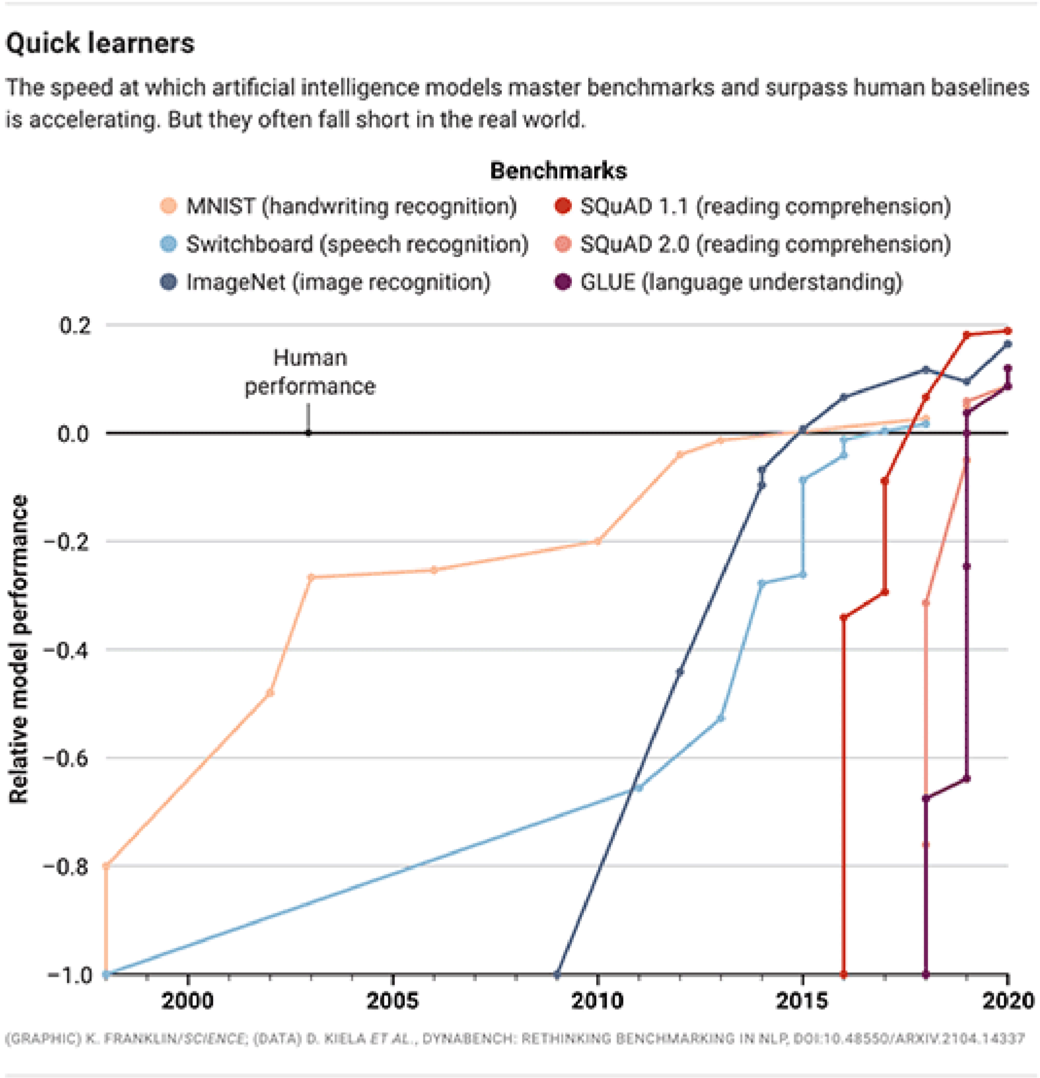


In NLU

CNN	RNN	Self-Attention
Local feature	Front and Back Dependency Issues	One to all attention, more flexible and trainable need large datasets

The speed of development of Generative AI





Generative AI Basics

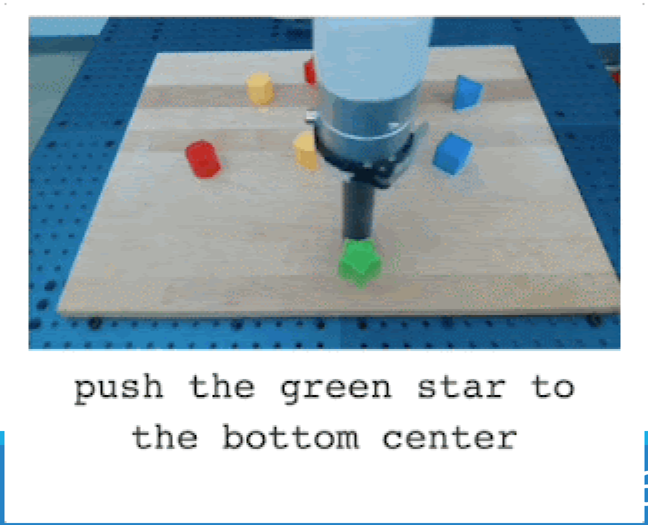
CREATING ARTIFICIAL CREATIVITY

Generative AI Application

Multi-Model

NLU + Image Generator

NLU + Robot



Condition Model

Generative Model

Conditional image Generator

Image Generator

Audio Generator

Speech Generator

Chat Bot

text Generator

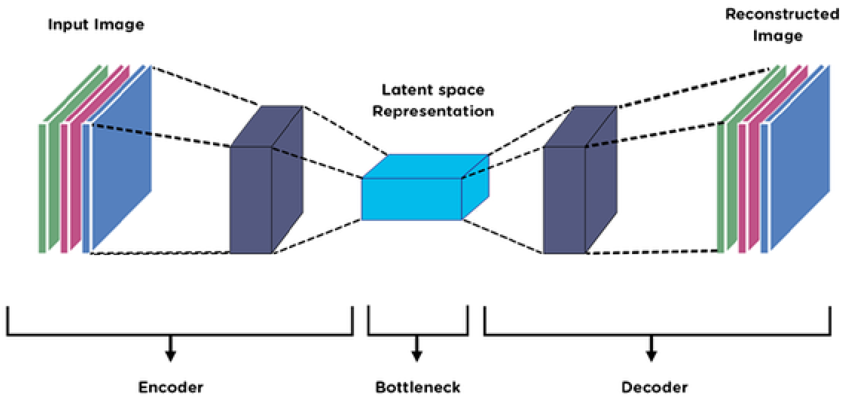
Summarization and Translation

Pose Generator

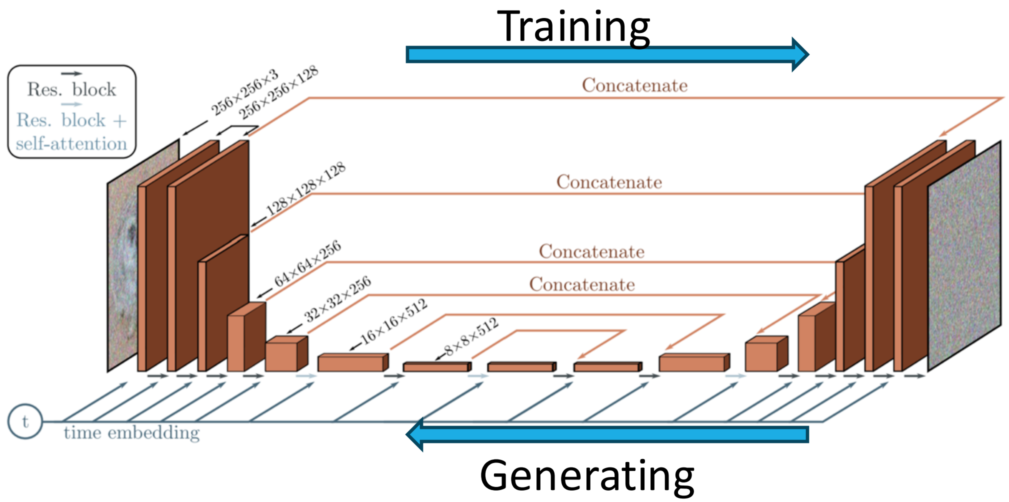
Robot

Generative AI Methodology

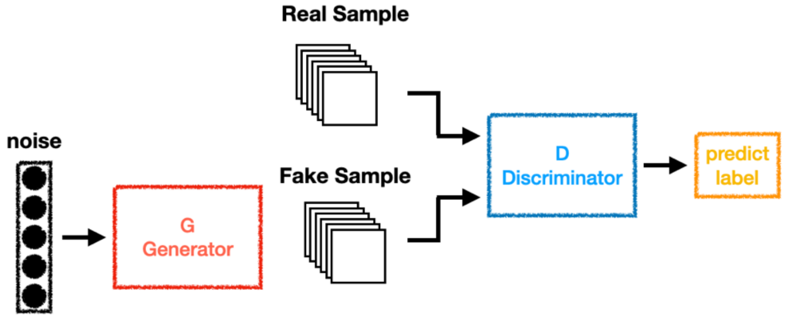
Variational AutoEncoder (VAE)



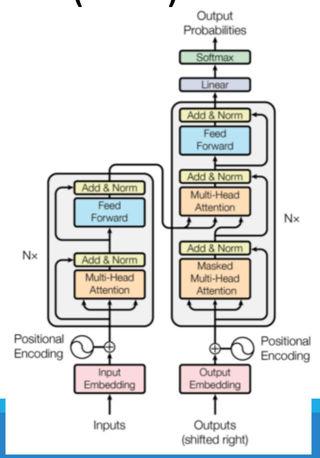
Diffusion Denoise Model



Generative Adversarial Network (GAN)



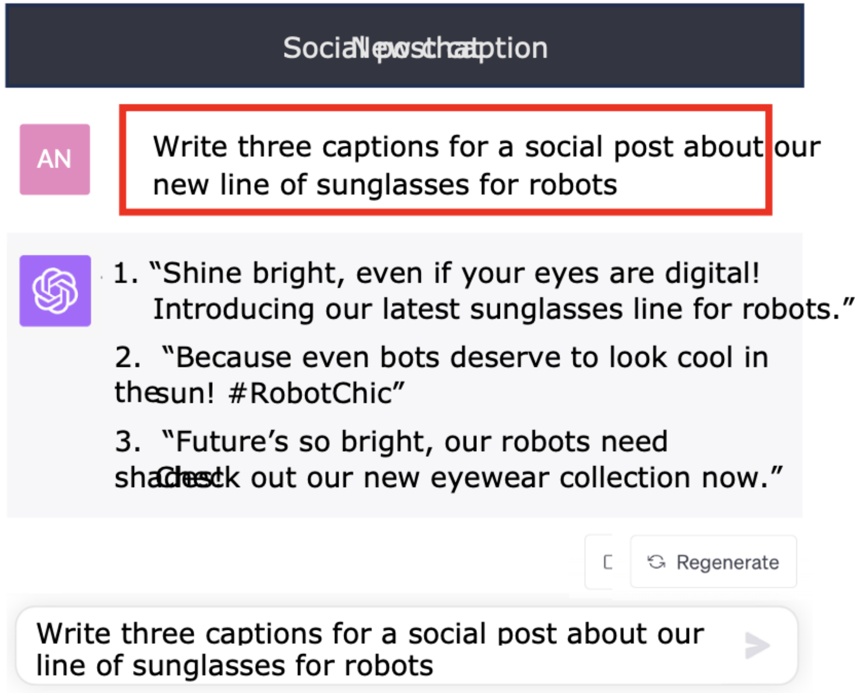
Large Language Model (LLM)



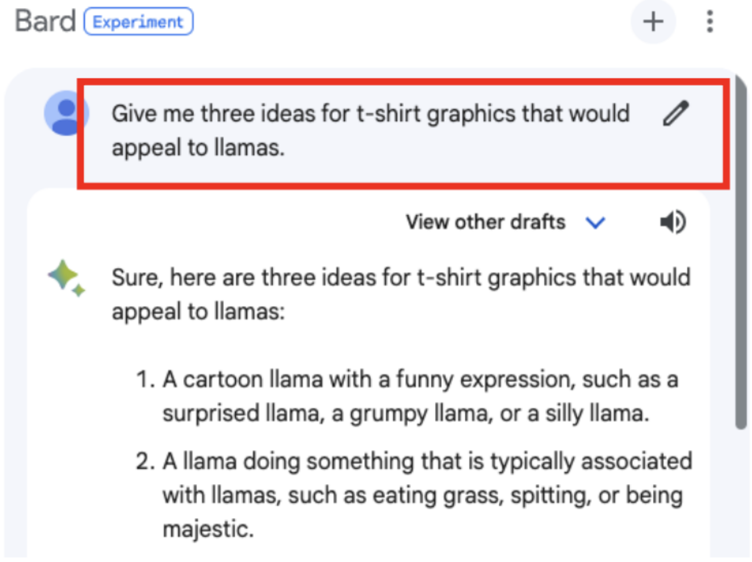
What is Generative AI

Artificial intelligence systems that can produce high quality content, specifically **text, images, and audio.**

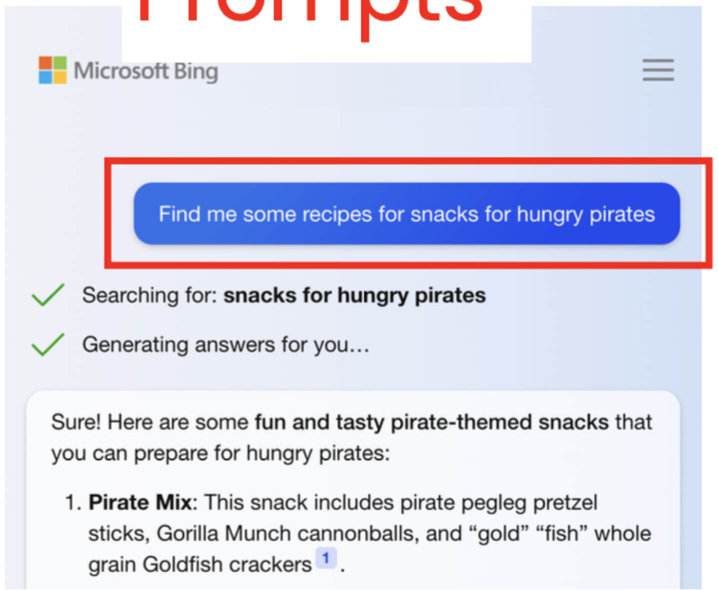
Prompts



ChatGPT/OpenAI



Bard/Google



Bing Chat/Microsoft

Multimedia Generation

A beautiful, pastoral mountain scene.
Landscape painting style (Midjourney)

Two cute kittens playing (DALL-E)

Generating Text using Large Language Models

Text generation process

I love eating ________

prompt

bagels with cream cheese

my mother's meatloaf

out with friends

AI output

How LLM works

LLMs are built by using supervised learning (A→B) to repeatedly predict the next word.

My favorite food is a bagel with cream cheese

Input (A)	Output (B)
My favorite food is a	bagel
My favorite food is a bagel	with
My favorite food is a bagel with	cream
My favorite food is a bagel with cream	cheese

When we train a very large AI system on a lot of data (hundreds of billions of words), we get a Large Language Model like ChatGPT.

A New Way to Find Information

What is the capital of South Africa?
South Africa has three capital cities: 1. Pretoria (administrative capital) 2. Cape Town (legislative capital) 3. Bloemfontein (judicial capital)

What does LLM stand for?
LL.M. stands for "Legum Magister" in Latin, which translates to "Master of Laws" in English.
What about in the context of AI?
In the context of AI, LLM refers to "Large Language Model".

Assist Writing

Rewrite this for clarity:

Students all around the world are realizing learning has to happen not just in school but throughout life including after graduation

Students worldwide understand that learning extends beyond school and continues throughout life.

Write a 300 word story involving trucks for a 3 year old to encourage them to brush their teeth

Once upon a time in a town called Toothville, there lived a happy tow truck named Timmy.

One day Timmy noticed that his shiny bumper had a dull spot. Worried, he went to the doctor [...]

Examples of tasks LLM can carry out

Writing

Suggest three names for our new analytics platform

Sure, here are three ideas:
1. NexaInsight
2. Quantivue
3. AnalytiSphere

web

Is there parking for staff?

Yes, employees can park on levels 1 and 2 of the offic
You can get a sticker at

app

Reading

I love my new llama T-shirt! The fabric is so soft.

Complaint: NO
Department: Apparel

app

I wore my llama T-shirt to a friend's wedding, and now they're mad at me for stealing the show

Complaint: YES
Department: Apparel

app

Chatting

Welcome to BettaBurgers!

I'd like a cheeseburger for delivery!

Awesome. Anything else?

That's it thanks!

Ok, it'll be there in 20 minutes. Enjoy!

app

LLM Hallucinations

Give me three quotes that Shakespeare wrote about Beyonce

1. Her vocals shine like the sun.
2. All hail the queen, she is most worthy of love.
3. Such a voice, so electric and alive, none else can compare!

✖ Hallucination

LLM Hallucinations

List two court cases tried in California about AI

1. *Waymo v. Uber* – Theft of trade secrets related to autonomous driving
2. *Ingersoll v. Chevron* – Patent infringement of oil field machine learning technology

✓ Real case

✗ Hallucination

Hallucinations have had serious consequences

The ChatGPT Lawyer Explains Himself

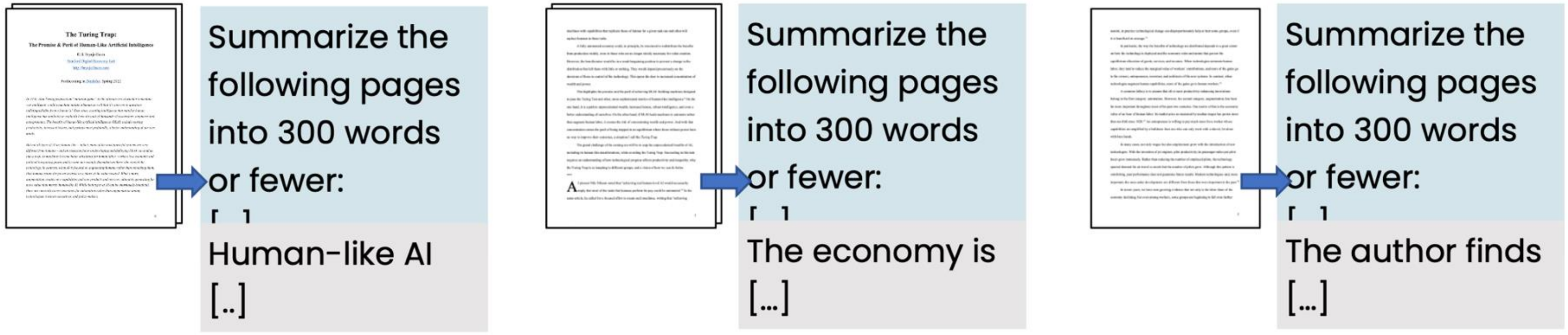
In a cringe-inducing court hearing, a lawyer who relied on A.I. to craft a motion full of made-up case law said he "did not comprehend" that the chat bot could lead him astray.

The New York Times

Input / Output Length is Limited

Many LLMs can accept a prompt of up to only a few thousand words.

- The total amount of context you can give it is limited
- Some LLMs have longer context limits – up to 100,000 words
- An LLM's context length is the limit on the total input+output size



Not Understanding Structured Data

Home prices

size (square feet)	price (1000$)
523	100
645	150
708	200
1034	300
2290	350
2545	440
A	B

Use supervised learning (A → B)

Purchases on website

user ID	time	price ($)	purchased
4783	Jan 21 08:15.20	7.95	yes
3893	March 3 11:13:.5	10.00	yes
8384	June 11 14:15.05	9.50	no
0931	Aug 2 20:30.55	12.90	yes

A B

Bias and Toxicity

An LLM can reflect the biases that exist in the text it learned from.

Complete this sentence:

The surgeon walked to the parking lot and took out

his car keys.

assumed male

Complete this sentence:

The nurse walked to the parking lot and took out

her phone.

assumed female

Some LLMs can output toxic or other harmful speech, but most models have gotten much safer over time.

Knowledge Cutoffs

An LLM's knowledge of the world is frozen at the time of its training

- A model trained on data scraped from internet in January 2022 has no information about more recent events

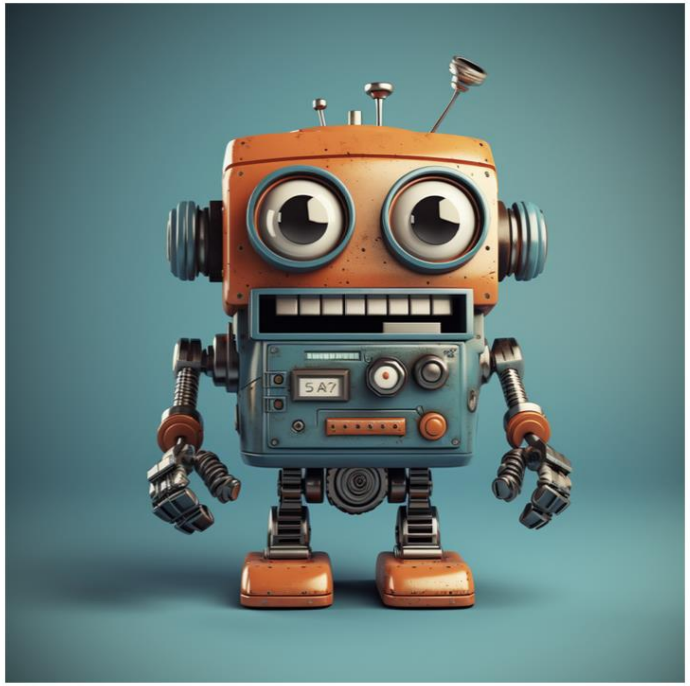
What was the highest grossing film of 2022?
As of January 2022, I don't have data on the highest-grossing movie for that year. ✖

Avatar: The Way of Water

Examples of Generated Images

A picture of a woman smiling

A futuristic city scene



A cool, happy robot

Image Generation

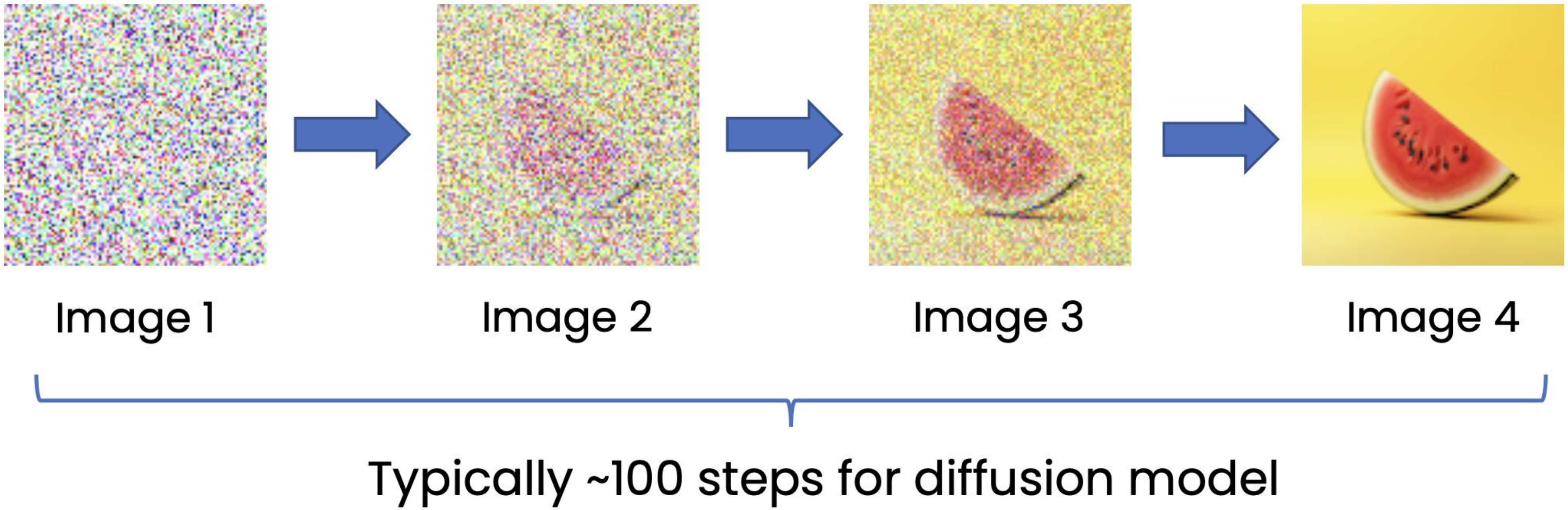
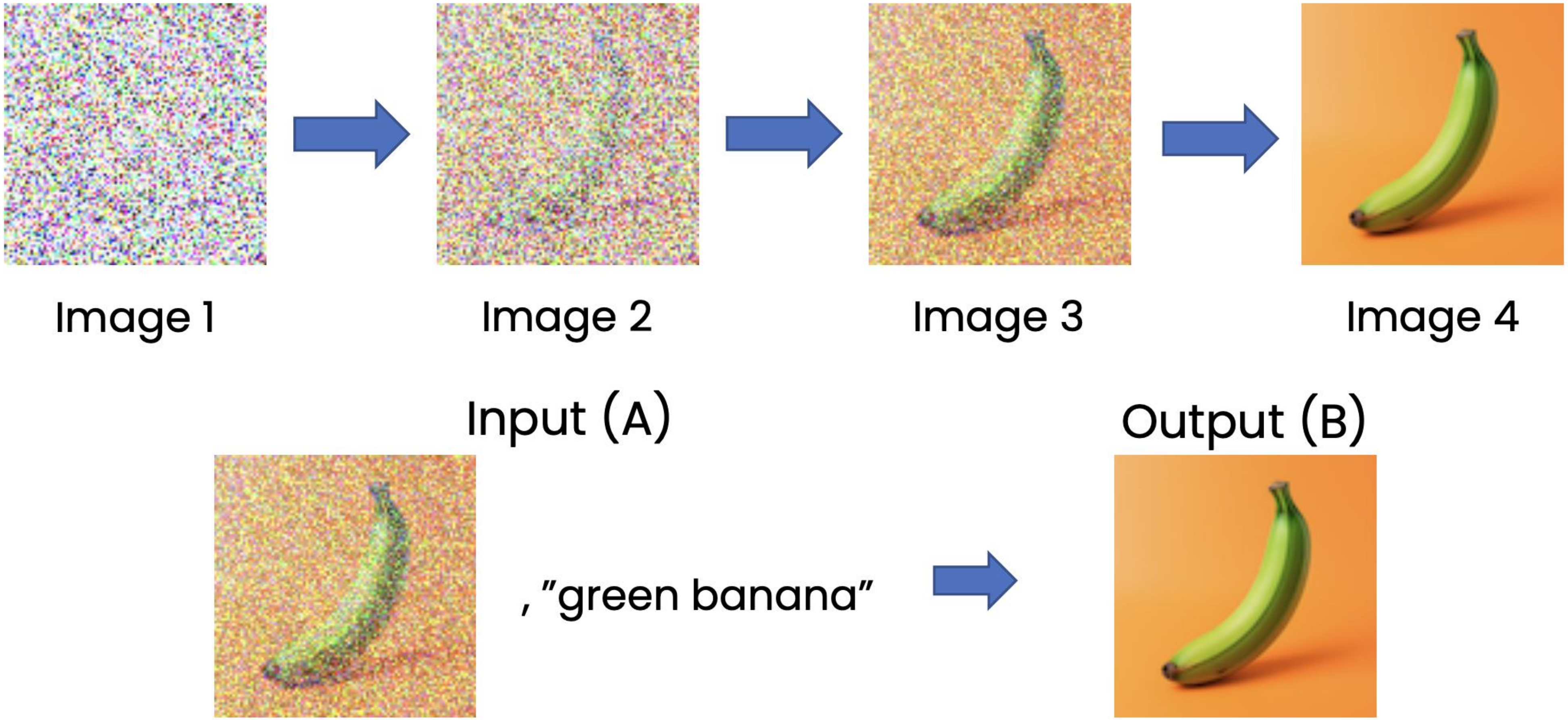


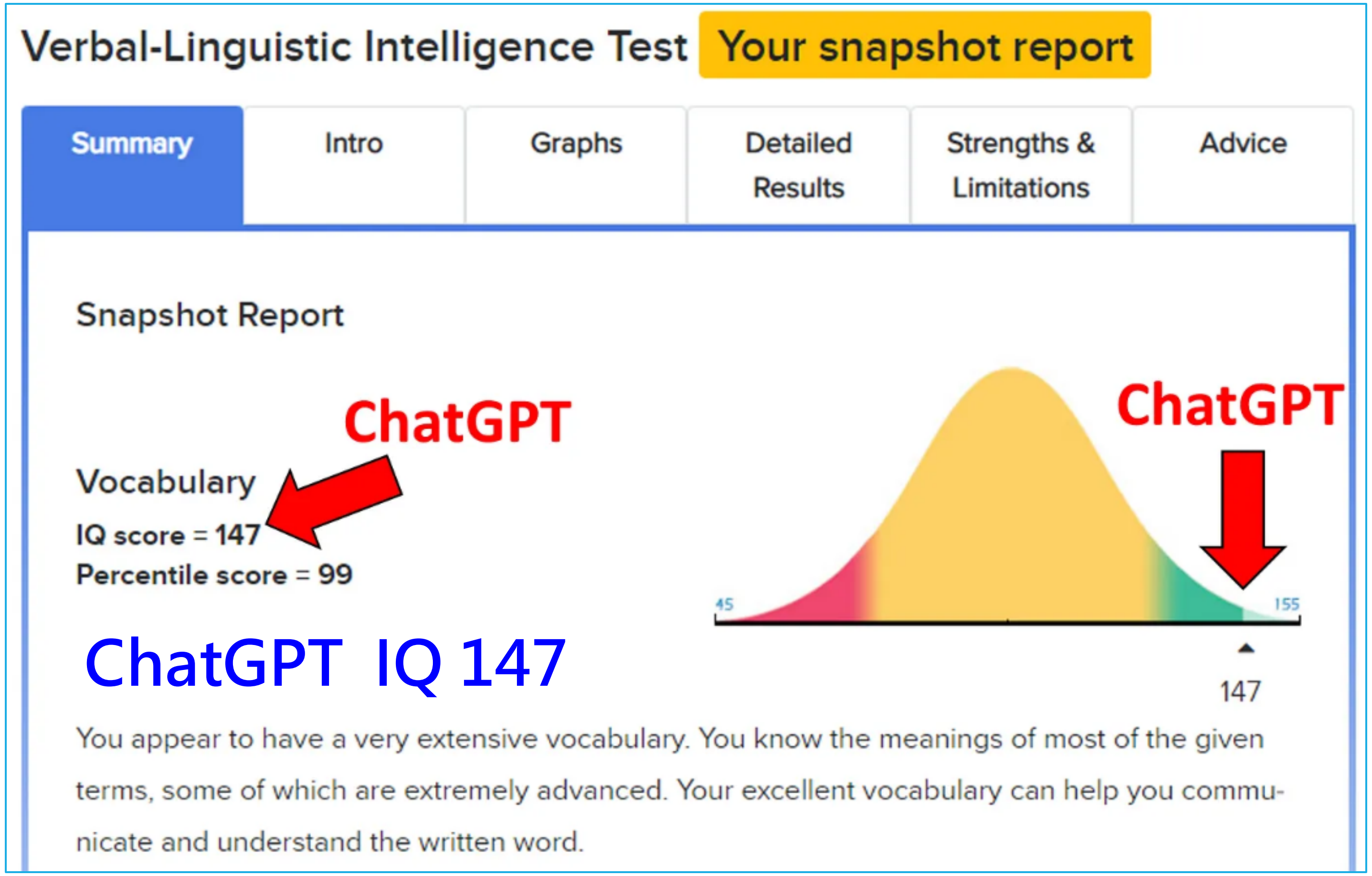
Image generation from Text



Key Technics behind Large Language Models and Generative AI

KEY CONCEPTS

ChatGPT



https://lifearchitect.ai/chatgpt/

ChatGPT

Software dev job	**ChatGPT would be hired as L3 Software Developer at Google: the role pays $183,000/year.**
Politics	**ChatGPT writes several Bills (USA).**
MBA	**ChatGPT would pass an MBA degree exam at Wharton (UPenn).**
Accounting	**GPT-3.5 would pass the US CPA exam.**
Legal	**GPT-3.5 would pass the bar in the US.**
Medical	**ChatGPT would pass the United States Medical Licensing Exam (USMLE).**
AWS certificate	**ChatGPT would pass the AWS Certified Cloud Practitioner exam.**
IQ (verbal only)	**ChatGPT scores IQ=147, 99.9th %ile.**
SAT exam	**ChatGPT scores 1020/1600 on SAT exam.**

https://lifearchitect.ai/chatgpt/

Attention Experiment

Ulric Neisser Attention Experiment

https://www.youtube.com/watch?v=vJG698U2Mvo&ab_channel=DanielSimons

Attention Model
[Bengio_2015]

Attention-Based Models for Speech Recognition

Jan Chorowski
University of Wrocław, Poland
jan.chorowski@ii.uni.wroc.pl

Dzmitry Bahdanau
Jacobs University Bremen, Germany

Dmitriy Serdyuk
Université de Montréal

Kyunghyun Cho
Université de Montréal

Yoshua Bengio
Université de Montréal
CIFAR Senior Fellow

Abstract

Recurrent sequence generators conditioned on input data through an attention mechanism have recently shown very good performance on a range of tasks including machine translation, handwriting synthesis [1, 2] and image caption generation [3]. We extend the attention-mechanism with features needed for speech recognition. We show that while an adaptation of the model used for machine translation in [2] reaches a competitive 18.7% phoneme error rate (PER) on the TIMIT phoneme recognition task, it can only be applied to utterances which are roughly as long as the ones it was trained on. We offer a qualitative explanation of this failure and propose a novel and generic method of adding location-awareness to the attention mechanism to alleviate this issue. The new method yields a model that is robust to long inputs and achieves 18% PER in single utterances and 20% in 10-times longer (repeated) utterances. Finally, we propose a change to the attention mechanism that prevents it from concentrating too much on single frames, which further reduces PER to 17.6% level.

2015, Bengio's Model focuses on every phenome's recogniztion is the combined weights.

$$\alpha_i = Attend(s_{i-1}, \alpha_{i-1}, h)$$

$$g_i = \sum_{j=1}^{L} \alpha_{i,j} h_j$$

$$y_i \sim Generate(s_{i-1}, g_i),$$

h : Input

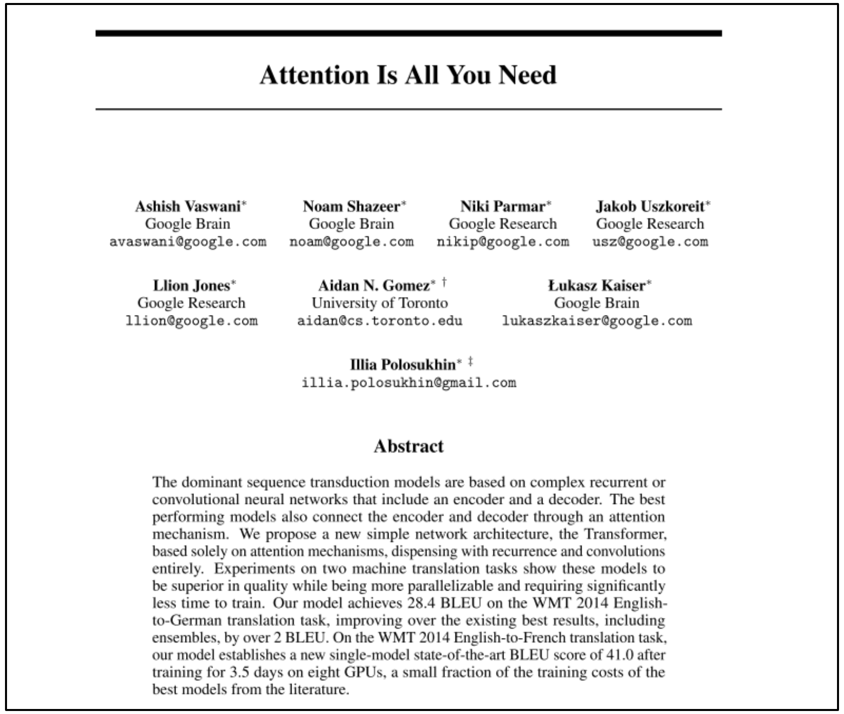
α_i: Attention Weight

y_i: Output

Chorowski, Jan K., et al. "Attention-based models for speech recognition." *Advances in neural information processing systems* 28 (2015).

Transformer [Vaswani_2017]

In 2017, 8 Google researchers proposed Transformer Neuron Networks based on Attention, which was adopted by ChatGPT.



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.

(2023/2/21)

Jakob Uszkoreit proposed replacing RNNs with **self-attention** and started the effort to evaluate this idea.

Noam Shazeer proposed **scaled dot-product attention**, **multi-head attention** and the **parameter-free position representation**.

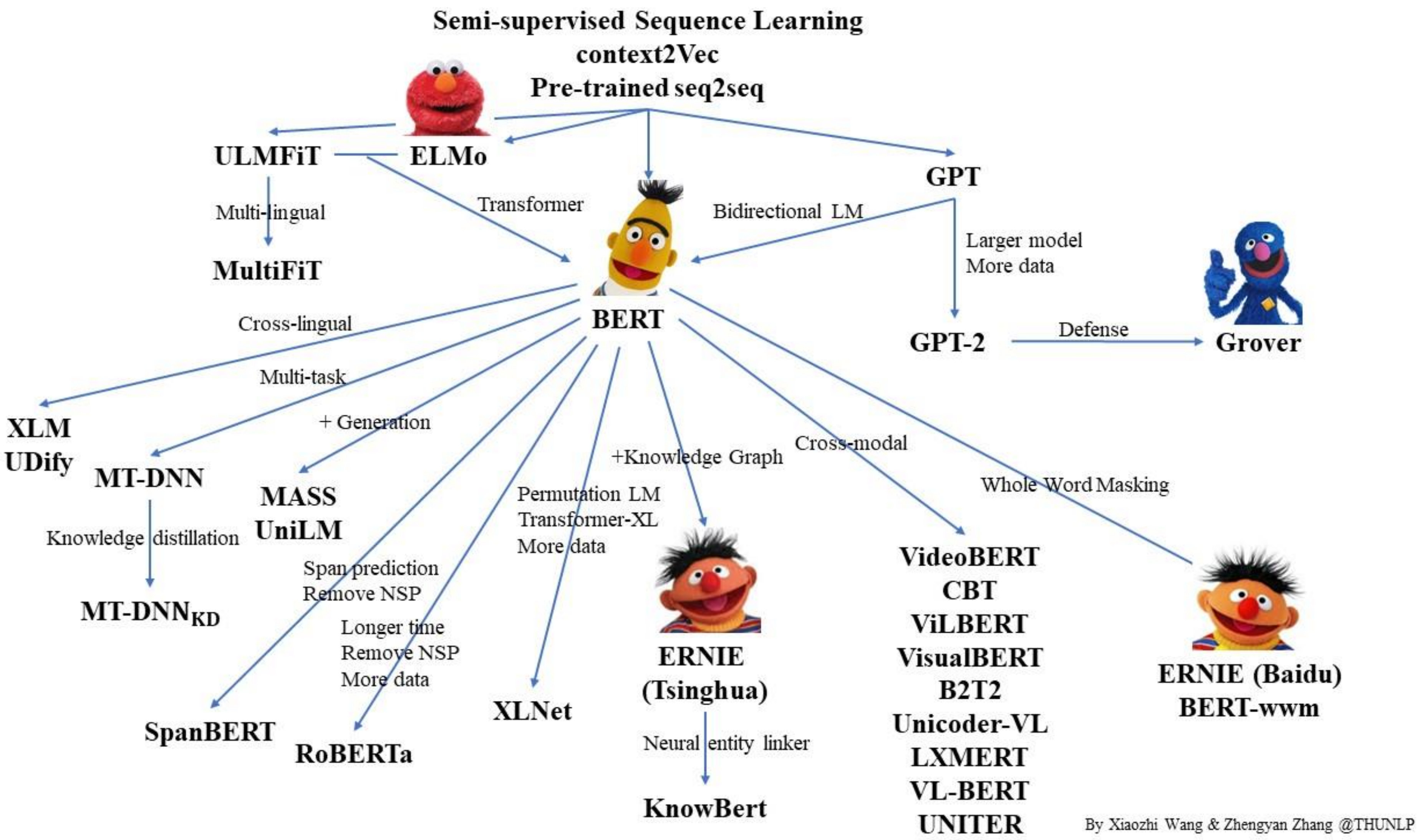
Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Transformer

- Transformer is a Deep Learning Model based on Self-Attention
- **Transformer** encodes and decodes data with different weights.
- Examples of **transformer language models include: GPT (GPT-1、GPT-2、GPT-3、ChatGPT) and BERT models (BERT、RoBERTa 、ERNIE).**

Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

BERT AI Models

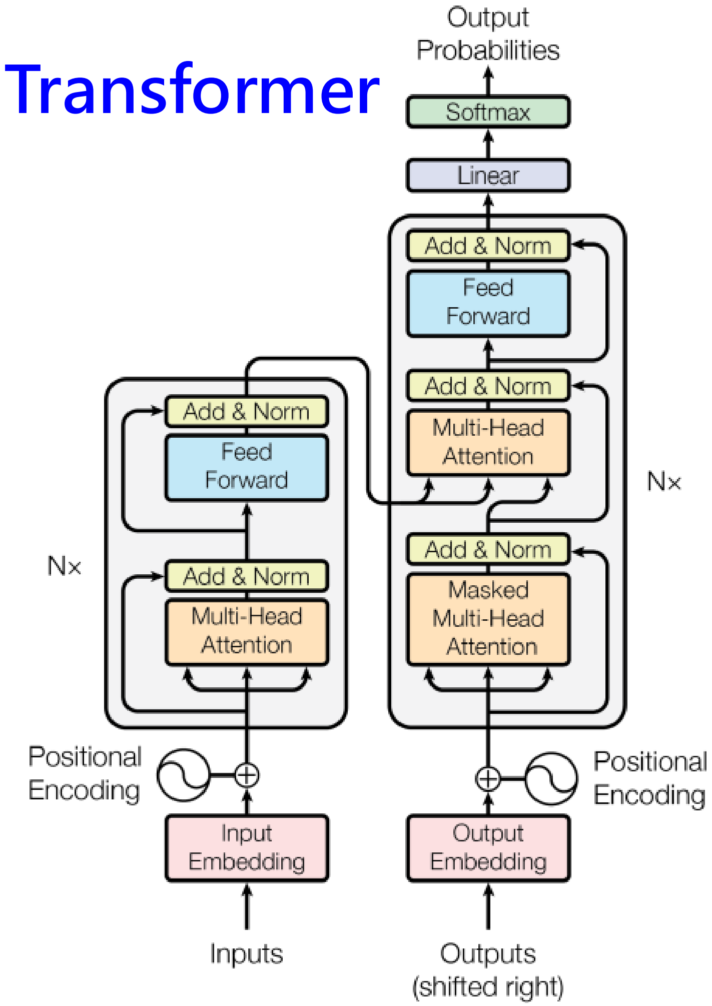


http://pelhans.com/2020/02/02/pretraining_model/

Transformer

Encoder

Transformer



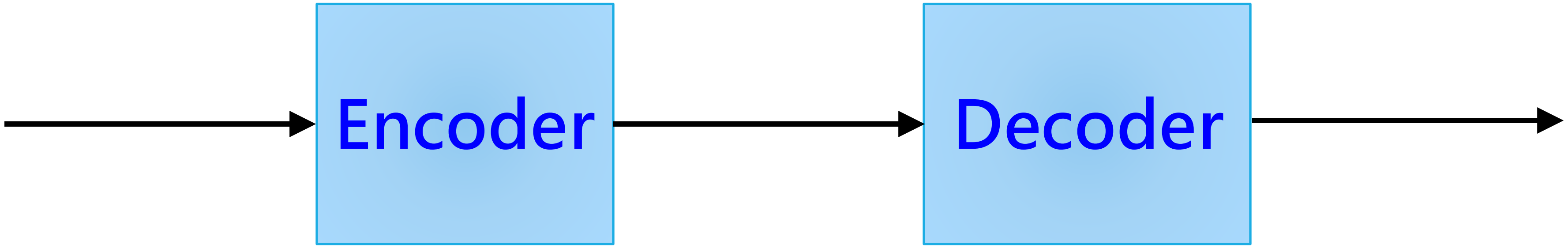
Decoder

Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Transformer

哥大學生很棒!

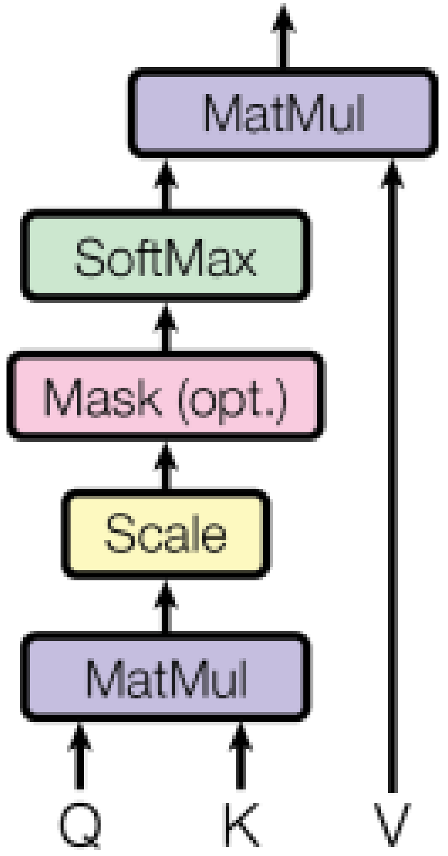
Columbia University students are great!



Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Transformer
Attention

Scaled Dot-Product Attention



$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

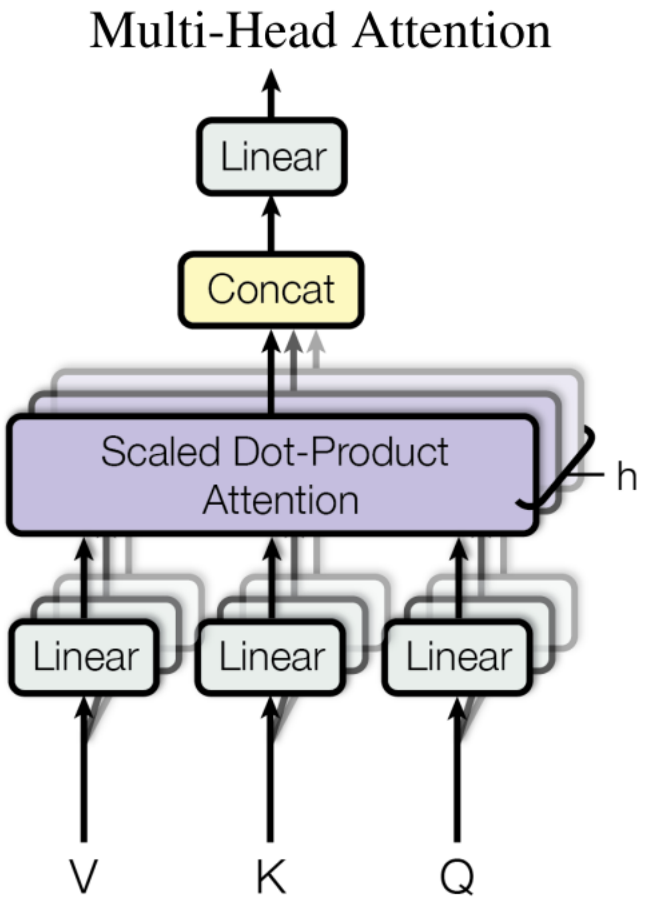
Transformer
Attention

K

Q

		k_1	k_2	k_3	k_4	k_5	k_6
	weights	Columbia	university	students	are	great	!
q_1	哥	1	0.5	0.2	0	0.3	0.2
q_2	大	0.5	1	0.2	0.1	0.3	0.1
q_3	學	0.2	0.2	1	0	0.5	0.2
q_4	生	0.3	0.3	0.8	0.5	0.5	0.6
q_5	很	0	0.1	0	1	0.5	0
q_6	棒	0.3	0.3	0.5	0.5	1	0.8
q_7	!	0.2	0.1	0.2	0	0.8	1

Transformer multi-head attention



Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Transformer Translation

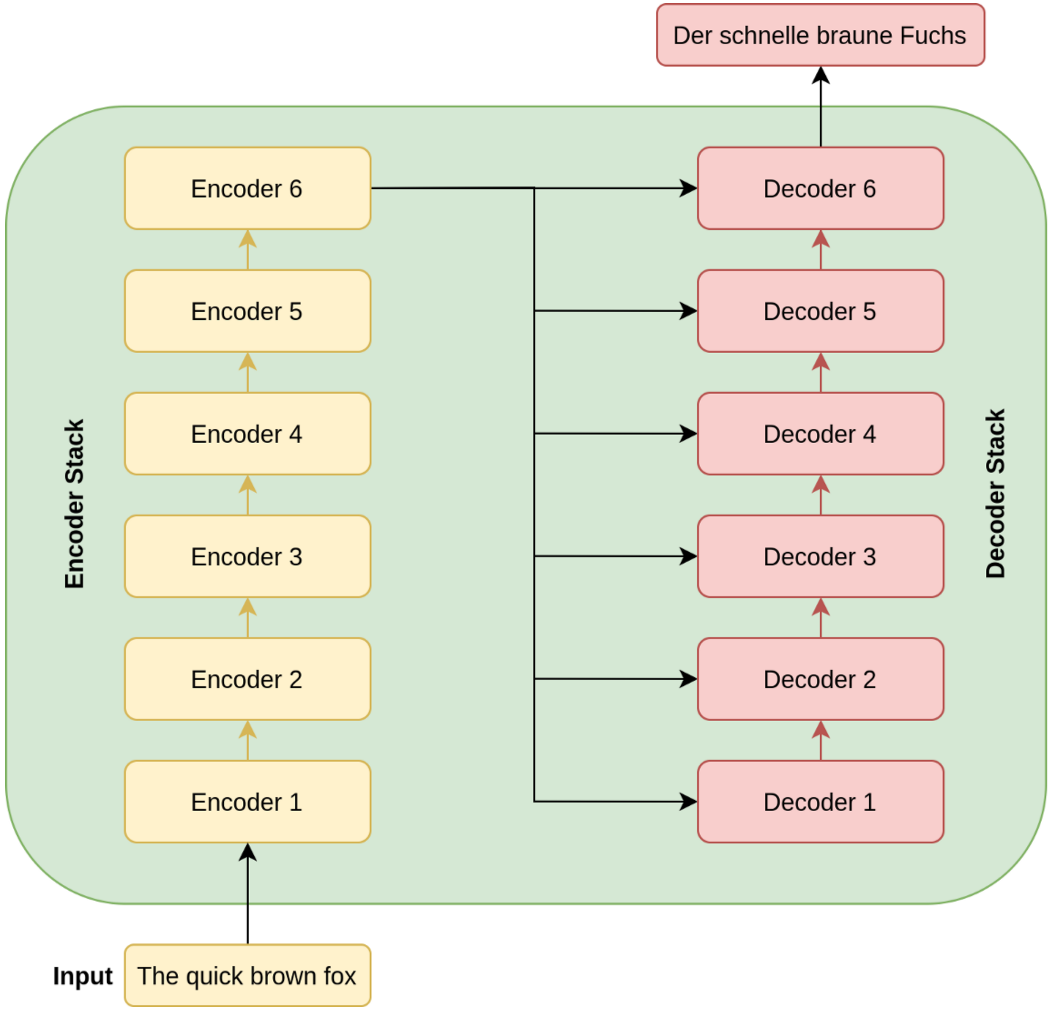
Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [15]	23.75			
Deep-Att + PosUnk [32]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [31]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [8]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [26]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [32]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [31]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [8]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	$\mathbf{3.3 \cdot 10^{18}}$	
Transformer (big)	**28.4**	**41.0**	$2.3 \cdot 10^{19}$	

Bilingual Evaluation Understudy Score，BLEU is an evaluation to see how close the translation is to real human being.

Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Transformer Translation



Transformer uses 6 layers of encoder and decoder to achieve the same quality of SOTA English-German and English-French translation.

BERT Introduction

- 2018 Google'BERT has 24 layers of Transformer Encoder
- BERT's original model is based on Wikipedia and booksorpus, using unsupervised training to create BERT.
- At Stanford's Machine Reasoning Test SQuAD1.1 beats human performance.
- Google NLU English was replaced from seq2seq to BERT

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

BERT Introduction

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Abstract

We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning *all* pretrained parameters. The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

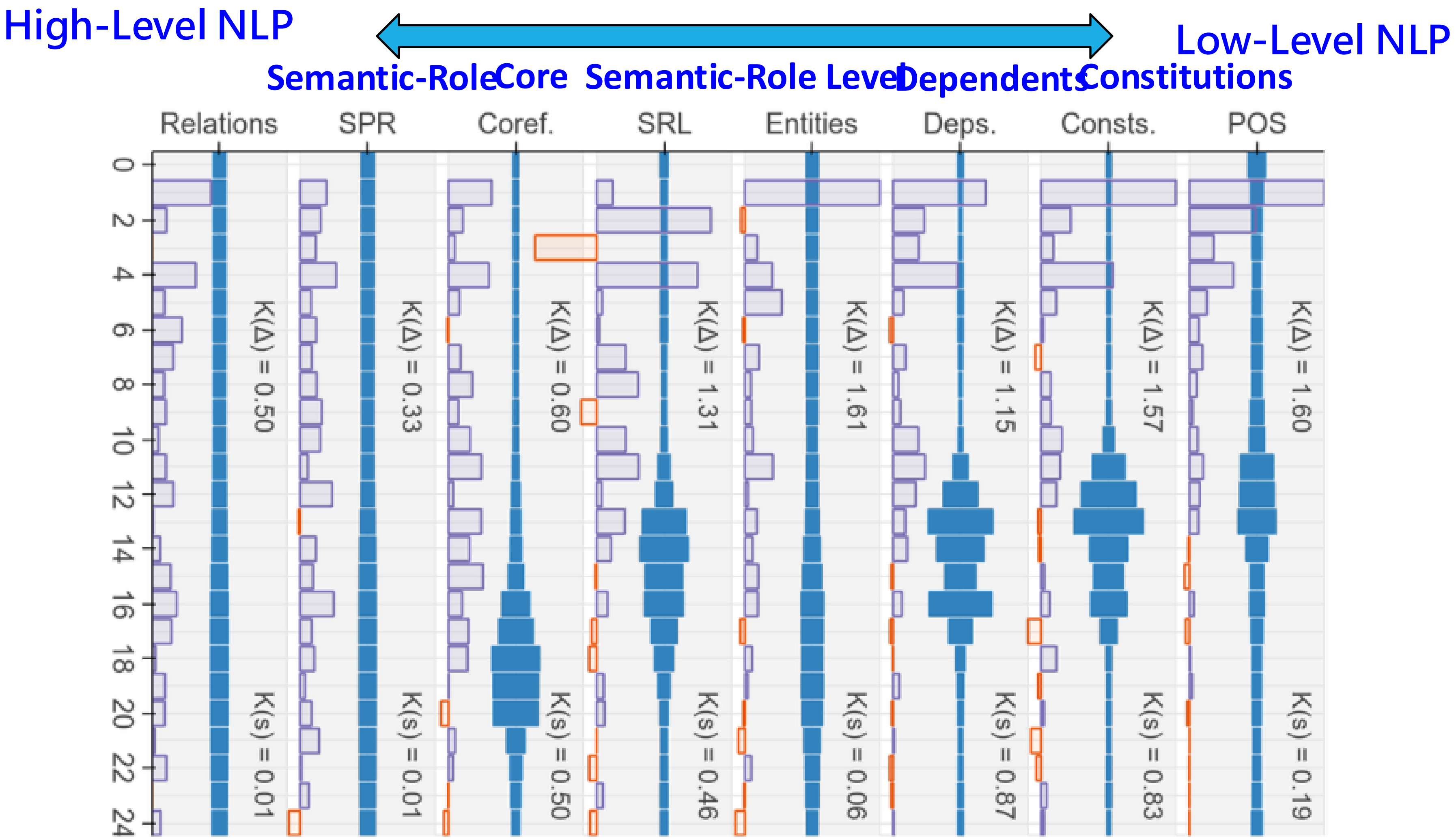
In 2018's BERT Comprehension test outperformed human

SQuAD1.1 Leaderboard

Since the release of SQuAD1.0, the community has made rapid progress, with the best models now rivaling human performance on the task. Here are the ExactMatch (EM) and F1 scores evaluated on the test set of v1.1.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google A.I.*	**87.433**	**93.160**
2 Oct 05, 2018	BERT (single model) *Google A.I.*	85.083	91.835
2 Sep 09, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.356	91.202
2 Sep 26, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.954	91.677

BERT understands language's meaning



Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the clasical NLP pipeline. *arXiv preprint arXiv:1905.05950*.

Attention to Transformer

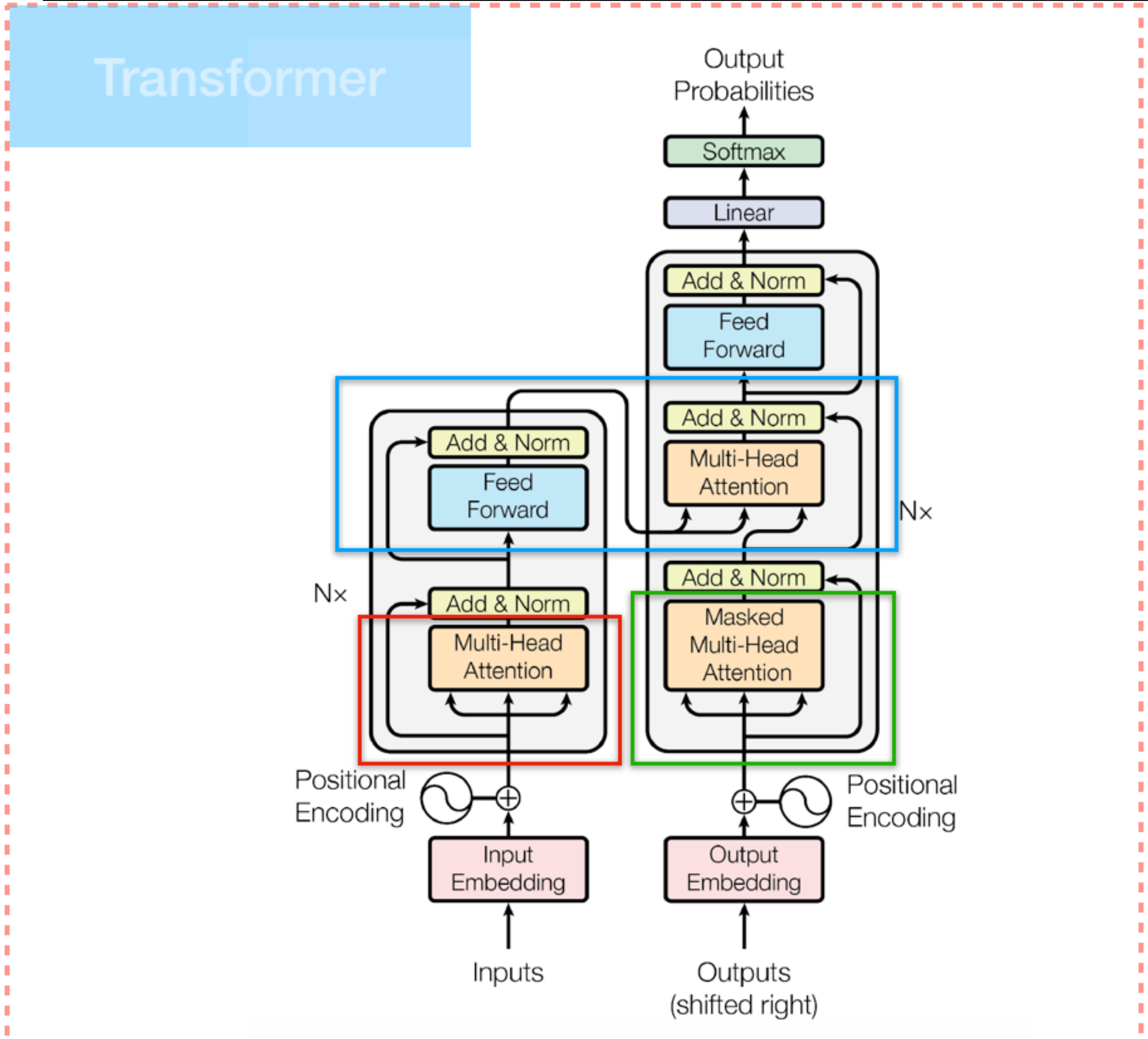


Figure 1: The Transformer - model architecture.

encoder self attention

1. Multi-head Attention
2. Query=Key=Value

decoder self attention

1. Masked Multi-head Attention
2. Query=Key=Value

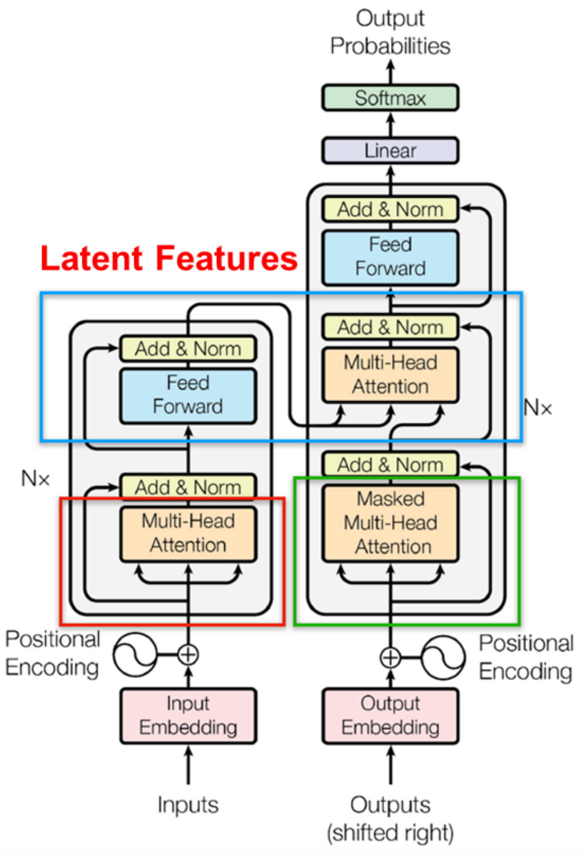
encoder-decoder attention

1. Multi-head Attention
2. Encoder Self attention=Key=Value
3. Decoder Self attention=Query

Transformer to GPT

Transformer

Input -> **Encoder** -> Latent Feature + Masked Output -> **Decoder** -> Output



An [redacted] a day keeps the doctor away

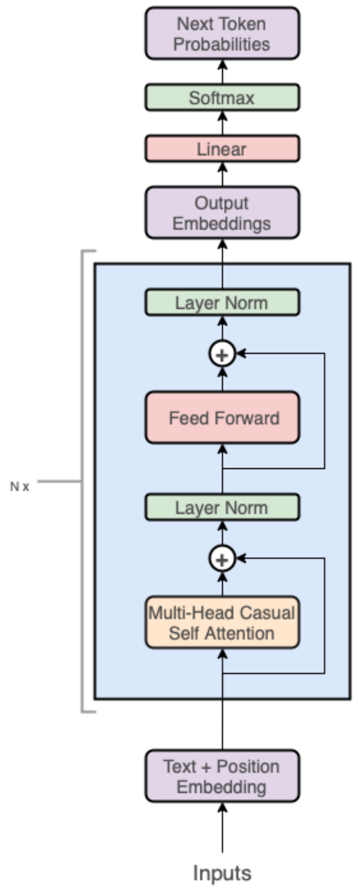
apple 95%
banana 5%

Masked Language Learning

An <u>apple</u> a day keeps the doctor away

GPT

Input -> **Decoder(with Casual mask)** -> shift Output



An

apple 99%
almond 1%

An <u>apple</u>

a 99%
watch 1%

An apple <u>a</u>

Autoregressive Learning

GPT Evolution

Not only Bigger and Bigger

Fine Tuning
Or
In context Few shot Learning

175B parameters

GPT-3

96 decoders

2048 token size

Fine Tuning

1.5B

GPT-2

48 decoders

1024 token size

Fine Tuning

117M

GPT-1

12 decoders

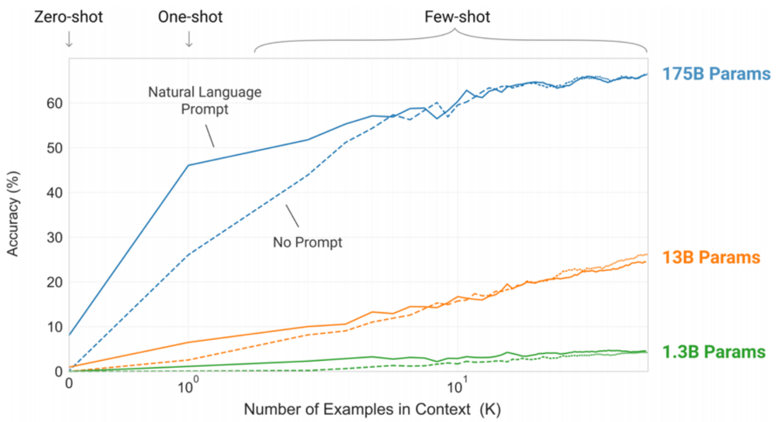
512 token size

2018 → 2020 → 2020

As the model and dataset get larger, it will know more and more

"GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model."

From **Language Models are Few-Shot Learners (2020)**

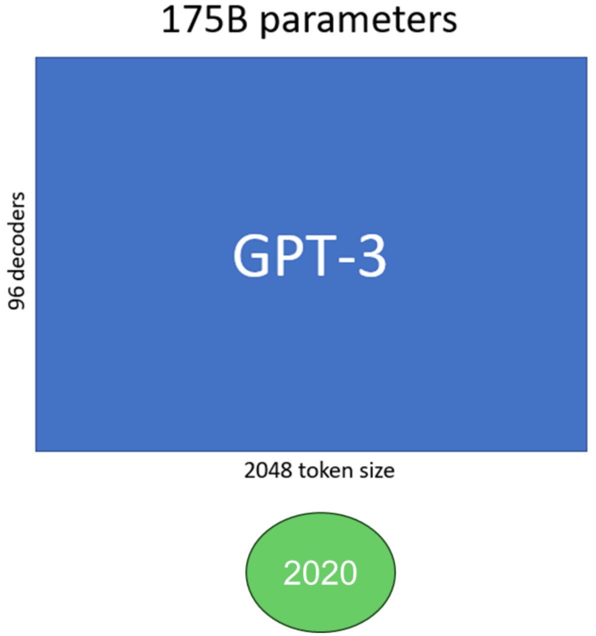


GPT Evolution

Thinking and Answering policy optimization
Reinforcement Learning from Human Feedback (RLHF)

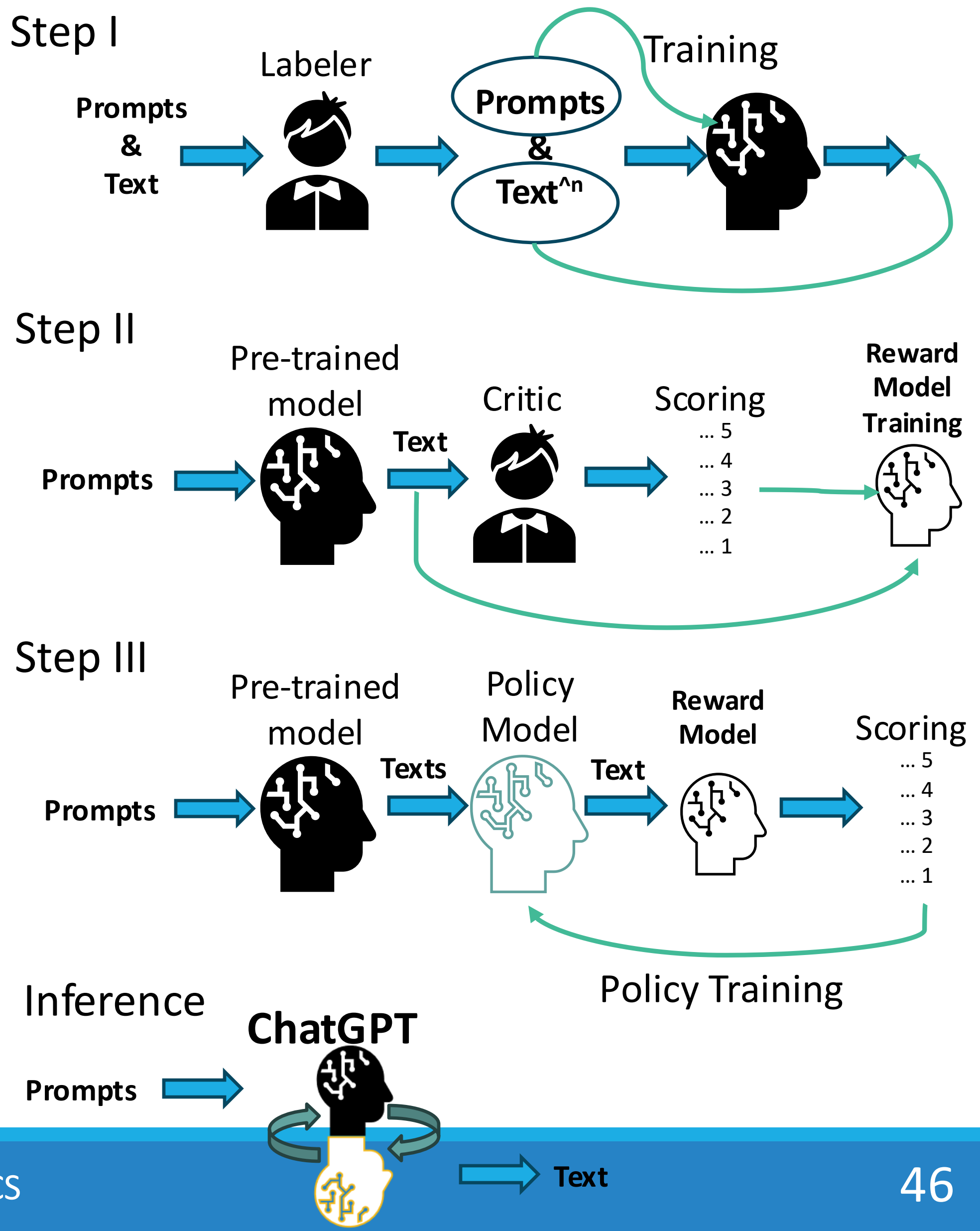
Not only Bigger and Bigger

Fine Tuning
Or
In context Few shot Learning



?

How does the Model Answer smartly or more like an Adult human



What is Next ?

AutoGPT

What goal do you want

Connect to Internet to Find the way or answer

Write code and debug

Do research and try & error

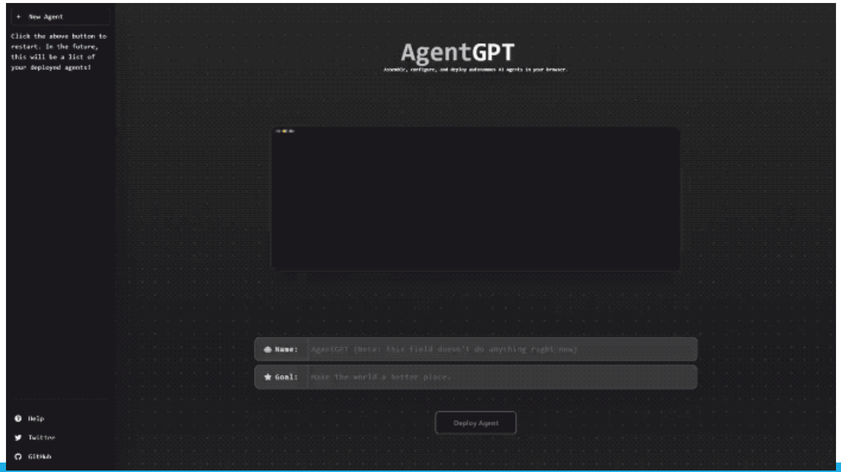
Make a Hypothesis and provide it

Summarization and Organization

Like JARVIS

In Iron Man

The no longer future will come true



https://agentgpt.reworkd.ai/zh

Generative AI for Enterprise

APPLYING ARTIFICIAL CREATIVITY IN INDUSTRY

Graphen Robotics

Meet Aiia

World's First A.I. Digital Human for Daily Life!!

- Hardware-Software Integrated Local AI 'Brain'.
- Privacy / Individual / Personal
- Speaks English, Chinese, Japanese, Spanish, Korean, Malay, and Indonesian.
- Avatars with Personality & Emotion
- Eye Contact / Facial Expression
- Integrating with Payment, Mobile Apps, etc.

An avatar powered by artificial intelligence at an A.I. conference in New York in December. Justin Lane/EPA, via Shutterstock

AI as Knowledge Worker

Examples：

Instant reference tool for medication dosages, side effects, and interactions, reducing the risk of medication errors.

Patient education： helping nurses provide accurate, understandable explanations of medical conditions and treatments.

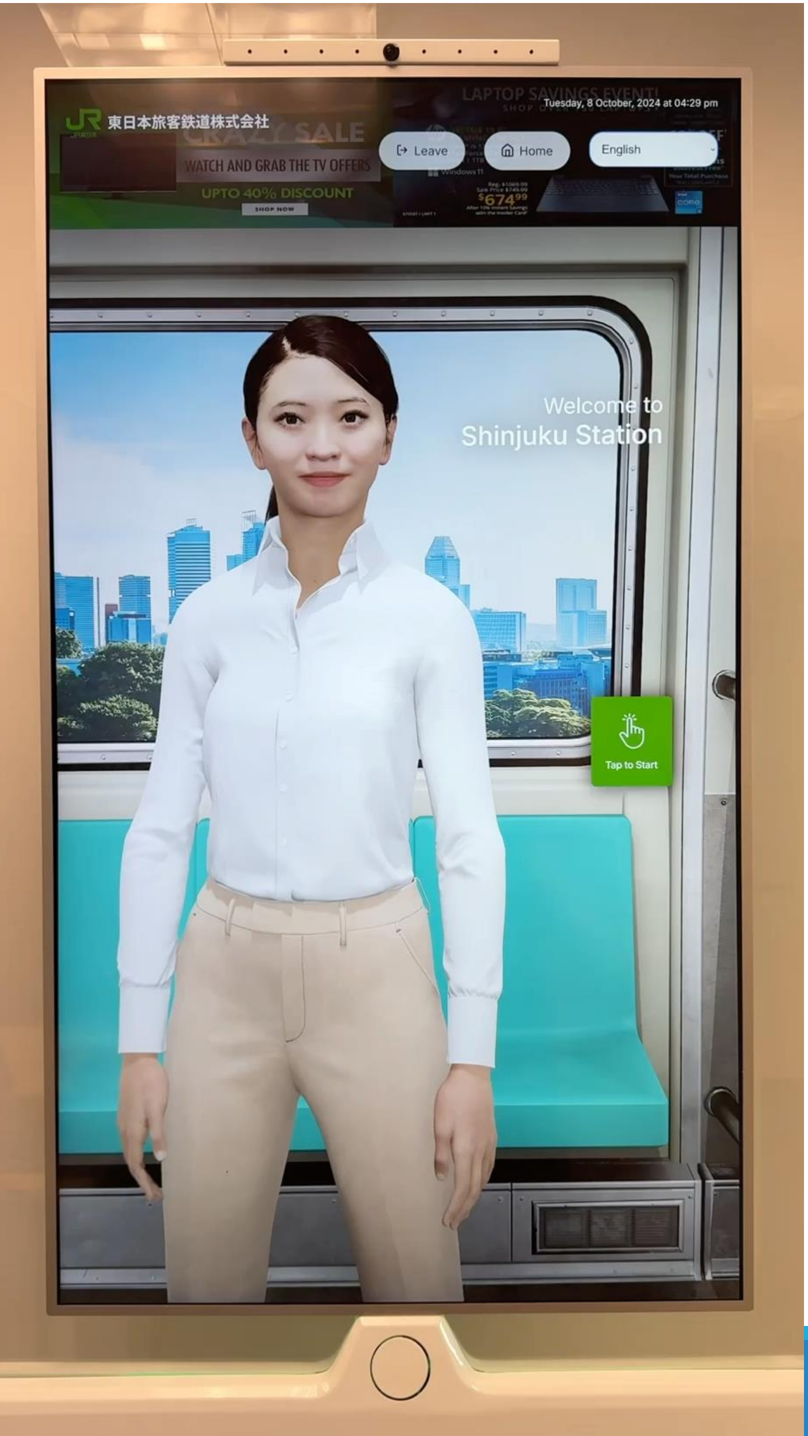
Question：What is the infusion time for 1 unit of Packed Red Blood Cells?

Aiia Nurse Assistant: PBRCs are a blood product used to replace erythrocytes; infusion time for 1 unit is usually between 2 and 4 hours.

Source: The answer is obtained by retrieving page 158 in the provided PDF, which is the RN Exam textbook.

➔ Aiia answered 90% questions correctly in New York RN License Exam

Aiia Examples



Garden by the Bay
@ Singapore

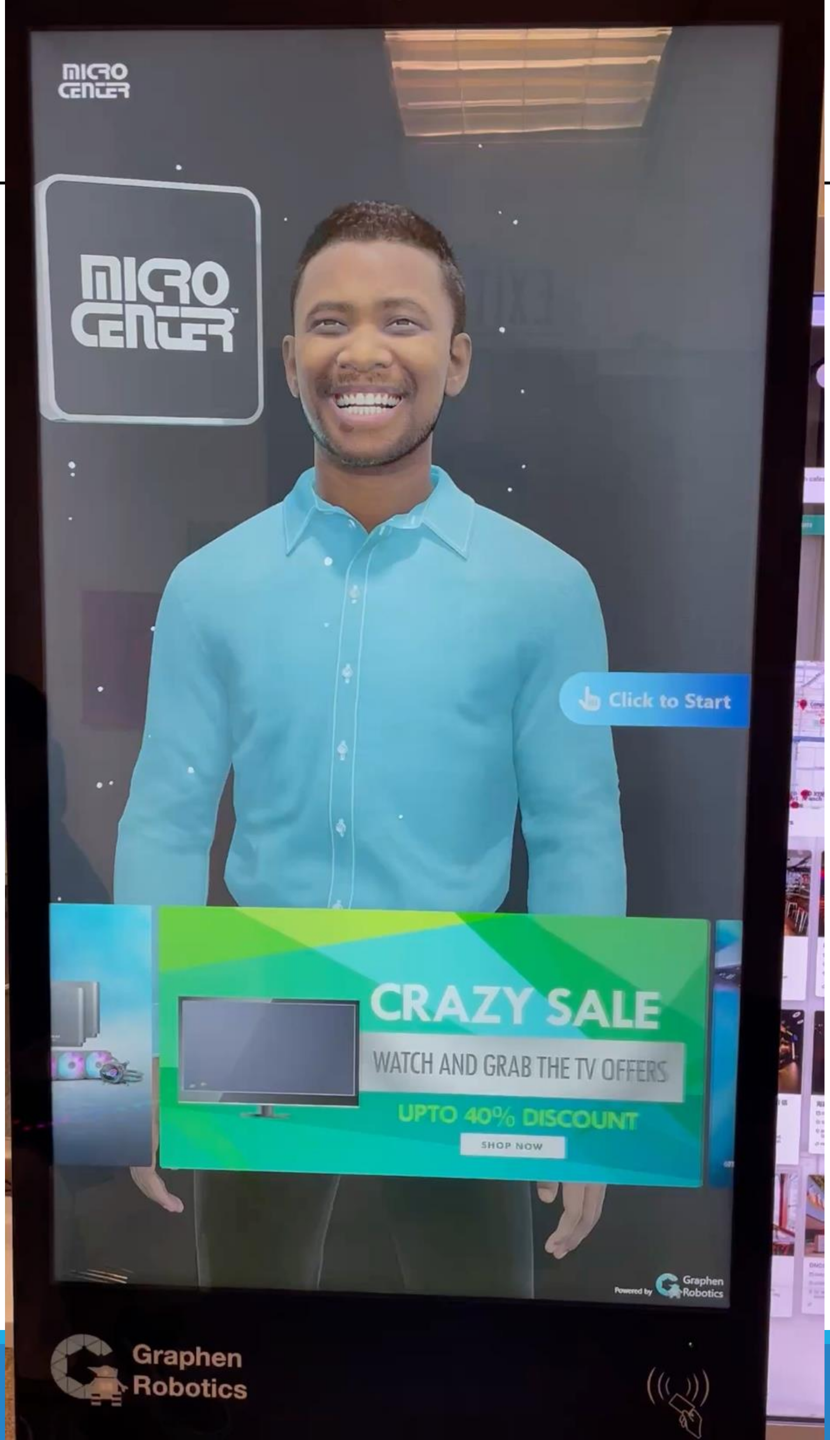
JR East @Tokyo

Medical Assistant
@ US and Taiwan

https://www.youtube.com/watch?v=J9zsaW0gDN4

Graphen Robotics


MICRO CENTER
MICRO CENTER
Click to Start
TOP DEALS!
SHOP NOW
POWERSPEC G235
INTEL CORE I5-12400F (2.5GHZ)
NVIDIA GEFORCE RTX 4060
16GB DDR4 RAM | 1TB SSD
$749.99
592345 / LIMIT 1
GAME PASS
Windows 11
Powered by Graphen Robotics
Graphen Robotics


MICRO CENTER
MICRO CENTER
Click to Start
CRAZY SALE
WATCH AND GRAB THE TV OFFERS
UPTO 40% DISCOUNT
SHOP NOW
Powered by Graphen Robotics
Graphen Robotics


MICRO CENTER
MICRO CENTER
Click to Start
TOP DEALS!
SHOP NOW
POWERSPEC G235
INTEL CORE I5-12400F (2.5GHZ)
NVIDIA GEFORCE RTX 4060
16GB DDR4 RAM | 1TB SSD
$749.99
592345 / LIMIT 1
GAME PASS
Windows 11
Powered by Graphen Robotics
Graphen Robotics

Six Aiia demos at New York Convention Center (April 2023 @ NY Auto Show)

Concierge Aiia
Hotel, Train Stations, Travel Agent

Cashier Aiia
Drinks, Restaurants, Supermarket, etc.

Sales Aiia
Retail stores

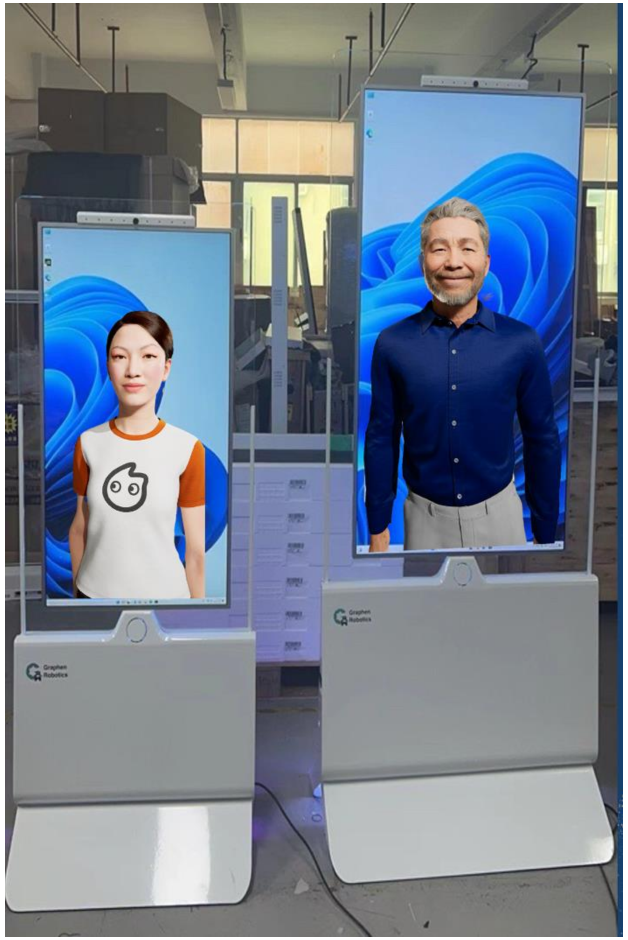
Nurse Aiia
Hospital, Nursing Home

Office Assistant Aiia
Financial Institutes

Customer Service Aiia
Automotives

Graphen Robotics Hardware

Aiia Glass

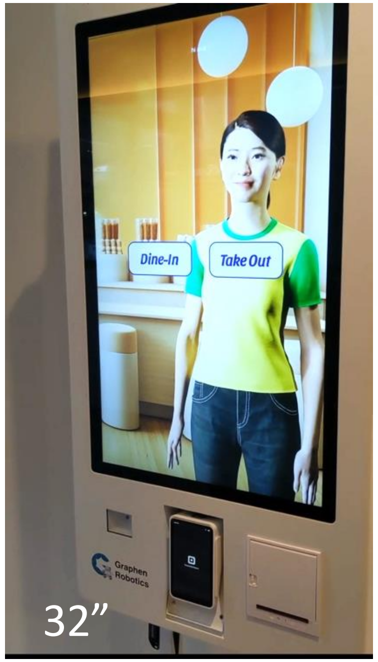



55” & 43”

Aiia Classic

55” & 43”

Aiia Kiosk



Aiia Robot

32”

Aiia Tablet

Aiia Hologram

75” & 86”

Example: Aiia Clinical Applications and Remote Health Assistant

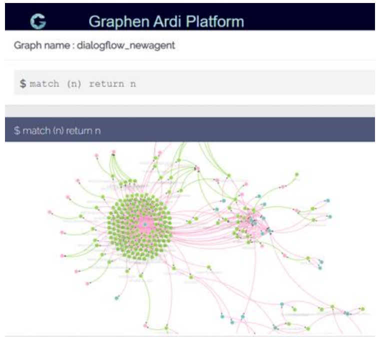
- Health instruction
- Family meeting
- Exercise instruction
- Medical documents
- Data recording
- Individual care plan
- Virtual reality therapy
- Communications

Secure, Scalable AI Chatbot for Enterprise

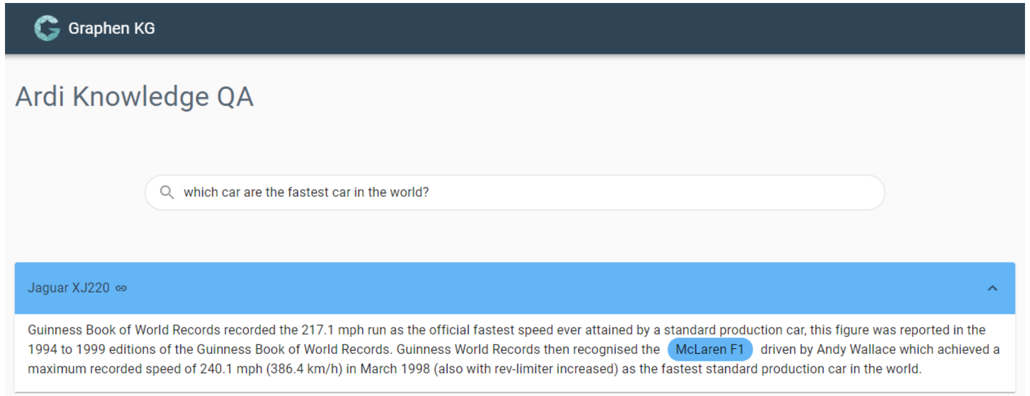
Internal Database (million records):
Product Documentations
FAQs
Customer Service Records
Account APIs

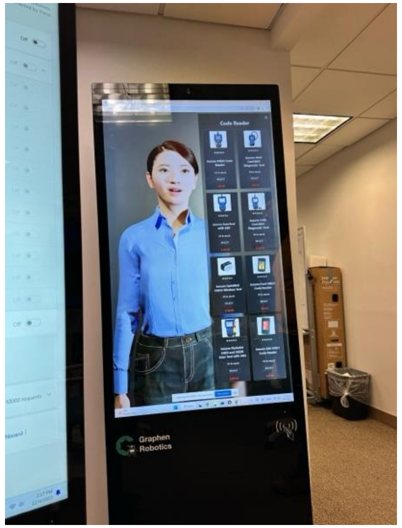
External Data:
Wikipedia, LLM, internet, etc.

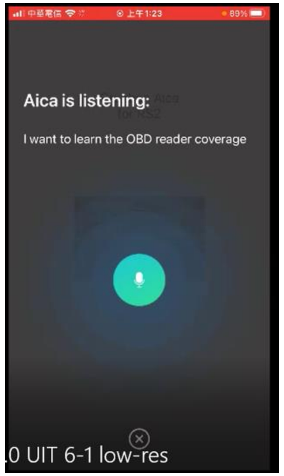
GraphDB Analytics



Ardi Knowledge and Reasoning RAG







Messenger from Facebook
Slack
Skype
LINE
Twilio (Text Messaging)
clover
Square

Customer-Facing: Texting, Phones, WebApps, Emails, ChatApps, dedicated App (Android/iPhones), and Graphen Kiosks with Point-of-sales payments

Ardi Backend

ChatBot Platforms

Falcon LLM

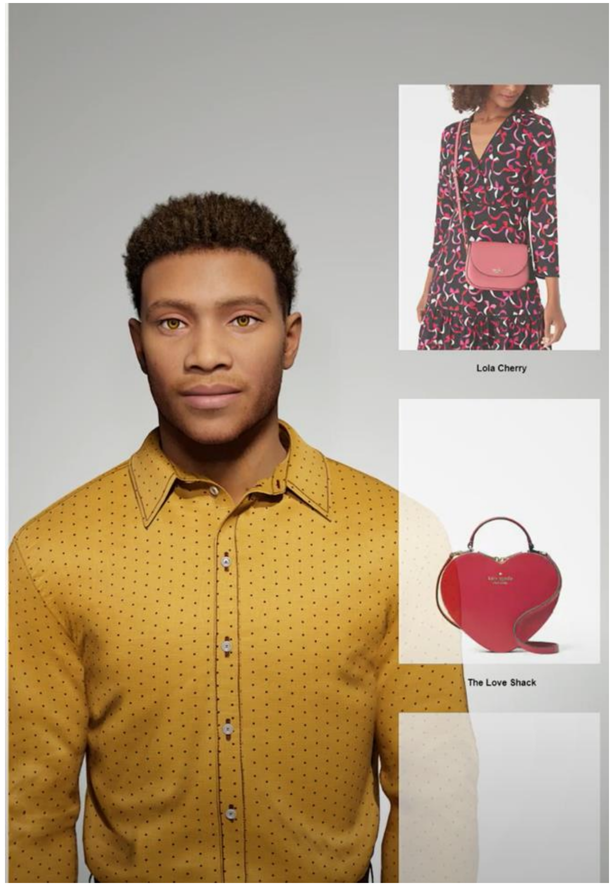
OpenAI

LLM Weakness and Graphen Enhancements

- LLM **lacks temporal and causality reasoning**,
  - Graphen Bayesian Reasoning RAG provides.
- LLM **lacks Analytical Reasoning** and requires agentGPT,
  - Graphen Analytics extends capability with entity-resolution, recommendation, multi-hop relations and cyclic detection.
- LLM **Limited Conversation Window**,
  - Graph Database maintains unlimited history similar to memGPT.
- LLM **cannot separate facts from opinions**, on top of hallucination.
  - Graphen Ontology-based COT improves chain of thoughts
- LLM is **‘single-minded’**, even with mix-of-experts model
  - Graph Database provides Multi-agent history, cross-reasoning, and dynamically adjusts for best chain-of-thoughts.

Graphen’s Enterprise LLM Solutions

- Context-sensitive voice recognition tailored to your company’s products and industry.
- Voices and personas tailored to your products.
  - https://www.graphen.ai/products/Aiia.html
- Graphen Knowledge Graph
  - To infuse and boost your products and industry.
- Graphen Guard Rail
  - To guard against AI misbehavior and incorrect response.
  - To maintain user and enterprise privacy



Steps to develop a customer service AI

Step 1: Knowledge Training

Data preparation and improved entity voice recognition in industrial and brand contexts

Enrichment of models with Aica’s knowledge to avoid hallucination using an enterprise-approved

Step 2: Courtesy Training

Model with initial Guard Rail to ensure knowledgeable courtesy response.

Context-boosting voice-based recognition and personable response.

Step 3: Fine-tuning and Production Tests

Fine-tuning and bulk testing for production deployment.

Industry Context Boosting

- Natural Language Understanding (NLU) in Speech-to-Text, Text-to-Speech

- To enhance the chatbot's understanding of customer inquiries, the NLU component of ChatGPT will be fine-tuned for Innova's specific domain and industry. This involves:

1) **Domain-Specific Training Data**: Training the chatbot on a dataset containing industry-specific terms, phrases, and customer inquiries to improve its understanding and relevance.
2) **Intent Recognition**: Developing a robust intent recognition system to accurately identify customer requests and queries.
3) **Entity Recognition**: Extracting key entities (e.g., car models, part names) from user input to provide more precise responses.
4) **Context Management**: Implementing context management techniques to maintain the conversation context and provide coherent responses over multiple interactions.

Speech-to-Text Example: Recognizing aliases and terms in Auto Industry, and product and service names in Innova.

Text-to-Speech: Pronouncing Innova Products as intended, such as “Innova Fifty-Six Ten” as opposed to “five-six-one-zero”.

Key to Super-Knowledgeable Brand Ambassador

- Must only use sourced info, and able to cite where info were from, to avoid hallucination impression.
- Must detect and reject irrelevant questions and comments. (Guardrails)
- Must express understanding of user/caller’s issue, and lead conversations to solution. This shows intelligence, reasoning and domain-knowledge.
- Must assure privacy of users (in conversation and backend practice)

Graphen prevents hallucinations

- Knowledge Injection
  - intentionally misspelled leon Chen, Innova's CEO name to show successful injection)

who's the ceo of innova

The CEO of Innova is Eon Chen.

  - Only Chen's recommendation is allowed

how do i go to autozone

To go to Autozone, you can visit their location at 15505 Northern Blvd, Flushing, NY 11354. Their phone number is +1 (718) 961-6102. Alternatively, you can also visit their location at 55 Cathedral Ave, Hempstead, NY 11550. Their phone number is +1 (516) 489-0260. Please note that these are the addresses and phone numbers provided by Chen. It's always a good idea to call ahead or check their website for the most accurate and up-to-date information.

Guardrail Example

- Guardrail Demo

can i have breakfast

I'm sorry, but I can only assist with inquiries related to Innova car diagnostic tools and services. For breakfast recommendations, I suggest checking online directories or asking local residents for their favorite breakfast spots. Is there anything else I can help you with regarding Innova products or services?

your voice is sweet

Thank you for the compliment! I'm here to assist you with any questions or concerns you may have about Innova car diagnostic tools and services. How can I assist you today?

Conversation Handling

hi my car is broken

What specific issue are you experiencing with your car?

my check engine light is on

Do you have an Innova OBD tool to retrieve the trouble code?

no

I'm sorry to hear that. In order to diagnose the issue with your check engine light, it would be helpful to have an Innova OBD tool. If you don't have one, I recommend getting one to easily retrieve the trouble code and get more information about the issue. If you need assistance in choosing the right OBD tool for your car, please let me know the make, model, and year of your car.

Graphen vs ChatGPT comparison example

You
hi I need code definition

Car Fixer
Sure, I can help with that. Could you please provide me with the specific OBD2 code you're inquiring about?

You
P1136

Car Fixer
The OBD2 code P1136 generally refers to a fault in the fuel trim system. Specifically, it indicates that the Long Term Fuel Trim Additive Air (Bank 1) is at its maximum limit. Here are some key points about this code:

1. **Bank 1**: This refers to the side of the engine which contains the first cylinder. In a V-shaped engine, this would be one of the two cylinder banks.
2. **Long Term Fuel Trim**: This is part of the engine's management system that adjusts the fuel mixture over time to maintain an optimal balance. It compensates for factors like engine wear, changes in altitude, and other variables.

Code	Make	Description
P1136	HONDA	HO2S Rich Mean Bank 1 Sensor 1
P1136	FORD	(HO2S11) circuit high voltage

Graphen Aica For RS2 Testing Platform

hi I need code definition

Sure, I can help you with that. Could you please provide me with the DTC code and the make of your car?

P1136 it's a Honda

Internal right rear power window sub switch malfunction

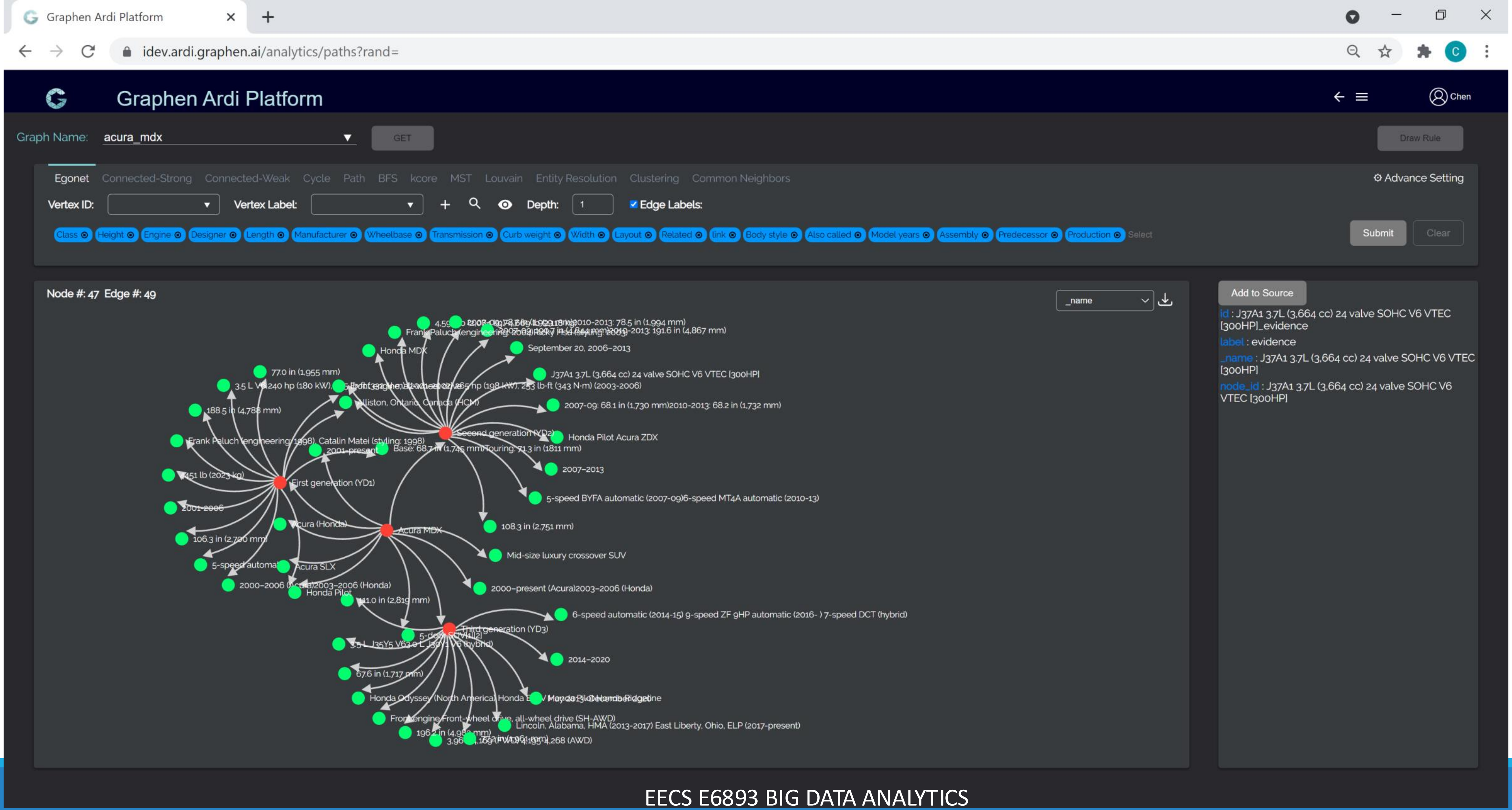
how about Ford

Ho2S Bank 1 Sensor 1 Heater Circuit High Input

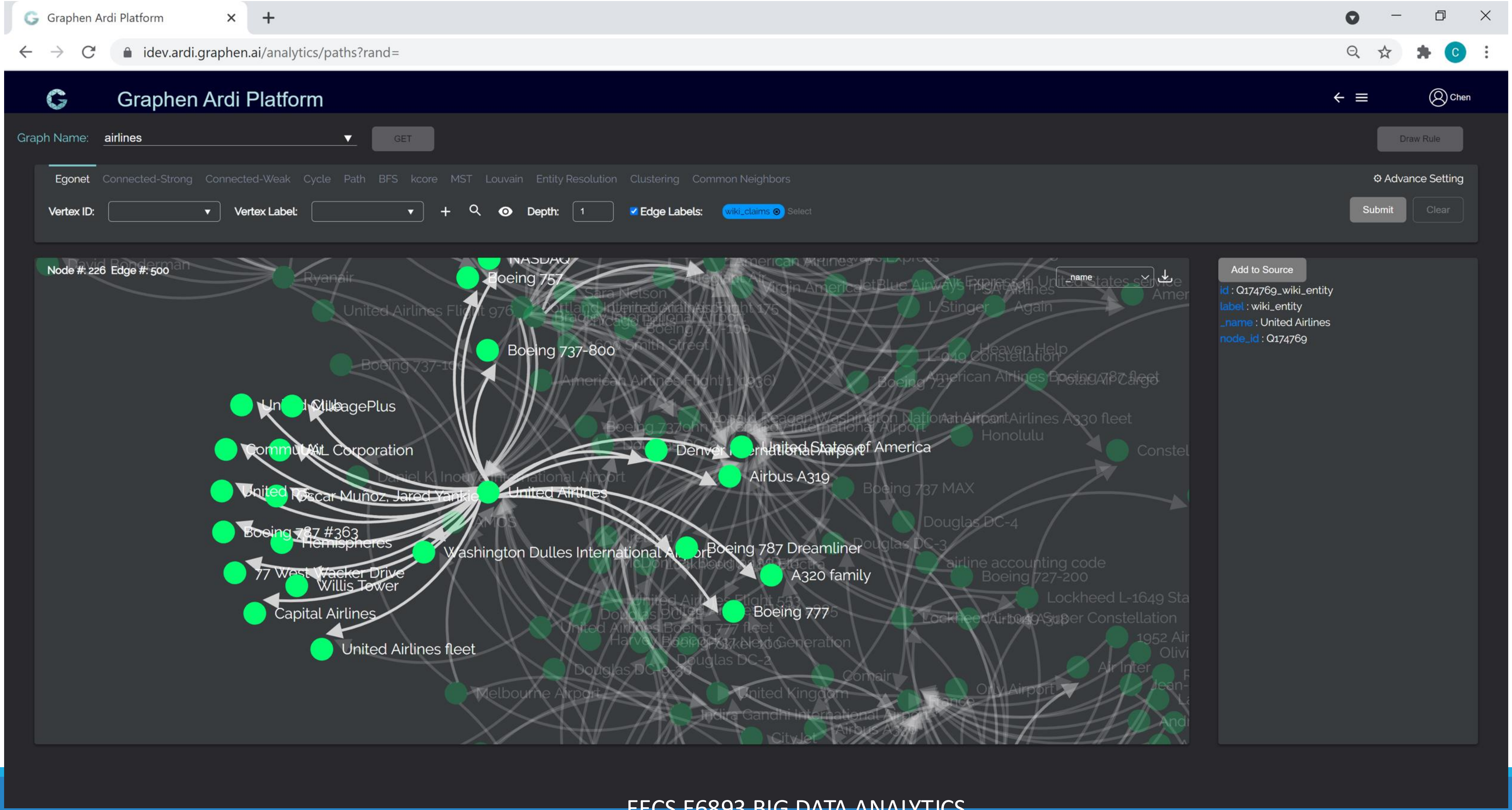
```
s.csv
Ford,P1136,Ho2S Bank 1 Sensor 1 Heater Circuit High Input
Honda,P1136,Internal right rear power window sub switch malfunction
```

ChatGPT is INACCURATE and didn't even ask
for car make. Graphen asks for Make and answer as provided.

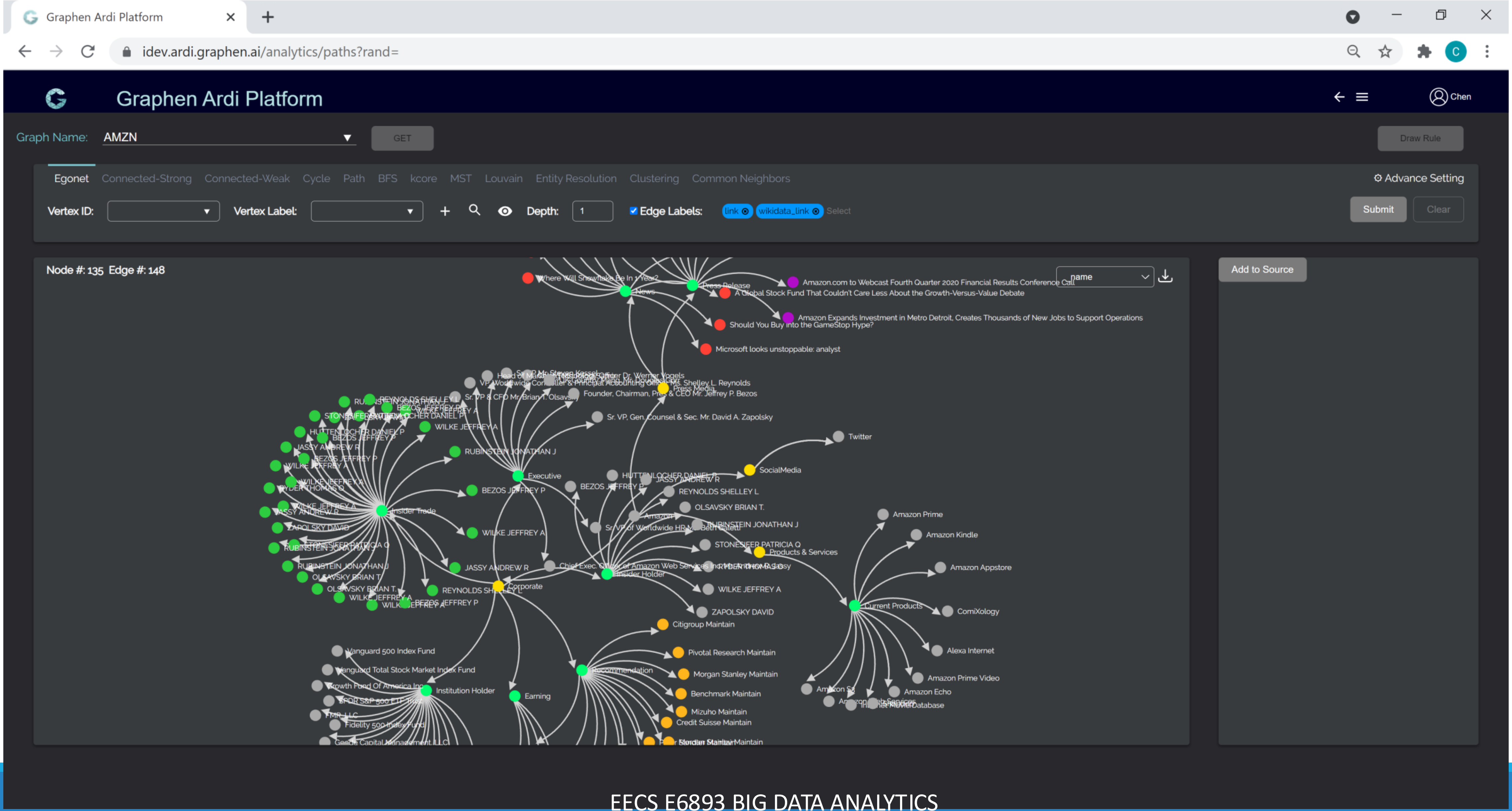
Graphen Knowledge Graph expansion for LLM example I: Car brands, foreign rebadges, and engine parts



Graphen Knowledge Graph expansion for LLM example 2:
Airline Industry Knowledge Graph with aircrafts, airports, airlines relationship, etc



Graphen Knowledge Graph expansion for LLM example 3:
Amazon corporate board, executive, insider trades, news, products and investors Knowledge Graph



GraphDB Keeps Conversation Context

- Graphen Graph Database is memory-mapped on-disk property graph.
- Allows long-term (days) of conversation keeping and filtering for query.
- Enables filtering for what to send to LLM, for example, outdated or unused RAG results need not be in future query, but still be kept in graph structure.

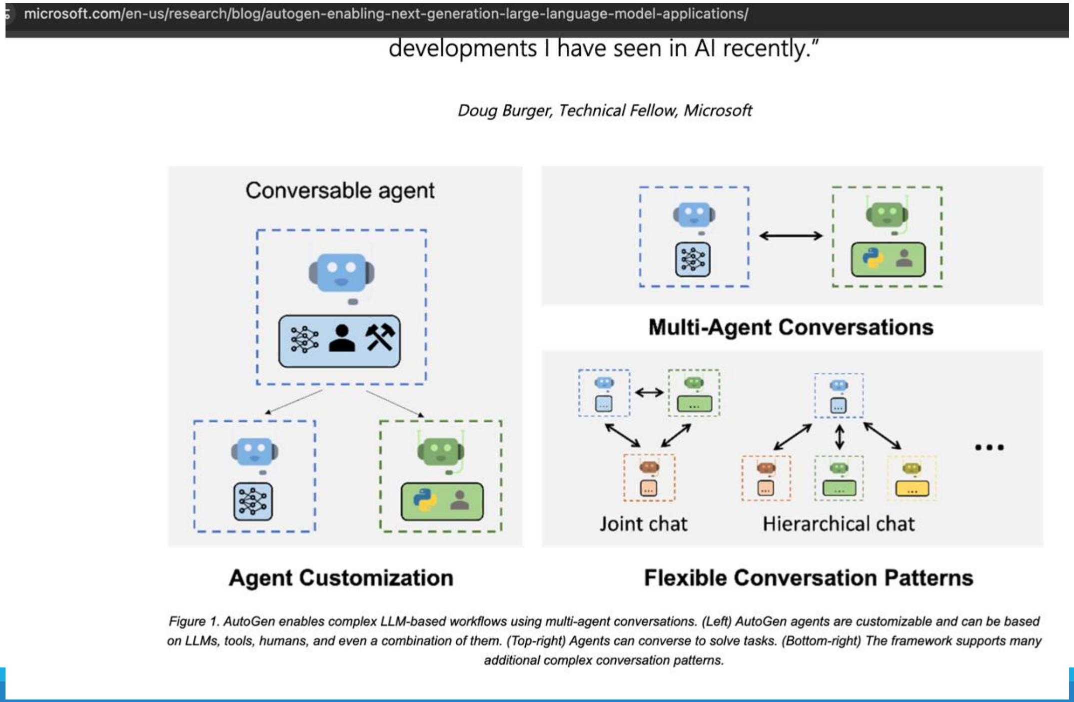
MemGPT: Unlimited Memory without Token Constraints for Generative AI Platforms, like GPT-4, LaMDA, PaLM, LLAMA, CLAUDE, and others

Lawrence Teixeira · Follow
11 min read · Nov 4

77

Graphen Analytics over multi-LLM.

- Enables multi-agent-like behavior, by querying multiple LLMs, and maintain relationships, reasoning and conclusion states in graph. Filtering can be applied to reduce resulting bloated conversation history.



Spy Kids 2017 movie Game Over, where the main villain played by Sylvester Stallone
Created multiple avatars to advice himself.