Generative AI

The talk of the town

Date : 21st June 2024 | Speaker : Ayon Roy |
Venue : KIET Group of Institutions, Uttar Pradesh, India

Visit - https://ayon-roy.netlify.app

Hello **World!**

I am **Ayon Roy**

Executive Data Scientist @ NielsenIQ

Z by HP Global Data Science Ambassador

- Mentored/Judged **100+** Hackathons
- Delivered **100+** Technical Talks
- Brought **Kaggle Days Meetup** Community in India for the 1st time

Scan to Connect

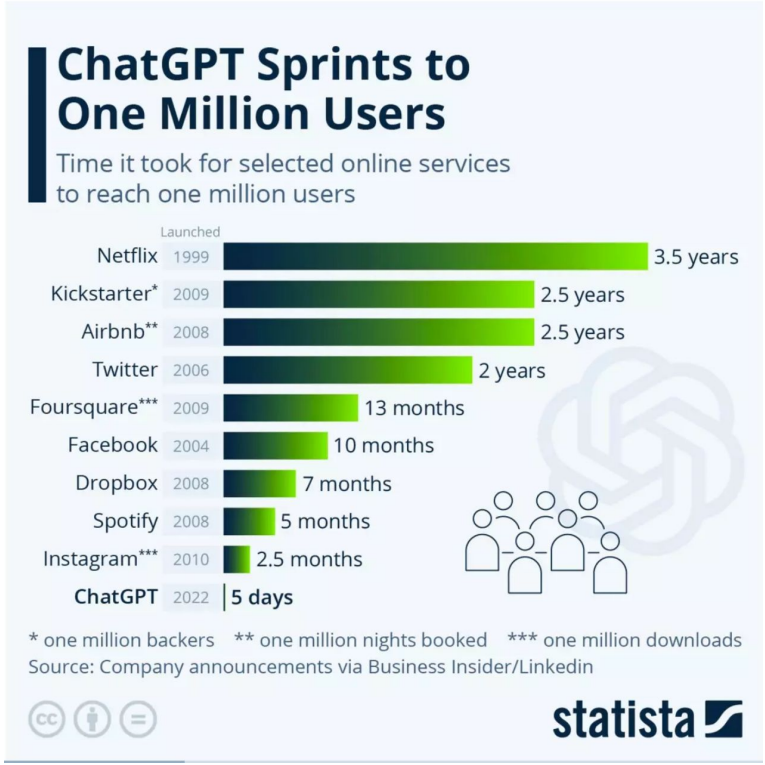
Visit - https://ayon-roy.netlify.app

Agenda

- Introduction to Generative AI (GenAI)
- How is GenAI different than Traditional AI ?
- How GenAI really works ?
- Impact of GenAI
- How to use GenAI effectively ?
- Challenges in GenAI

Generative AI

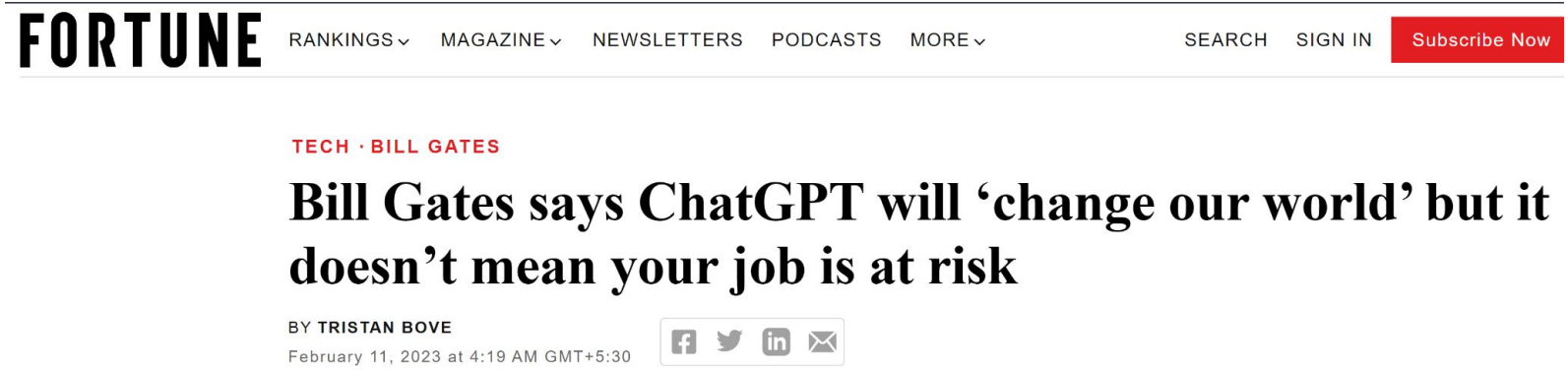
The Big Bang Moment of Generative AI



30% of outbound messages from business will be written by AI by 2025 (Gartner)

$110bn Generative AI market size by the end of the decade

Visit - https://ayon-roy.netlify.app



62% of consumers would prefer to use a customer service bot rather than waiting for a human agent to answer their queries.

Source : https://www.tidio.com/blog/chatbot-statistics/

A few popular Generative AI tools

ChatGPT by Open AI

Bard by Google

BLOOM by BigScience

Galactica AI by META

DALL-E2 by OpenAI

Imagen by Google

Stable Diffusion by Stability AI

MidJourney v4 by MidJourney

Vall-Eby Microsoft for speech

PointEby OpenAI for 3D objects

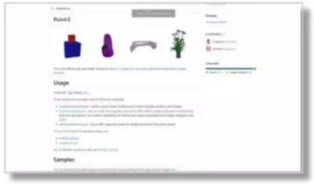


Imagen Video by Google

Chinchilla by Deep Mind for text

Visit - https://ayon-roy.netlify.app

Introduction to Generative AI

- Generative AI is a subset of artificial intelligence (AI) that involves creating models capable of generating new data or content.
- This is different from discriminative AI models that are trained to classify or predict existing data.
- Generative AI models can create new images, music, text, or other types of content that resemble or expand on the data they were trained on.

- Some popular generative AI models include Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and autoregressive models.

- The use of generative AI is growing rapidly across a variety of industries, including art, music, fashion, gaming, and healthcare.

- Generative AI has the potential to revolutionize the way we create and consume content, and can enable new forms of human-machine interaction.

How Generative AI is different from Traditional AI ?

Traditional AI	Generative AI
Rules-based system	Machine learning-based system
Pre-defined set of rules and algorithms	Learns from data and can generate new outputs
Limited ability to adapt to new situations or contexts	Can adapt to new situations and generate contextually relevant outputs
Primarily used for specific, well-defined tasks	Can be used for a wide range of tasks, including creative and open-ended ones
Requires human intervention and supervision	Can operate autonomously, with minimal human intervention
Focuses on accuracy and efficiency	Focuses on generating plausible and diverse outputs
Examples: Expert systems, rule-based chatbots	Examples: GANs, language models (e.g., GPT-3)

How Generative AI really works ?

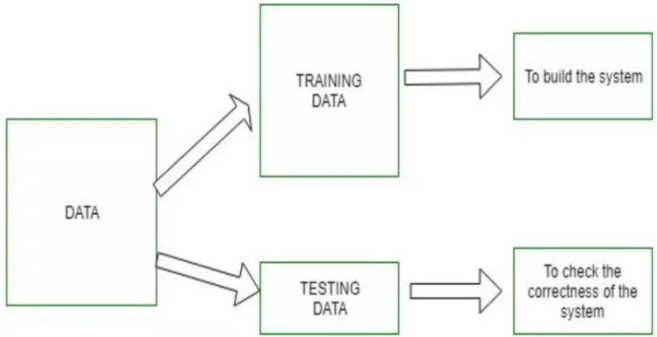
- Generative AI involves training models to learn patterns and capture the underlying structure of a given dataset.
- These trained models are then capable of generating new and original content that is similar to the training data but not an exact replica.
- Examples : Generative AI can be applied to various domains, such as image synthesis, text generation, music composition, and more, to create content that is realistic, diverse, and creative.

When you hear **“AI,”** think **“statistical pattern-matching”**

- Oracle describes AI this way:

AI has become a catchall term for applications that perform complex tasks that once required human input, such as communicating with customers online or playing chess.

*The term is often used interchangeably with … **machine learning (ML)** and deep learning.*



The data is **“tokenized”** *(= made into “chunks” of words, punctuation marks, pixels, etc.)* during this process - remember this for later

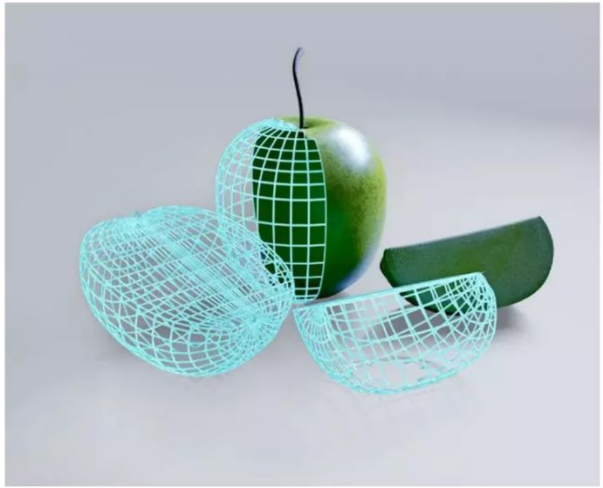
Text from What is Artificial Intelligence (AI)? Oracle, n.d. Retrieved May 16, 2023 from https://www.oracle.com/artificial-intelligence/what-is-ai/
Image from Pattern Recognition. GeeksforGeeks. Retrieved May 16, 2023 from https://www.geeksforgeeks.org/pattern-recognition-introduction/

Now, AI can synthesize *part* or *all* of a creative work

- McKinsey defines generative AI as:

  … Algorithms (such as ChatGPT) that can be used to create new content, including audio, code, images, text, simulations, and videos.

  Recent breakthroughs in the field have the potential to drastically change the way we approach content creation.



Text and image from What is generative AI? McKinsey. Retrieved May 16, 2023 from https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai

Foundational Basics of Generative AI tools

- **Training Data**: GenAI tools are trained on a massive corpus of data from various sources, including books, websites, and online forums. For eg : The training data for ChatGPT consisted of over 45 terabytes of text, making it one of the largest language models ever created.

- **Transformer Architecture**: GenAI tools at the core uses a transformer-based architecture, which is a type of deep learning model that excels at processing sequential data, such as natural language.

  The transformer architecture allows ChatGPT to learn the statistical patterns and structures of language and generate contextually relevant responses.

- **Feedback Loop**: GenAI tools continuously learn and improve their responses based on feedback from users. The model can learn from corrections, suggestions, and ratings provided by users to generate more accurate and relevant responses over time.

Simple probabilities (but trillions of them) are at the root of all large language models (LLMs)

Is it "aardvark"?

Or "birthday"?

The system generates text or images using its previously built model of the statistical distributions of **tokens** *(= "chunks" of words, punctuation marks, pixels, etc.)* created from its *very large* training dataset.

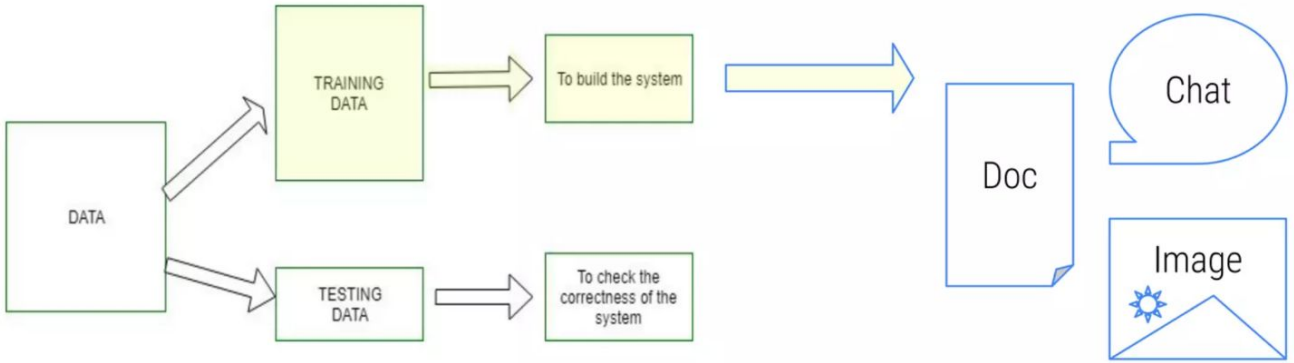
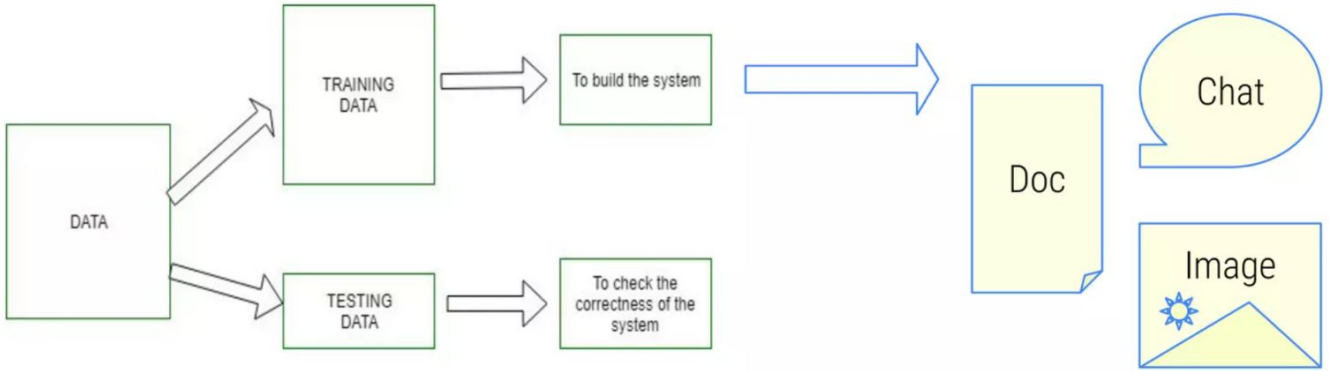


Image from Pattern Recognition. GeeksforGeeks. Retrieved May 16, 2023 from https://www.geeksforgeeks.org/pattern-recognition-introduction/
Murray Shanahan. 2022. Talking About Large Language Models. arXiv [cs.CL]. Retrieved from http://arxiv.org/abs/2212.03551
Bea Stollnitz. How generative language models work. Retrieved May 10, 2023 from https://bea.stollnitz.com/blog/how-gpt-works/

It might make mistakes or "hallucinate" based on the limitations of its process, but the output still might look like what you wanted.

Ted Chiang's analogy = "unreliable photocopier" or a "blurry JPEG"



Ted Chiang. 2023. ChatGPT Is a Blurry JPEG of the Web. *The New Yorker*. Retrieved May 10, 2023 from https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web
Murray Shanahan. 2022. Talking About Large Language Models. arXiv [cs.CL]. Retrieved from http://arxiv.org/abs/2212.03551
Bea Stollnitz. How generative language models work. Retrieved May 10, 2023 from https://bea.stollnitz.com/blog/how-gpt-works/

We can ask it questions - but a very specific type of question known as **prompts,** following this structure:

"Here's a fragment of text.

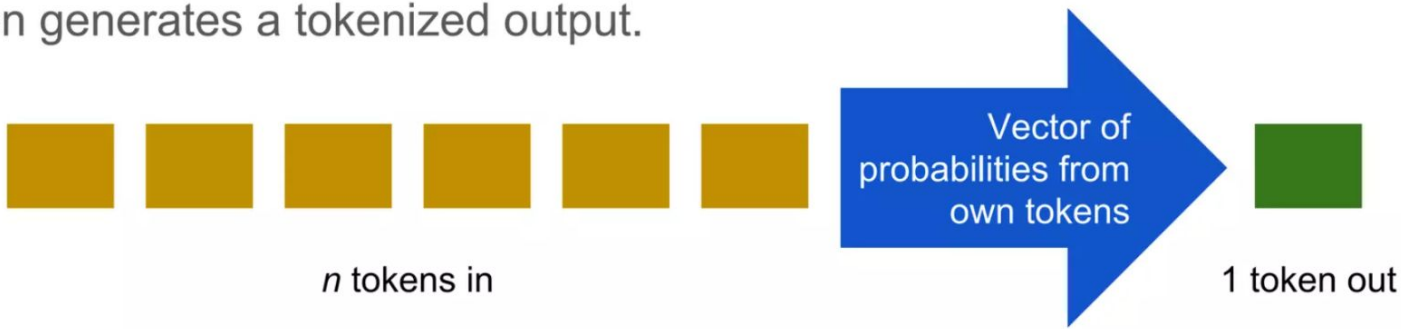
Tell me how this fragment might *<continue on in this language, or suggest a particular image>*.

According to your model of the statistics of *<human language, or human-handled images>*, what *<words, or pixels>* are likely to come next?"

Murray Shanahan. 2022. Talking About Large Language Models. arXiv [cs.CL]. Retrieved from http://arxiv.org/abs/2212.03551

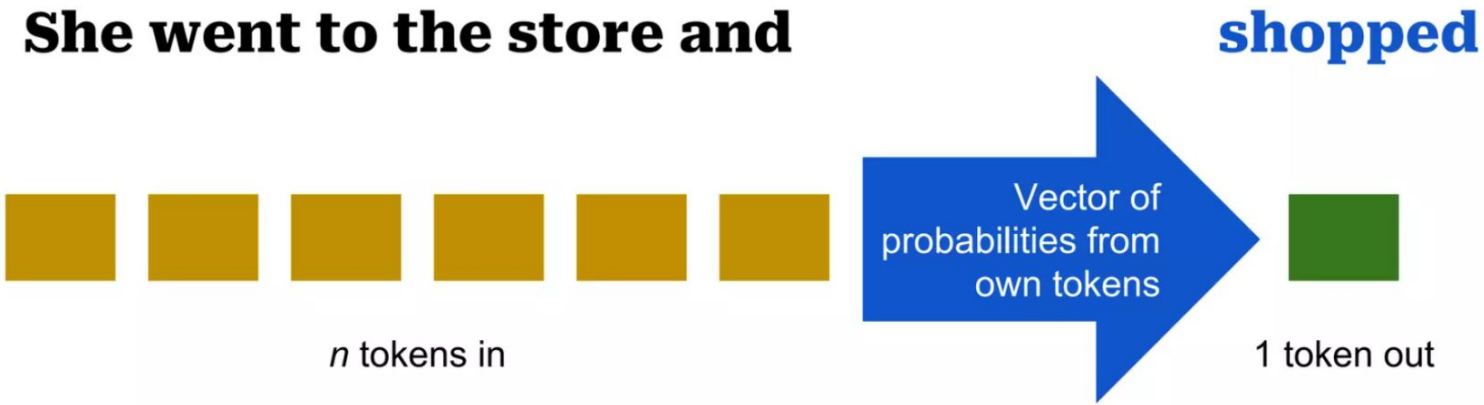
The prompts are converted into tokens *(= "chunks" of words, punctuation marks, pixels, etc.)*, then the system analyzes what is likely to come next, based on the tokens in its own dataset (as many as 32,000 in GPT-4!).

It then generates a tokenized output.



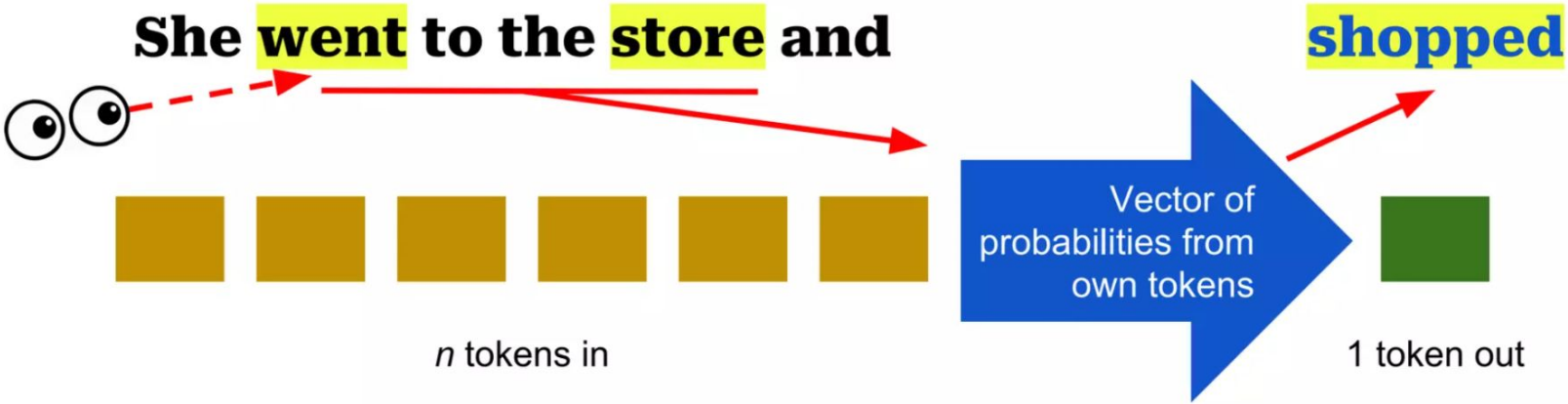
Murray Shanahan. 2022. Talking About Large Language Models. arXiv [cs.CL]. Retrieved from http://arxiv.org/abs/2212.03551
Bea Stollnitz. How generative language models work. Retrieved May 10, 2023 from https://bea.stollnitz.com/blog/how-gpt-works/

With each output, it keeps re-analyzing the probabilities to decide next tokens.



Murray Shanahan. 2022. Talking About Large Language Models. arXiv [cs.CL]. Retrieved from http://arxiv.org/abs/2212.03551
Bea Stollnitz. How generative language models work. Retrieved May 10, 2023 from https://bea.stollnitz.com/blog/how-gpt-works/

Transformers (the “T in “GPT”) know how to **direct attention to specific parts of the input** to guide their selection of the output - such as verb tenses, objects.



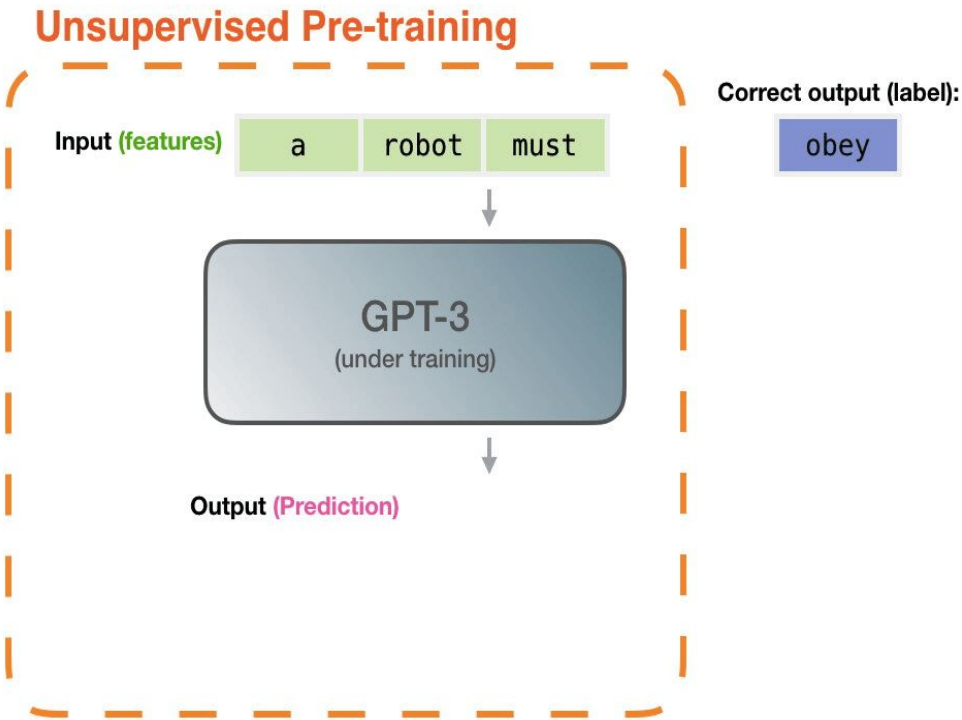
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. arXiv [cs.CL]. Retrieved from http://arxiv.org/abs/1706.03762
Bea Stollnitz. How generative language models work. Retrieved May 10, 2023 from https://bea.stollnitz.com/blog/how-gpt-works/

The system can give you different answers to the same inputs:

Murray Shanahan. 2022. Talking About Large Language Models. arXiv [cs.CL]. Retrieved from http://arxiv.org/abs/2212.03551
Bea Stollnitz. How generative language models work. Retrieved May 10, 2023 from https://bea.stollnitz.com/blog/how-gpt-works/

“Hallucinations” - when the output doesn’t seem to make sense - are why it is important not to accept everything it outputs at face value.

Murray Shanahan. 2022. Talking About Large Language Models. arXiv [cs.CL]. Retrieved from http://arxiv.org/abs/2212.03551
Bea Stollnitz. How generative language models work. Retrieved May 10, 2023 from https://bea.stollnitz.com/blog/how-gpt-works/

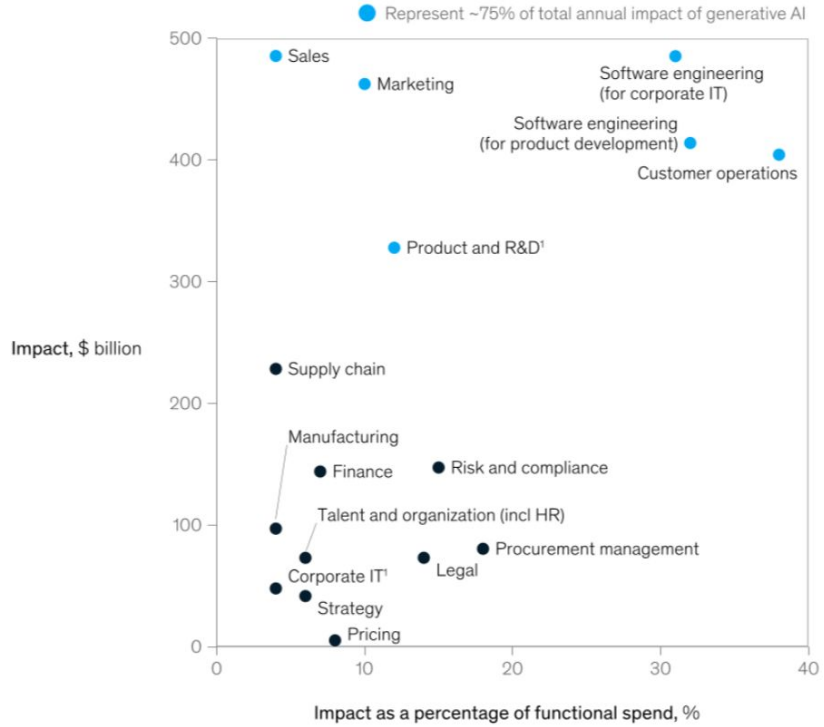

Unsupervised Pre-training
Input (features)
a
robot
must
Correct output (label):
obey
GPT-3
(under training)
Output (Prediction)

Various Generative AI tools

- https://www.artbreeder.com/ : An online platform that allows users to create and manipulate images using GAN technology. Users can mix and match different styles and genres to generate unique and surrealistic images. [
- https://www.aiva.ai/ : An AI-powered music composition platform that allows users to create original music tracks in a variety of styles and genres. AIVA uses deep learning algorithms to analyze and mimic the musical patterns and structures of different composers.
- https://runwayml.com/ : A cloud-based platform that offers a range of generative AI tools for various creative applications, including image and video synthesis, text-to-image conversion, and style transfer.
- https://tryellie.com/ : Email Reply Assistant
- https://fireflies.ai/ : Meeting Notes Transcriber & Analyzer

Impact of Generative AI

Using generative AI in just a few functions could drive most of the technology's impact across potential corporate use cases.



Note: Impact is averaged.
[1]Excluding software engineering.
Source: Comparative Industry Service (CIS), IHS Markit; Oxford Economics; McKinsey Corporate and Business Functions database; McKinsey Manufacturing and Supply Chain 360; McKinsey Sales Navigator; Ignite, a McKinsey database; McKinsey analysis

McKinsey & Company

Visit - https://ayon-roy.netlify.app

Domain wise Generative AI impact

Generative AI use cases will have different impacts on business functions across industries.

Generative AI productivity impact by business functions[1]

Low impact — High impact

	Total, % of industry revenue	**Total,** $ billion	Marketing and sales	Customer operations	Product and R&D	Software engineering	Supply chain and operations	Risk and legal	Strategy and finance	Corporate IT[2]	Talent and organization
			760–1,200	340–470	230–420	580–1,200	280–530	180–260	120–260	40–50	60–90
Administrative and professional services	**0.9–1.4**	150–250									
Advanced electronics and semiconductors	**1.3–2.3**	100–170									
Advanced manufacturing[3]	**1.4–2.4**	170–290									
Agriculture	**0.6–1.0**	40–70									
Banking	**2.8–4.7**	200–340									
Basic materials	**0.7– 1.2**	120–200									
Chemical	**0.8–1.3**	80–140									
Construction	**0.7–1.2**	90–150									
Consumer packaged goods	**1.4–2.3**	160–270									
Education	**2.2–4.0**	120–230									
Energy	**1.0– 1.6**	150–240									
Healthcare	**1.8–3.2**	150–260									
High tech	**4.8–9.3**	240–460									
Insurance	**1.8– 2.8**	50–70									
Media and entertainment	**1.5– 2.6**	60–110									
Pharmaceuticals and medical products	**2.6–4.5**	60–110									
Public and social sector	**0.5–0.9**	70–110									
Real estate	**1.0–1.7**	110–180									
Retail[4]	**1.2–1.9**	240–390									
Telecommunications	**2.3–3.7**	60–100									
Travel, transport, and logistics	**1.2–2.0**	180–300									
		2,600–4,400									

Note: Figures may not sum to 100%, because of rounding.
[1]Excludes implementation costs (eg, training, licenses).
[2]Excluding software engineering.
[3]Includes aerospace, defense, and auto manufacturing.
[4]Including auto retail.
Source: Comparative Industry Service (CIS), IHS Markit; Oxford Economics; McKinsey Corporate and Business Functions database; McKinsey Manufacturing and Supply Chain 360; McKinsey Sales Navigator; Ignite, a McKinsey database; McKinsey analysis

McKinsey & Company

Visit - https://ayon-roy.netlify.app

What to do now ?

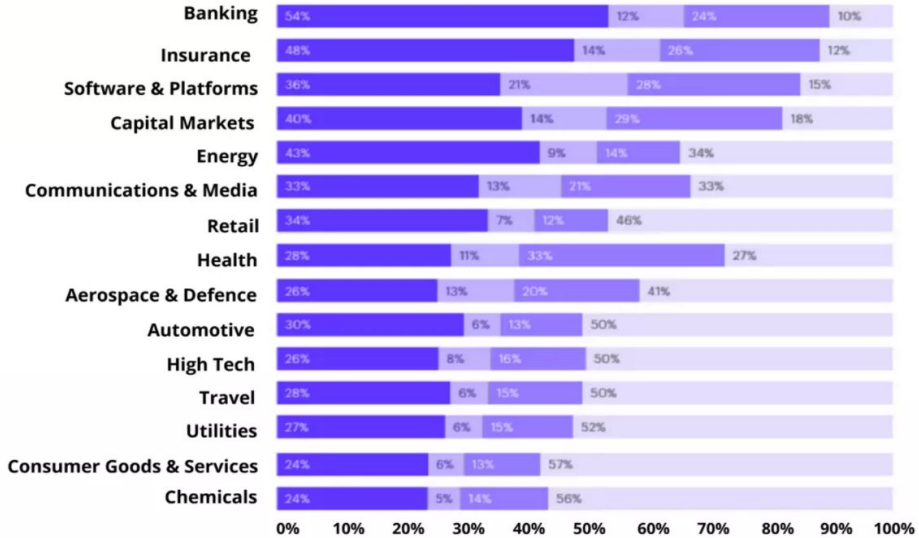
Companies must reinvent work to find a path to generative AI value. Business leaders must lead the change, starting now, in job redesign, task redesign and reskilling people.

Nearly 6 in 10 organisations plan to use ChatGPT for learning purposes and over half are planning pilot cases in 2023. Over 4 in 10 want to make a large investment.

40% of working hours across industries can be impacted by Large Language Models (LLMs)

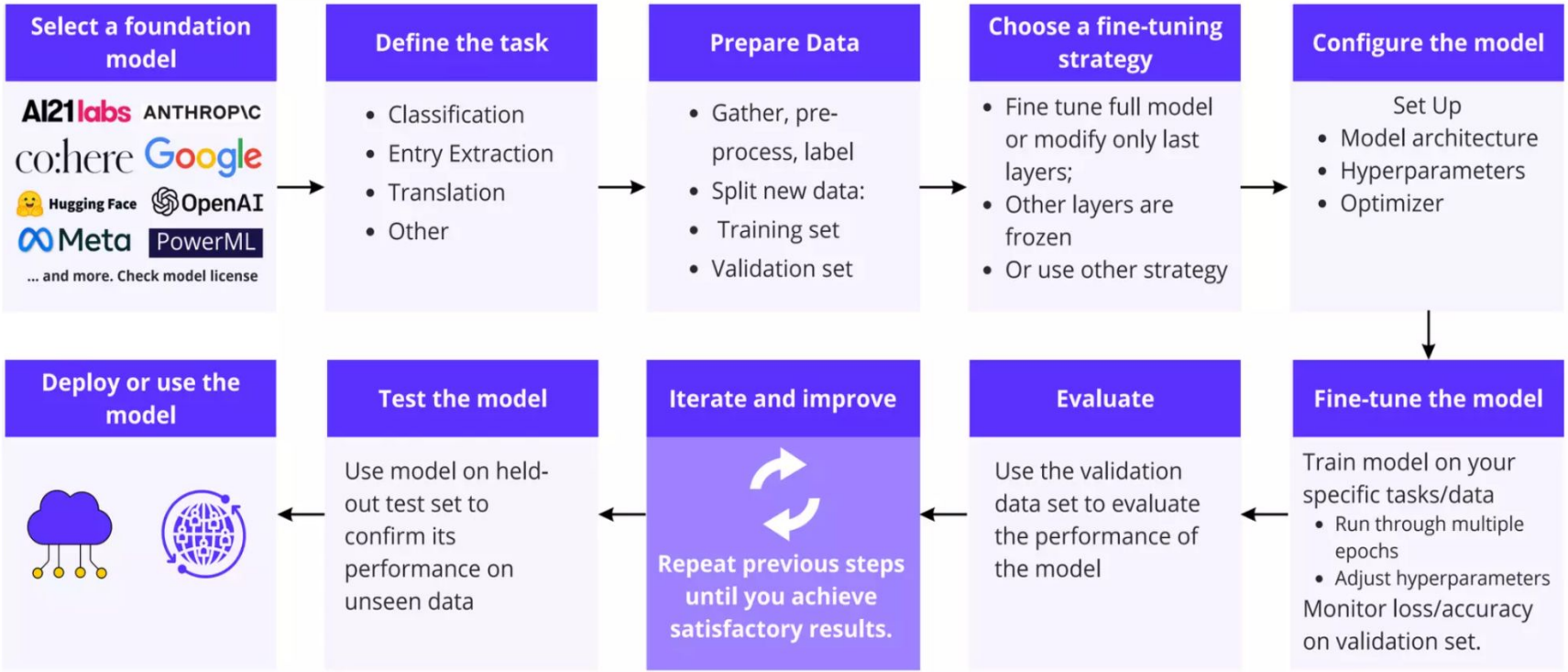
Why is this the case? Language tasks account for 62% of total worked time in the US. Of the overall share of language tasks, 65% have high potential to be automated or augmented by LLMs.

katonic.ai



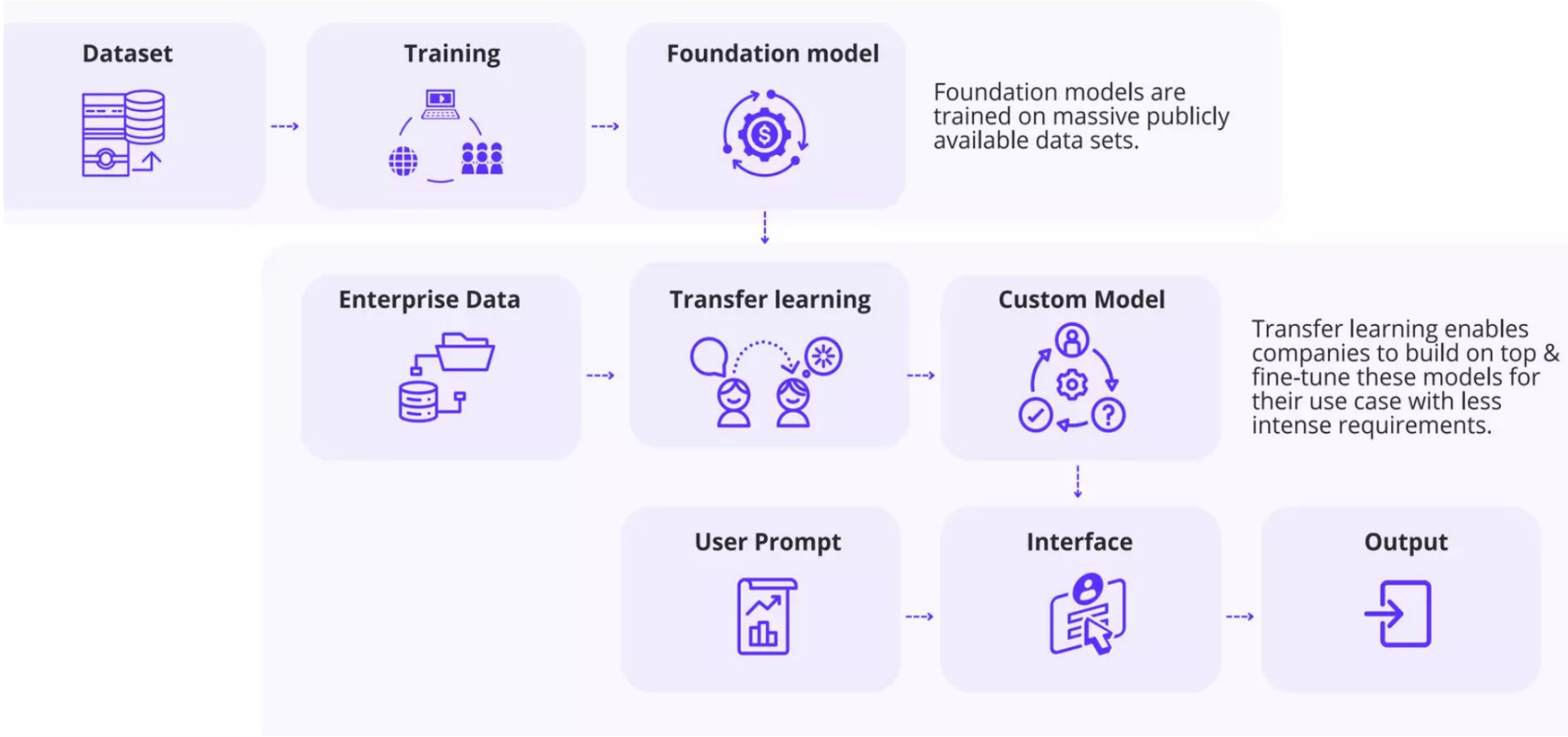
Visit - https://ayon-roy.netlify.app

How to use Generative AI models effectively ?



Visit - https://ayon-roy.netlify.app

Fine Tuning Generative AI models



Visit - https://ayon-roy.netlify.app

Options to use Generative AI

	Option	Explanation	Cost
Off the Shelf	*Subscribe and Use*	Embracing off-the-shelf tools leveraging LLMs that are already available.	Minimal
Customise	*Consume with Guardrails*	Build Guard Rails by adding pre and post-processing restrictions to off the shelf LLM's	$
	Augment	Use database lookups to tailor LLMs to an organisation's needs.	$$
	Fine Tune	Using fine-tuning to tailor LLMs to an organisation's needs	$$$
Train your own	*Build your own*	Build and Train your model from scratch with your data	$$$$

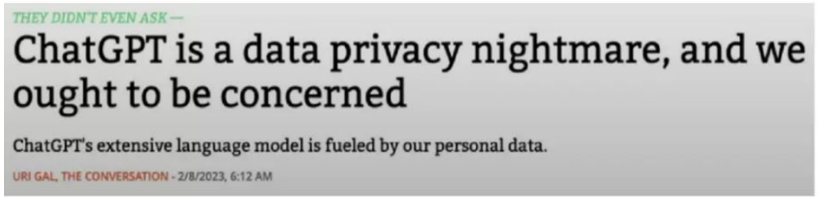
Potential Opportunities

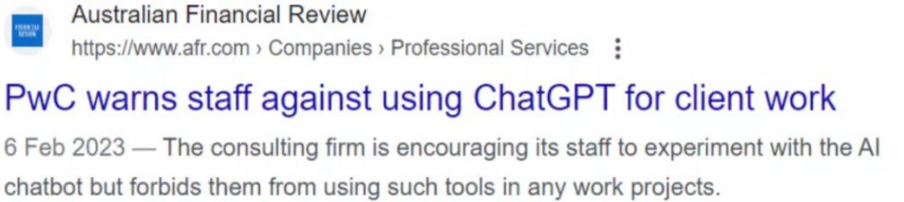
- **Finance & Legal:** Draft and review documents, patents and contracts; find, summarise and highlight important points in regulatory documents; find and answer specific queries from large documents; scan through historical data to recommend a course of action.
- **Marketing & Sales:** Automate SEO-optimised content generation, enhance ad bids, hyper-personalise communication and deployment, create product user guides by persona, analyse & segment customer feedback, hyper-capable chatbots for upsell and cross-sell .
- **Customer service:** Natural-sounding, personalised chatbots and virtual assistants can handle customer inquiries, recommend swift resolution, and guide customers to the information they need.
- **HR & Recruitment:** Smart-shortlist of candidates, risk assessment of candidates, self-service of HR functions via chatbots and automation
- **Information technology:** Advanced code writing code and documentation, code review and error detection, and accelerated software development, auto-complete data tables, generate synthetic data.

Challenges in Generative AI

- Faces a challenge where the AI model may produce limited and repetitive output, lacking diversity and creativity.
- The training process of generative models can be sensitive to settings and may require careful adjustments. It can sometimes be challenging to achieve stable and consistent results.
- Generative AI models can unintentionally reflect biases present in the training data, potentially resulting in biased or unfair output. Addressing these biases and ensuring ethical use of AI is an important challenge.
- Training and utilizing generative models often require significant computational power and resources, which can pose challenges for individuals or organizations with limited access to such resources.
- Understanding how and why generative AI models generate specific content can be difficult. These models often lack transparency and explanations, making it challenging to trust and interpret their output.

Data Security, Privacy & Cost aspect of Generative AI





Visit - https://ayon-roy.netlify.app

- Violations of intellectual property
  - Check the Terms of Service - will your inputs or prompts be used as training data?
- Violations of academic integrity
  - Do a spot check of outputs, using a search engine, to see if any are wholly from another work
  - Analyze submitted work using Open AI's AI Text Classifier or the multi-service GPTZero

- Generative AI tools are great for PRODUCTIVITY - they can be nifty shortcuts to dispose of low-value tasks and / or to jumpstart creativity
- Generative AI tools should always be used - *and taught to be used* - with a critical mind, because they are prone to mistakes and "hallucinations"

Is your job in danger ?

Visit - https://ayon-roy.netlify.app

Humans' #1 skill set will continue to be *communication*

Screenshot from
https://twitter.com/TheRealOllieLaw/status/1656605938374307840?s=20

Let me answer your Questions now.

Finally, it’s your time to speak.

Visit - https://ayon-roy.netlify.app

Danke Schoen

Questions ? Any Feedbacks ? Did you like the talk?
Tell me about it.

If you think I can help you,
connect with me via

Email : ayon-roy@outlook.com

LinkedIn : https://www.linkedin.com/in/ayon-roy

Website : https://AYON-ROY.NETLIFY.APP/

Scan to Download the slides