

MASTERING LARGE LANGUAGE MODELS

Yogesh Haribhau Kulkarni



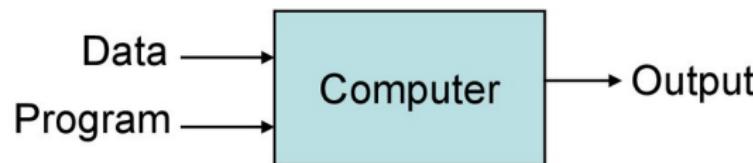
Outline

1 INTRODUCTION

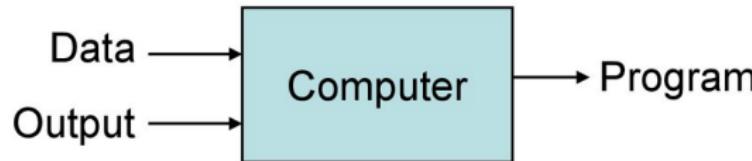
Background

Traditional vs. Machine Learning?

Traditional Programming



Machine Learning



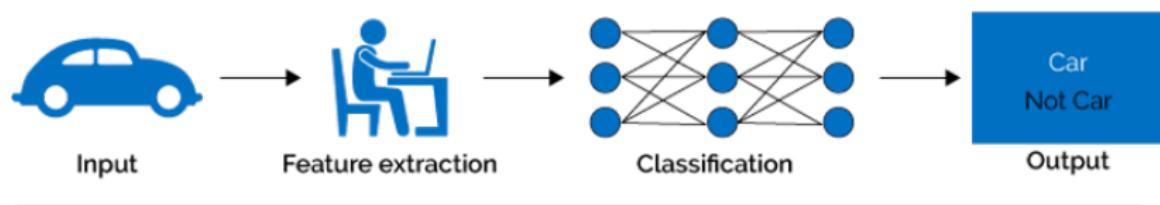
Why Machine Learning?

- ▶ Problems with High Dimensionality
- ▶ Hard/Expensive to program manually
- ▶ Techniques to model 'ANY' function given 'ENOUGH' data.
- ▶ Job \$\$\$

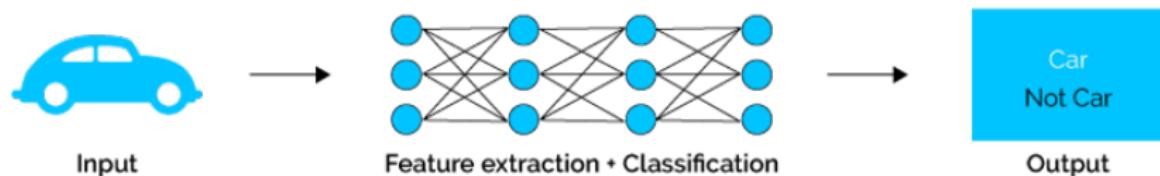
ML vs DL: What's the difference?

Deep learning algorithms attempt to learn (multiple levels of) representation by using a hierarchy of multiple layers

Machine Learning



Deep Learning



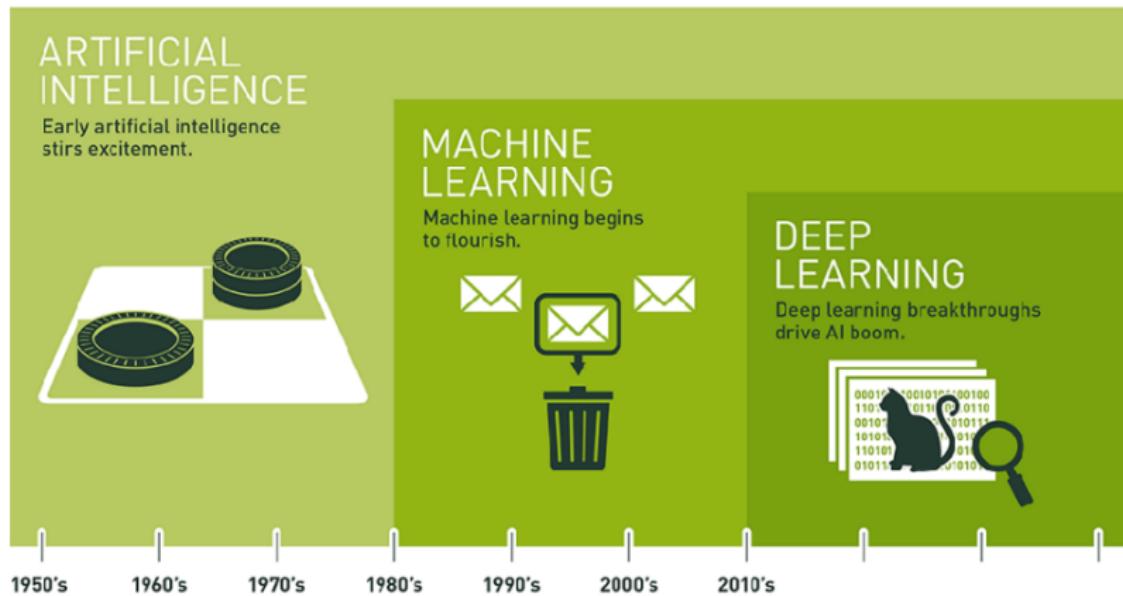
(Reference: <https://www.xenonstack.com/blog/static/public/uploads/media/machine-learning-vs-deep-learning.png>)

Use Deep Learning When ...

- ▶ You have lots of data (about 10k+ examples)
- ▶ The problem is “complex” - speech, vision, natural language
- ▶ The data is unstructured
- ▶ You need the absolute “best” model

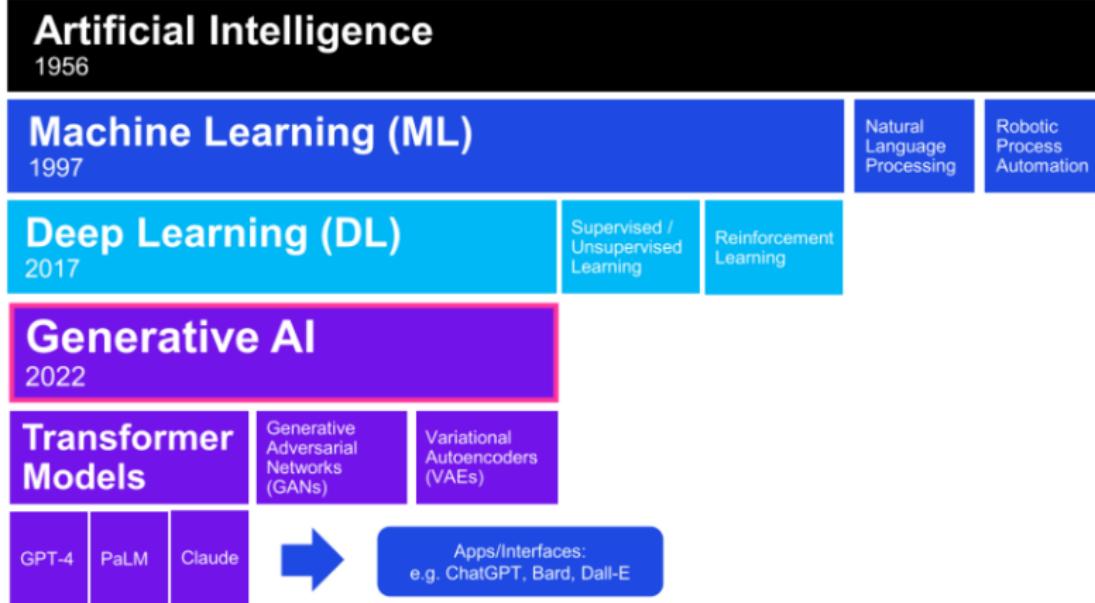
(Ref: Introduction to TensorFlow 2.0 - Brad Miro)

Relationship between AI, ML, DL



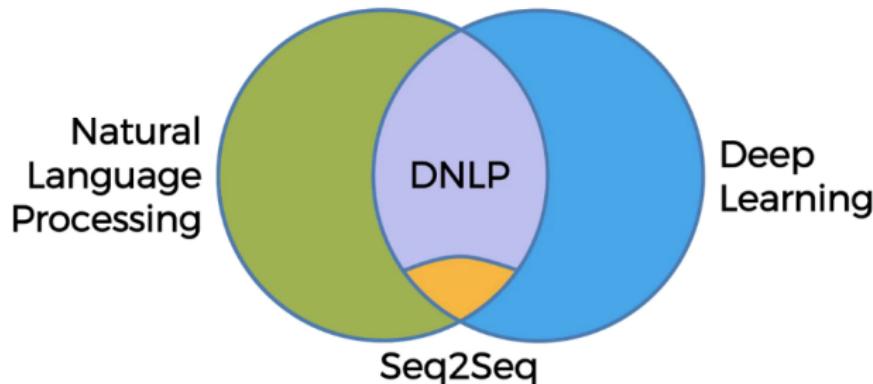
(Ref: <https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/>)

Get Time Sense



(Ref: Preparing your board for Generative AI - KPMG)

What is Deep NLP



(Ref: Deep Learning and NLP A-Z - Kirill Eremenko)
(Note: Size is not indicative of importance)

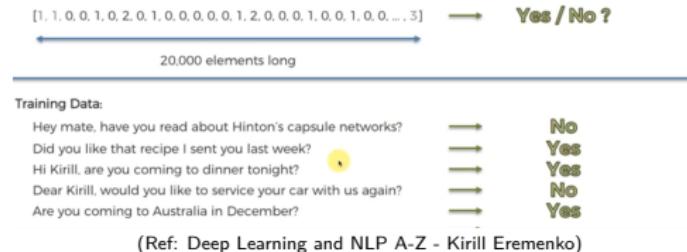
- ▶ Green part is NLP (rule based, linguistic)
- ▶ Blue part is Deep Learning not applied to NLP
- ▶ Purple is Deep NLP (DNLP), NN applied for NLP use cases
- ▶ Seq2Seq is heavily used technique of DNLP for sequence to sequence modeling, eg Translation, Q & A, etc.

Overview of Large Language Models

Typical Machine Learning Classification

- ▶ Questions are converted to bag of words (a vocab long vector, having frequency of specific words at their places)
- ▶ Each question thus gets converted to fixed size vector, which acts as list of features.
- ▶ In training, weights are computed based on the given target.
- ▶ Once model is ready, it is able to answer Yes or No to the question.

Hello Kirill, Checking if you are back to Oz. Let me know if you are around ... Cheers, V



Context

Representing words by their context



- Distributional semantics: A word's meaning is given by the words that frequently appear close-by
 - “*You shall know a word by the company it keeps*” (J. R. Firth 1957)
 - One of the most successful ideas of modern statistical NLP!
- When a word w appears in a text, its **context** is the set of words that appear nearby (within a fixed-size window).
- Use the many contexts of w to build up a representation of w

...government debt problems turning into **banking** crises as happened in 2009...

...saying that Europe needs unified **banking** regulation to replace the hodgepodge...

...India has just given its **banking** system a shot in the arm...

These **context words** will represent **banking**

Evolution of Vectorization

Vectors can be statistical (frequency based) or Machine/Deep Learning (supervised) based. Simple to complex.



(Ref: Analytics Vidhya <https://editor.analyticsvidhya.com/uploads/59483evolution.of.NLP.png>)

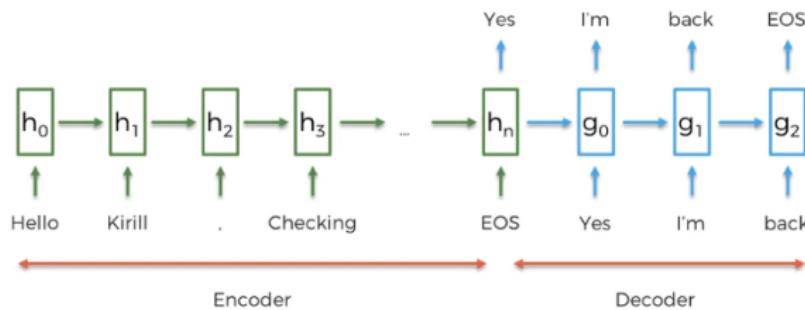
Word vectors

- ▶ Dense vector for each word
- ▶ Called distributed representation, word embeddings or word representations
- ▶ Test: similar to vectors of words that appear in similar contexts

$$\text{banking} = \begin{pmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \end{pmatrix}$$

Seq2Seq architecture

Hello Kirill, Checking if you are back to Oz.

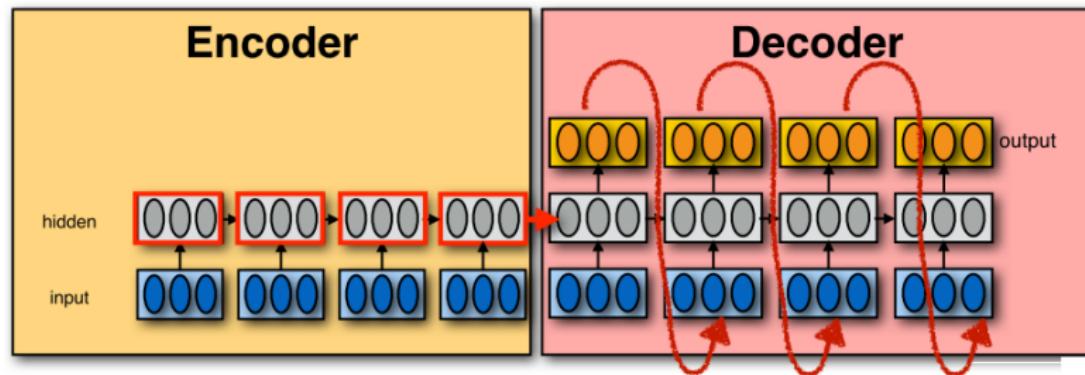


(Ref: Deep Learning and NLP A-Z - Kirill Eremenko)

During training, Encoder is fed with Questions and decoder with Answers. Weights in gates, hidden states get settled. During testing for each sequence of input, encoder results in to a combo vector. Decoder takes this and starts spitting out words one by one, probabilistically.

Encoder-Decoder (seq2seq) model

- ▶ The decoder is a language model that generates an output sequence conditioned on the input sequence.
 - ▶ Vanilla RNN: condition on the last hidden state
 - ▶ Attention: condition on all hidden states



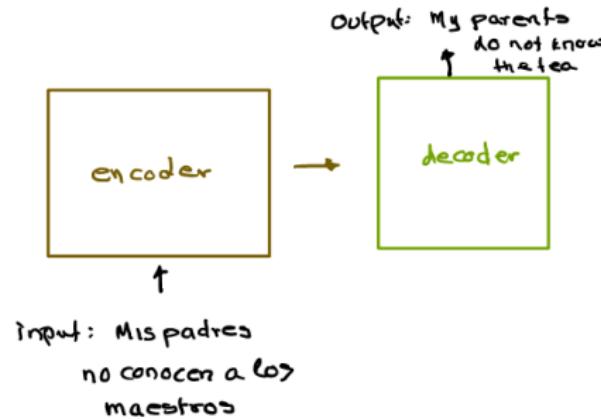
(Ref: CS447 Natural Language Processing (J. Hockenmaier)

Transformers use Self-Attention

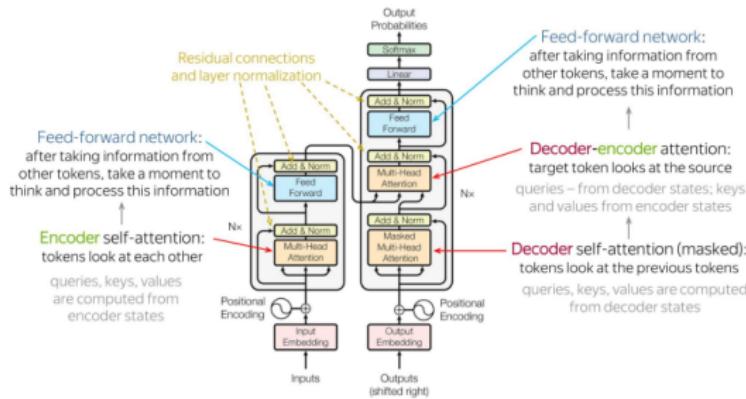
- ▶ Attention so far (in seq2seq architectures): In the decoder (which has access to the complete input sequence), compute attention weights over encoder positions that depend on each decoder position
- ▶ Self-attention: If the encoder has access to the complete input sequence, we can also compute attention weights over encoder positions that depend on each encoder position

Transformers

- ▶ The Transformer is a model that uses attention to boost the speed with which seq2seq with attention models can be trained. The biggest benefit, however, comes from how The Transformer lends itself to parallelization.
- ▶ In its heart it contains an encoding component, a decoding component, and connections between them.



Transformer Models



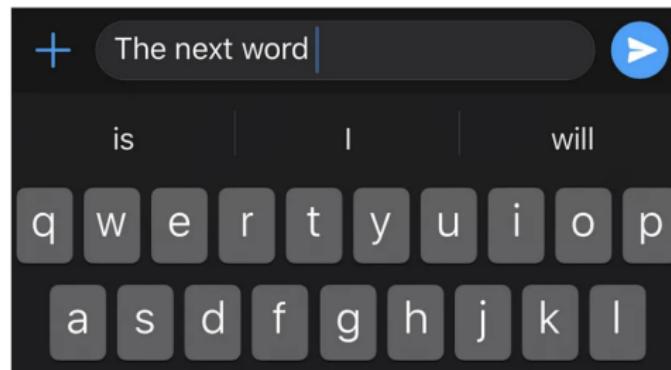
(Ref: The Complete Prompt Engineering for AI Bootcamp (2023))

- ▶ No hidden states, no recurrence, so parallelization possible
- ▶ Context information captured via attention and positional encodings
- ▶ Consists of stacks of layers with various sublayers

Transformers are basis of (the most) Large Language Models

What is a Language Models?

- ▶ While typing SMS, have you seen it suggests next word?
- ▶ While typing email, have you seen next few words are suggested?
- ▶ How does it suggest? (suggestions are not random, right?)
- ▶ In the past, for "Lets go for a ... ", if you have typed 'coffee' 15 times, 'movie' say 4 times, then it learns that. Machine/Statistical Learning.
- ▶ Next time, when you type "Lets go for a ", what will be suggested? why?
- ▶ This is called Language Model. Predicting the next word. When done continuously, one after other, it spits sentence, called Generative Model.



Next word prediction using language modeling in keyboards(Mandar Deshpande)

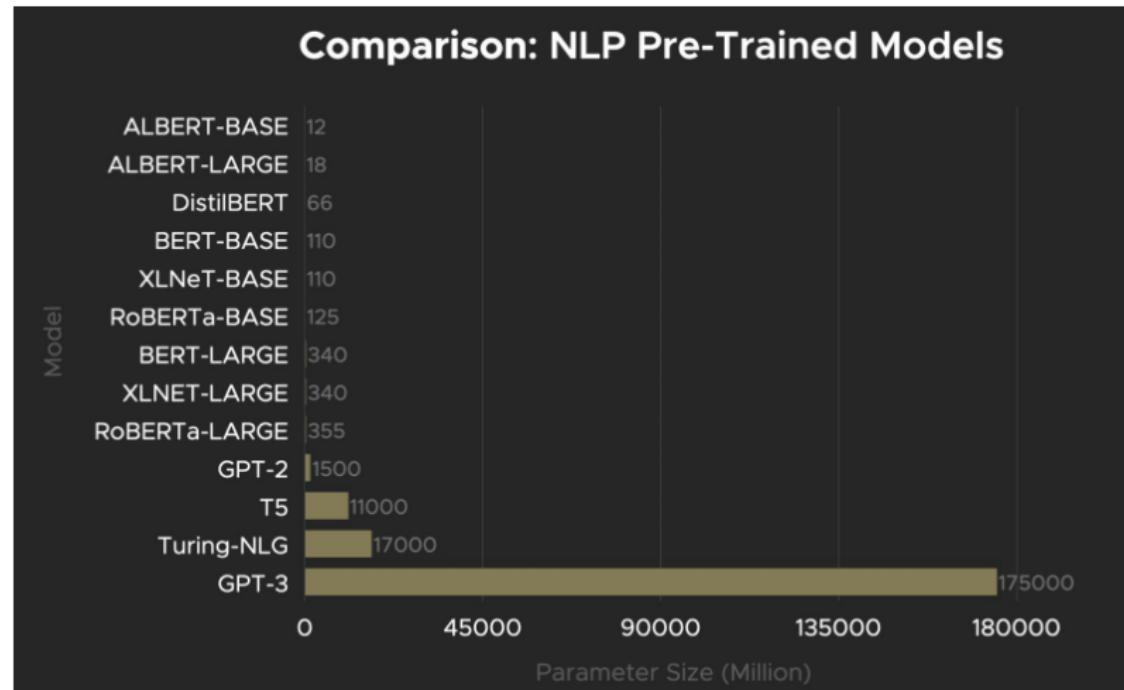
YHK

Overview of LLM

- ▶ Large Language Models (LLMs) are deep neural networks (e.g., GPT-3, BERT) based on the Transformer architecture.
- ▶ LLMs are foundation models trained on large amounts of unsupervised and unstructured data.
- ▶ The Transformer architecture consists of an encoder and decoder, both mostly identical with a few differences.
- ▶ LLMs compute a probability distribution over a vocabulary (list of tokens) given an input prompt.
- ▶ LLMs have limitations like hallucination and issues in chain of thought reasoning, but recent improvements have been made.
- ▶ LLMs are trained for statistical language modeling, which involves predicting the next token based on context.

(Ref: Overview of Large Language Models - Aman AI)

Large Language Models - Comparison



(Ref: Deus.ai <https://www.deus.ai/post/gpt-3-what-is-all-the-excitement-about>)

Example LLMs with Front-Ends

GPT3

The screenshot shows the OpenAI Playground interface. At the top, there's a header with 'Playground' and a search bar containing 'IMD - Long Descriptions'. Below the header are buttons for 'Save', 'View code', and 'Share'. A dropdown menu shows 'Model: text-davinci-003'. On the right side, there are several sliders and input fields for tuning the model's behavior:

- Temperature: 0.7
- Maximum length: 1790
- Stop sequences: (empty)
- Top P: 1
- Frequency penalty: 0.2
- Presence penalty: 0.65
- Best of: 12
- Request start text: (checkbox checked)
- Request restart text: (checkbox checked)

At the bottom left, there's a message: 'Looking for ChatGPT? Try it now!'. The bottom right corner has a small 'Close' button.

Created by OpenAI Access it
with code or without
(Playground
<https://platform.openai.com/playground>)

(Ref: The Complete Prompt Engineering for AI Bootcamp (2023))

GPT3 Features

Highly Configurable

Mode: Complete (Top P: 1, Frequency penalty: 0.2, Presence penalty: 0.43)

Model: text-davinci-003

Temperature: 0.7 (Best of: 12)

Maximum length: 1730

Stop sequences: Enter sequence and press Tab

Buttons: Inject start text (checked), Inject restart text (checked), Show probabilities

Text Generation

Instructions:
 Create a small linkedin post with emojis on "the benefits of side coding projects".
 Character limit: 2000 characters
 Include Emojis: true
 Tone of voice: Professional
 # End of Instructions

Are you looking to stay ahead of the curve in your career? 😊 Side coding projects help learn cutting-edge technologies and hone your coding skills, but they can also help a few of the benefits of side coding projects:

- ✓ You get to work with the latest technologies: Working on side coding projects gives you the opportunity to work with the latest trends in the industry. 🚀
- ✓ You can refine your coding skills: Working on side coding projects can help you apply for jobs. 💼
- ✓ You can showcase your work: Side coding projects also give you the opportunity to showcase your work to potential employers. 🎯

So, if you're looking to stay ahead of the curve, why not give side coding projects a try? #careerdevelopment 🚀

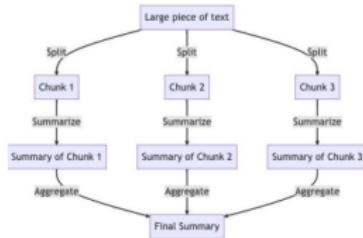
Presets

PPC - Long Descriptions
PPC - Long Descriptions
PPC - Short Descriptions
PPC - Long Headlines
PPC - Headline Generation
Landing pages - Meta Description
Landing pages - Title Tag
Landing pages - FAQS
Landing pages - Change JSON (benefits)
Landing pages - Sub-services List
Landing pages - Subsections copy generation

(Ref: The Complete Prompt Engineering for AI Bootcamp (2023))

GPT3 Usecases

Custom Workflows



Marketing

You are responsible for making landing pages. Firstly we will be making a landing page for "Web scraping Services".

Write an SEO-optimised title tag. The company is called understandingdata. But don't include the company name.

"Data Extraction & Web Scraping Services | understandingdata"

Automation



(Ref: The Complete Prompt Engineering for AI Bootcamp (2023))

GPT3 Limitations

- ▶ Unlike ChatGPT, GPT-3 doesn't store chat history, which can lead to you needing to re-inject the extra context into the original prompt.
- ▶ Tokens are currently limited to 4000 tokens per output. This includes both your initial prompt and the text generated from ChatGPT.

(Ref: The Complete Prompt Engineering for AI Bootcamp (2023))



Bard

Product description: A home milkshake maker
Seed words: fast, healthy, compact
Product names: HomeShaker, Fit Shaker, QuickShake, Shake Maker

Product description: A pair of shoes that can fit any foot size.
Seed words: adaptable, fit, omni-fit
Product names:

Here are some product names for a pair of shoes that can fit any foot size:

- Omni-Fit Shoes
- Adaptable Shoes
- Fit-All Shoes
- Size-Free Shoes
- Universal Shoes

Created by Google Access it via chat <https://bard.google.com/> or encounter it in search results

(Ref: The Complete Prompt Engineering for AI Bootcamp (2023))

Bard Features

Web Browsing

- Nike Air Max 90: The Nike Air Max 90 is a classic style.
- Nike Air Vapormax Flyknit 3: The Nike Air Vapormax Flyknit 3 has a modern design and its comfortable cushioning.
- Jordan I High OG: The Jordan I High OG is a classic and comfortable fit.
- Jordan XIII: The Jordan XIII is a popular sneaker with excellent cushioning.

These are just a few of the many popular shoes that will perfectly suit your needs.

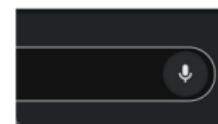
Sources

1. mvmagazine.com/nike-air-force-1-e-la-p...

Drafts

The screenshot shows the 'Drafts' section of the Bard interface. It contains two separate boxes, each labeled 'Draft 1' and 'Draft 2'.
Draft 1: 'Here are some product names for a pair of shoes that can fit any foot size:
• Casual Chucks'
Draft 2: 'Here are some product names for a home milkshake maker that are fast, healthy, and compact:
• HomeShaker
• Fit Shaker
• QuickShake
• Shake Maker
• Smoothie Maker
• Milkshake Maker
• Healthy Shake Maker
• Compact Shake Maker'

Voice



(Ref: The Complete Prompt Engineering for AI Bootcamp (2023))

Bard Usecases

Research

- Nike Air Max 90: The Nike Air Max 90 is a classic style.
- Nike Air Vapormax Flyknit 3: The Nike Air Vapor design and its comfortable cushioning.
- Jordan I High OG: The Jordan I High OG is a classic comfortable fit.
- Jordan XIII: The Jordan XIII is a popular sneaker cushioning.

These are just a few of the many popular shoes that perfect pair of shoes to suit your needs.

Sources

1. mvcmagazine.com/nike-air-force-1-e-e-la-p...

SEO

Search "marketing attribution" and tell me what the top 5 results have in common.

[View other drafts](#)

- The top 5 results for "marketing attribution" all have the following in common:
- They define marketing attribution as the process of determining which marketing channels and messages are most effective in driving conversions.
 - They discuss the importance of marketing attribution for businesses of all sizes.
 - They provide different methods for calculating marketing attribution, including last-touch attribution, first-touch attribution, and multi-touch attribution.

News

What happened with Nvidia's stock price?

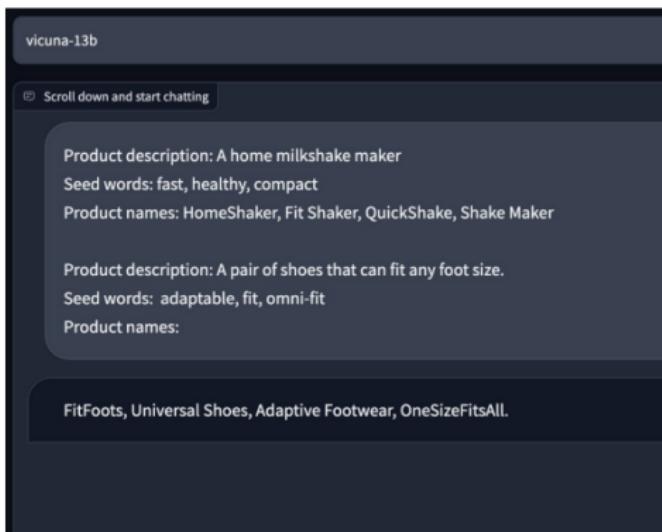
[View other drafts](#)

- Nvidia's stock price rose by 24.37% on May 25, 2023, closing at \$379.80. This was the largest single-day percentage gain for the stock since February 2021. There are a few possible reasons for the sharp rise in the stock price.

- Strong earnings report: Nvidia reported strong earnings for the first quarter of 2023, beating analyst expectations. The company's revenue grew by 84% year-over-year to \$8.29 billion, and its earnings per share grew by 83% to \$1.92.
- Positive outlook: Nvidia gave a positive outlook for the second quarter of 2023, expecting revenue to grow by 68% year-over-year to \$8.1

(Ref: The Complete Prompt Engineering for AI Bootcamp (2023))

Meta LLaMA



(Ref: The Complete Prompt Engineering for AI Bootcamp (2023))

Created by Meta and open-sourced / leaked Access via hugging face or lmsys <https://chat.lmsys.org/> Use one of the fine-tuned models (i.e. Vicuna) for better results

Meta LLaMA

- ▶ Open-Source. Need to build a UX and any advanced functionality around it, and may need to fine-tune it.
- ▶ Many use-cases in the enterprise can't use OpenAI for fear of sensitive data leaking or being used to train the model (though OpenAI claims to keep API data private).
- ▶ If you have 200+ examples fine-tuning beats prompt engineering for a specific defined task.

(Ref: The Complete Prompt Engineering for AI Bootcamp (2023))

Anthropic Claude

The screenshot shows a user interface for generating product names. At the top, there's a header with a square icon, three dots, and a search bar containing the text "Fit Shaker: Fast and Healthy". Below the search bar are two buttons: "Chat" and "Edit Chat". A purple sidebar on the right contains the generated product names: "HomeShaker, Fit Shaker, QuickShake, Shake Maker". The main text area displays the product description: "A pair of shoes that can fit any foot size." and seed words: "adaptable, fit, omni-fit". A text input field says "Product names:" followed by a blue circular button with a white letter "M". Below this, a gray box contains the text: "Here are some suggested product names based on the seed words:". A bulleted list of suggestions follows:

- OmniFit
- AdaptiShoe
- UniSole
- One Size Fits All
- FlexiFit

At the bottom, there's a text input field with placeholder text "Write a message..." and a blue send button with a white arrow pointing right.

(Ref: The Complete Prompt Engineering for AI Bootcamp (2023))

Created by Anthropic Access
via Claude console
<https://console.anthropic.com/>
or API Uses Constitutional AI
rather than RLHF

Anthropic Claude Features

The screenshot shows a user interface for a large language model. At the top, there are icons for a square, a magnifying glass, and three dots. Below that, a header bar has the text "Synopsis for Marketing Memetics" on the left and "Chat" and "Edit Chat" on the right. A light blue sidebar on the left contains text about memetics and its potential. A main text area below the sidebar contains a summary of the book "Marketing Memetics: Reverse-Engineering Creativity To Drive Brand Performance". At the bottom, there is a message input field with the placeholder "Write a message..." and a send button with a play icon.

Synopsis for Marketing Memetics

Chat Edit Chat

M

Memetics shows great promise, but is still pre-Mendel.

Nobody has meticulously traced the inheritance of thousands of memes. Nobody has worked out the math to predict meme transmission. Nobody has revealed the source code that would make memes programmable.

With advancements in machine learning, virtual reality, and neuroimaging, perhaps memetics' own Mendel, Fisher, Crick & Watson are right around the corner. ``

Synopsis:

Here is a synopsis for the book "Marketing Memetics: Reverse-Engineering Creativity To Drive Brand Performance":

- Memetics is the study of how ideas spread between people, as proposed by Richard Dawkins in 1976. Memes are the cultural equivalent of genes; units of cultural information that propagate in the pool of human culture.

Write a message... ▶

(Ref: The Complete Prompt Engineering for AI Bootcamp (2023))



Microsoft Bing (GPT 4)

A screenshot of the Microsoft Bing AI interface. At the top, there's a blue button labeled "Help me plan a trip to London". Below it, a green checkmark icon and the text "Generating answers for you...". The main text area says: "Sure, I can help you plan a trip to London. London is a great city to visit, with iconic attractions, historic museums, and vibrant neighborhoods. You can find things to do in London, from sightseeing tours to fun-filled days out, on [visitlondon.com](#) ¹. You can also book a London Pass to make the most of popular attractions ² or a one-day hop-on hop-off bus tour to explore the sights at your own pace ³. You can find cheap flights to London from \$99 on [KAYAK](#) ⁴ and London vacation packages from \$849 on [Expedia](#) ⁵.
Learn more: [1. visitlondon.com](#) +4 more". At the bottom, there's a "Ask me anything" input field and a microphone icon.

(Ref: The Complete Prompt Engineering for AI Bootcamp (2023))

Powered by OpenAI's GPT-4
Access it via chat
<https://www.microsoft.com/en-gb/bing> or encounter it in search results

Microsoft Bing (GPT 4) Features

Web Browsing

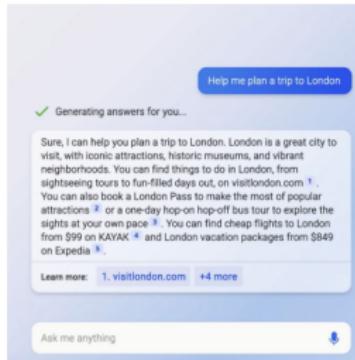
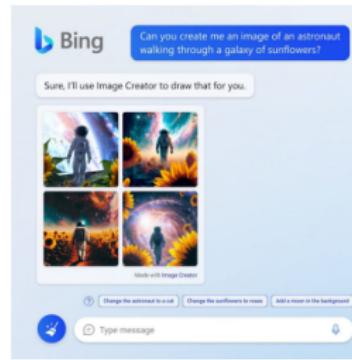


Image Generation



Voice



(Ref: The Complete Prompt Engineering for AI Bootcamp (2023))

Microsoft Bing (GPT 4) Use Cases

Research

Sure, I can help you with that. Based on my research, some of the latest trend in designer earrings are:

- Colorful earrings:** These are moon-shaped earrings featuring pearls, rubies, and sapphires, inspired by floral motifs.
- Dangle earrings:** These are elegant and modern earrings with colored stones that complement any outfit.
- Modern pearl earrings:** These are earrings that update the classic pearl with unique shapes, colors, and textures.
- Colorful statement earrings:** These are earrings that add a pop of color to your spring or summer ensemble with pearls, blues, greens and more.
- Geometric earrings:** These are earrings that feature geometric shapes and metalwork for a modern, edgy look.
- Mismatched earrings:** These are earrings that create an unexpected and playful look by using different styles, textures, and colors.
- Celestial earrings:** These are earrings that feature moons and stars for a whimsical and celestial theme.
- Chain link earrings:** These are earrings that have an edge yet elegant, appear with chain links in gold or silver.

Based on these trends, I have created a detailed description for Bing Image Creator to show you a prototype of an on-trend design:

Learn more: [1.gearpatrol.com](#) [2.menshealth.com](#) [3.vogue.co.uk](#) [+3 more](#)

Learning

Hello, once a student shows an appropriate level of understanding given their learning need, ask them to explain it concisely in their own words. This is the best way for you to know what they understand, or ask them for examples. When a student demonstrates that they understand the concept you can move the conversation to a closer and tell them you're here to help if they have further questions.

Help me learn about basic marketing for a new business

✓ Searching for basic marketing for a new business
✓ Searching for learning needs
✓ Generating answers for you...

That's a great topic to learn about. Marketing is the process of co-creating desirable products or services and distributing them in a way that increases their attractiveness. Marketing can help you reach your target audience, boost your customer base, and ultimately increase your bottom line. These are key concepts of marketing, such as product, price, place, promotion, people, process, and physical evidence.

Before we dive into the details of marketing, I'd like to know more about your learning level. Are you a high school student, a college student or a professional? This will help me tailor my explanations and examples to your needs and interests.

Learn more: [1.binghelpcenter.com](#) [2.braithow.com](#) [3.forbes.com](#) [+3 more](#)

Product Ideas

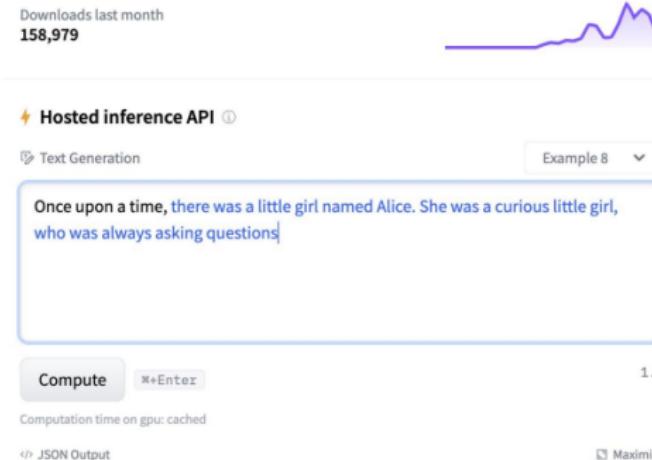
100 meters. The strap is made of black leather with white stitching and has a stainless steel buckle. The watch has a sporty yet elegant design that can suit any occasion.

Learn more: [1.gearpatrol.com](#) [2.menshealth.com](#) [3.vogue.co.uk](#) [+3 more](#)

"A round stainless steel case with a diameter of 43mm."
Made by Bing Image Creator Powered by DALL-E

(Ref: The Complete Prompt Engineering for AI Bootcamp (2023))

Falcon



(Ref: The Complete Prompt Engineering for AI Bootcamp (2023))

Access it via HuggingFace
transformers library
<https://share.descript.com/view/Qp0>
7B and 40B models as well as
instruct fine-tuned
Features:

- ▶ Free for commercial use
- ▶ Open source
- ▶ Possible to fine-tune

Leader board (May 2023)

Rank	Model	Elo Rating	Description
1	💡 gpt-4	1225	ChatGPT-4 by OpenAI
2	💡 claude-v1	1195	Claude by Anthropic
3	💡 claude-instant-v1	1153	Claude Instant by Anthropic
4	gpt-3.5-turbo	1143	ChatGPT-3.5 by OpenAI
5	vicuna-13b	1054	a chat assistant fine-tuned from LLaMA on user-shared conversations by LMSYS
6	palm-2	1042	PaLM 2 for Chat (chat-bison@001) by Google
7	vicuna-7b	1007	a chat assistant fine-tuned from LLaMA on user-shared conversations by LMSYS

(Ref: The Complete Prompt Engineering for AI Bootcamp (2023))

Landscape of LLMs & Quiz

- ▶ Types of models - Foundation models, LLM, SLM, VLMs, etc.
- ▶ Common LLM terms - Prompts, Temperature, Hallucinations, Tokens, etc.
- ▶ LLM lifecycle stages - Pre-training, Supervised Fine Tuning, RLHF, etc.
- ▶ LLM evaluations - ROUGE, BLEU, BIG-bench, GLUE, etc.
- ▶ LLM architecture - Encoder, Decoder, Transformer, Attention, etc.
- ▶ Retrieval augmented generation - Vector DBs, Chunking, Evaluations, etc.
- ▶ LLM agents - Memory, Planning, ReAct, CoT, ToT, etc.
- ▶ Cost efficiency - GPU, PEFT, LoRA, Quantization, etc.
- ▶ LLM security - Prompt Injection, Data poisoning, etc.
- ▶ Deployment & inference - Pruning, Distillation, Flash Attention, etc.
- ▶ Platforms supporting LLMOps

(Ref: LinkedIn post by Abhinav Kimothi - 23 Jan 2024)



Thanks ...

- ▶ Search "**Yogesh Haribhau Kulkarni**" on Google and follow me on LinkedIn and Medium
- ▶ Office Hours: Saturdays, 2 to 5pm (IST); Free-Open to all; email for appointment.
- ▶ Email: yogeshkulkarni at yahoo dot com



(Generated by Hugging Face QR-code-AI-art-generator,
with prompt as "Follow me")