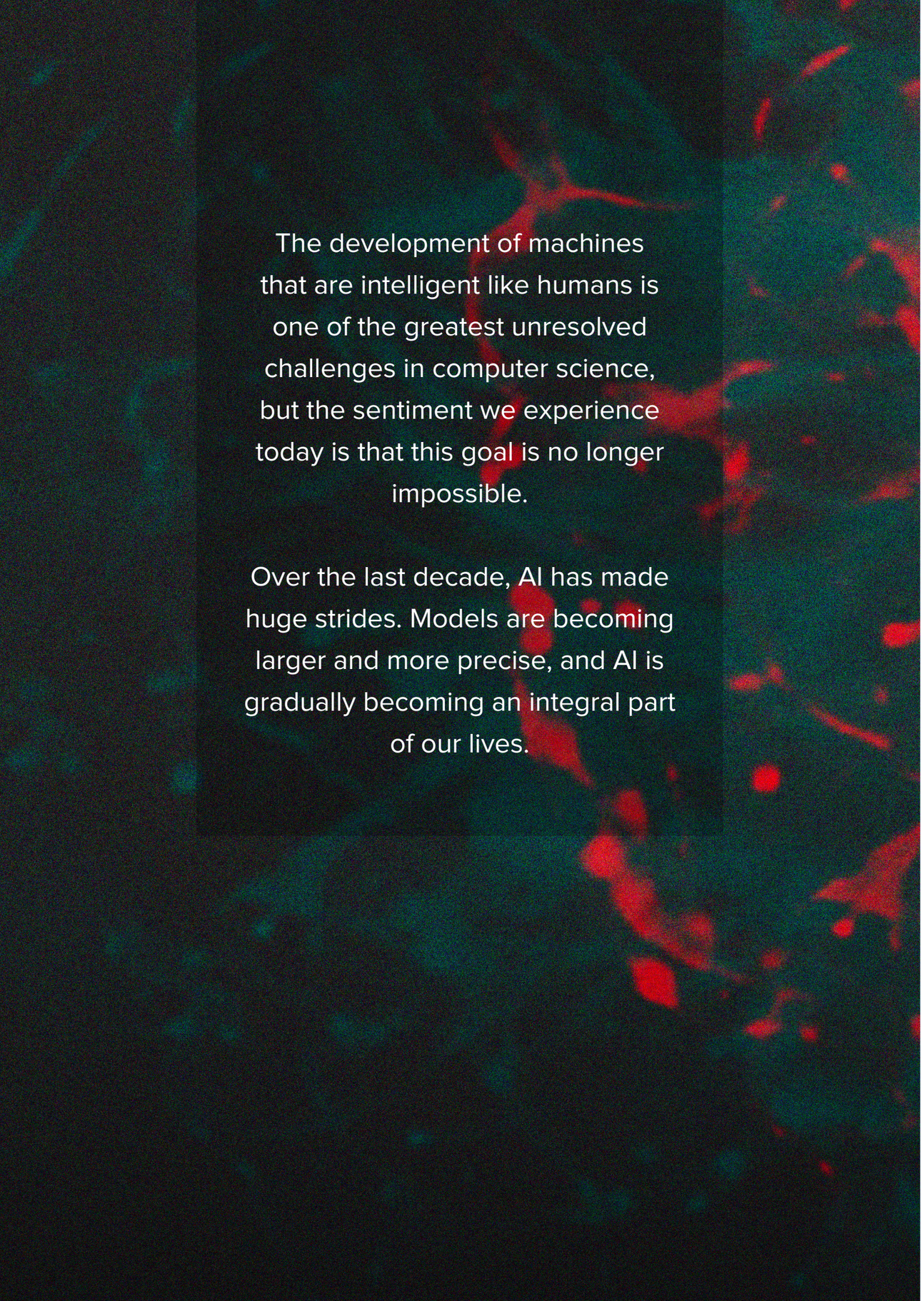
GPT-3: UNDERSTANDING ITS POTENTIAL

Autoregressive language models have become a dominant trend in the general Artificial Intelligence (AI) field. OpenAI's Generative Pre-trained Transformer 3 (GPT-3) leverages deep learning to perform diverse tasks. Reply is actively researching GPT-3's potential use cases across different fields to verify their effectiveness in real-world contexts.



WHAT IS GENERAL AI?

There is no universal definition of Artificial General Intelligence (AGI). This is because the scientific community has not yet been able to come to a consensus on a unified understanding of the concept of general intelligence.

Demis Hassabis, Co-founder and CEO of DeepMind, one of the leading authorities in this field, has stated that AGI is characterized by two key concepts: ‘learning’ and ‘general’.

LEARNING

AGI focuses on learning algorithms capable of mastering tasks from raw data. Therefore, the main feature of this system is that it can learn by itself.

GENERALITY

The system must operate over a wide range of tasks, including novel tasks, potentially right out of the box.

So, what would a system with these attributes look like? We have an example of such a system right in front of us: the human brain. In a nutshell, AGI systems are systems that possess intelligence equal to, or greater than, that of humans.

AGI technologies are based on complex networks with millions of parameters. Despite being frequently associated with science fiction, there are a number of existing applications, ranging from image generation to Natural Language Processing (NLP). Some of the most renowned models include:

AI MODEL	TRAINABLE PARAMETERS	AREA	IP OWNER
BERT	110 Million	NLP	Google
BIG GAN	1.75 Billion	CV	DeepMind
GATO	1.2 Billion	RL	DeepMind
GPT-3	175 Billion	NLP	OpenAI
BLOOM	176 Billion	NLP	Open source
WUDAO 2.0	1.75 Trillion	NLP	Beijing Academy of AI

Obtained from The London AI Summit 2022

These technologies are based on transformer architectures. A transformer is a deep learning model that mimics cognitive attention; in other words, the model focuses more on the most important parts of the input data, diminishing the importance of other parts. Additionally, transformers allow for parallelized training, meaning they can process more words at a time, reducing training times.

Transformers were introduced by Google Brain in 2017. Their architecture is divided into encoders and decoders.

ENCODERS

Encoders take an input and encode it with attention information (which mainly concerns the position of the words in a sentence).

DECODERS

Decoders ensure the output is focused on the relevant words from the input, keeping it relevant and coherent.

This architecture has become a well-established standard in deep learning frameworks, having been implemented in TensorFlow and PyTorch. Additionally, it has demonstrated great success in the field of NLP, where it is rapidly gaining traction as the model of choice for developers.

Natural Language Processing (NLP) is a well-established subfield of Artificial Intelligence, with the aim of building systems capable of “understanding” human language. This entails not only recognizing the individual meanings of words, but also the contextual nuances of the language they are used in. This allows the model to provide useful insights, categorise documents, extract key details, and much more.

Many current AI applications are driven by NLP, such as sentiment analysis, text summarisation, and chatbots, which have seen growing adoption.

25%

Of companies will primarily rely on chatbots for their customer services by 2027

Gartner

Most NLP applications require a key component: a Language Model. These models can be seen as statistical prediction engines, such as the text auto-completion feature found on most smartphones.

This feature tries to predict the next word a user will type, based on a statistical analysis of existing text sequences. When these models have many parameters and are trained on very large datasets, they are referred to as Large Language Models (LLMs).

In the last few years, LLM sizes has increased by

10x

NVIDIA

As language models (LLMs) become more complex and larger, so do their applications. A significant increase in the number of parameters used to train the model results in drastic changes in its behavior. Consequently, LLMs are currently a major development in AI, with the potential to revolutionize entire industries.

Generative Pre-Trained Transformer (GPT-3) is the third-generation AI language model released by OpenAI in 2020. It is their most advanced model available for business implementations.

Although language models are usually designed to carry out one task, GPT-3 was the first LLM to demonstrate impressive capabilities in various NLP tasks.

GPT-3 consists of

175 bn

parameters

GPT-3 offers a remarkably large number of applications, such as chatbots, search, text summarisation, and code generation. However, some of its capabilities are likely to remain unknown until more users take advantage of the technology.

Thus, it is difficult to estimate the impact that its uses (and potential misuses) may have on the economy.

WHAT IS GPT-3?

In order to understand GPT-3, it is key to understand the key words that it stands for: **“Generative Pre-Trained Transformer”**.

A GENERATIVE MODEL...

This feature is related to its ability to generate text. GPT-3 utilises statistical modelling to construct its output text. It encodes the probability of various outputs before producing the result. The main principle behind Generative Models is to generate new data similar to existing data.

A PRE-TRAINED MODEL...

GPT-3 has already been trained on four substantial databases: Common Crawl, which contains petabytes of data from web pages and other text from web crawling; WebText2, which was created by scraping high-quality web pages; Books1 and Books2, which are comprised of tens of thousands of books; and Wikipedia.
The result is a model with 175 billion parameters.

A TRANSFORMER MODEL...

This architecture enables GPT-3 to “pay attention” in a similar manner as humans do. Transformers are leading-edge in NLP, allowing the model to filter out extraneous information and focus on the relevant words based on probabilities. GPT-3 is a decoder-only architecture, which has demonstrated superior performance on long sequences of text compared to classic models.

These characteristics make GPT-3 one of the most straightforward models available for building AI applications.

THE POTENTIAL WITHIN GPT-3

GPT-3 has demonstrated its potential across a wide range of fields. Below, we explore the four main reasons why organisations should adopt GPT-3 to begin their journey towards process automation, thus adding value to their businesses.

BUSINESS VALUE

Large language models aim to extend the capabilities of systems when working with text, adding value in innovative ways that were previously inconceivable.

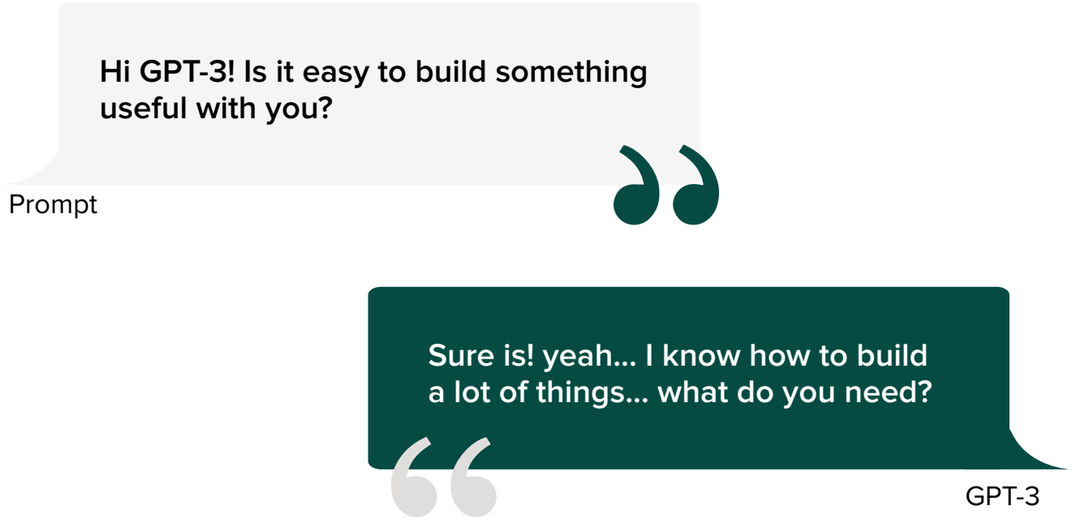
These models, trained on immense text corpora of data, are able to adapt to countless domains of application. The **interdisciplinarity** inherent in this family of models, to which GPT-3 belongs, allows it to create simple stories for children, but also reformat text for a scientific publication or suggest hints for professional developers.

For tasks such as classification and regression, we can exploit some intrinsic properties of the model with no need to fine-tune it. It is enough to engineer the task properly and ask the model in a zero-shot setting. GPT-3 is **ready** for any use case that requires some level of cognitive skill. Simple proof-of-concepts can be realized and validated within minutes.

The ease with which GPT-3 is **accessible** has enabled its wide adoption. The API release created a paradigm shift in NLP, attracting a large number of beta testers. This has spurred innovations and start-ups at a rapid rate, with commentators labeling it the “Fifth Industrial Revolution”.

Although Large Language Models require considerable resources for training, their structure makes them highly efficient for inference. Solutions based on GPT-3 are designed to **scale** up as required. Moreover, the affordability and prevalence of these technologies has enabled the market to become more competitive, offering GPT-3-based solutions at increasingly lower costs.

Despite the above, GPT-3 still has some limitations, such as being primarily trained on English-language text and content found on the internet. The research community is actively addressing these issues, and numerous alternative solutions have been proposed to address these limitations.



Emphasizing all these properties, the research world is uncovering new potentials of GPT-3 every day by identifying new features and minimizing the computational impact of the model. GPT-3 and its subsequent versions promise further amazing results to be utilized for scenarios that have never been considered before!

GPT-3 IN THE MARKET TODAY

GPT-3's ease of use and accessibility has led to remarkably fast adoption. According to OpenAI, within the first nine months of its release, more than 300 businesses were built around it.
It is no surprise then that there are a number of successful companies that have been created around it.

MessageBird

MessageBird focuses on Telco APIs, providing an omni-channel communications platform that supports calls, messages, and notifications. Leveraging GPT-3 for reliability, the product enhances all interactions, from personalised support to custom notifications.

> 25k clients
Including Google, Facebook, Uber and Adobe

$1 Bn funding
Over 3 billion valuation

algolia

Algolia's mission is to "empower every company to create delightful Search & Discovery experiences". It utilizes GPT-3 to identify what a user is searching for within a company's product catalogue and serve the most accurate results.

> 10k clients
including Slack, Stripe, Zendesk and Lacoste

$2.25 Bn
Over 1 billion valuation

Debuild

Debuild leverages a 'seemingly' concealed capability of GPT-3; its capacity to generate high-quality code. It is a low-code tool that can assist you in constructing a web application from a straightforward description. It can generate React components and SQL code with a user-friendly UI in a matter of seconds.

Currently generating revenue

10 investors
Including the CEO of Repl.it and the CTO of Github

viable

Viable is a tool designed to extract relevant outcomes from user feedback. Utilizing AI and GPT-3, this product automates qualitative data analysis without sacrificing quality, allowing companies to gain insights into their customers with minimal effort.

$8.9 M
Over 5 million in funding

4 use cases
Product Management, Marketing, Customer Experience and Customer Research

FABLE STUDIO VirtualBeings

FableStudio has developed a product called VirtualBeings, an AI-generated animated avatar that utilizes GPT-3 for its dialogue generation. This technology aims to bring people's favorite characters to life, allowing for interactions in a "human-like" manner.

Premiered at the Sundance, Cannes and Tribeca Film Festivals

Emmy-Award
For Outstanding Achievement in Interactive Media

Replier.ai

Replier.ai addresses an additional issue arising from receiving user feedback: the responses. Replier generates responses for customer reviews with an exceptional degree of quality, thanks to GPT-3's attention to detail.

Acquired
By Tailwind (leading small business marketing software platform)

testing.ai

AB Testing is a product designed to optimise the results of your landing page. It combines GPT-3 with statistical analysis methods to generate text variants and determine which one drives the most user interactions, as well as providing a justification.

7
Major clients, including Toyota, Suzuki and Movistar

Epsilon Code

Epsilon Code provides a valuable resource for Python developers, utilizing GPT-3 to generate Python code from plain-text descriptions. It also provides debugging assistance, similar to the help available on Stack Overflow.

Open-source
Available to test in GitHub

Additionally, established companies are also designing internal applications of GPT-3, such as Slack with Grok, which summarises a day's messages. The applications of GPT-3 developed by these companies range from text summarisation to film dialogue generation, highlighting its immense potential and economic feasibility.

REPLY'S GPT-3 USE CASES

Reply is actively exploring the use of GPT-3 for several use cases.

A. SENTIMENT ANALYSIS

This use case aimed to assess GPT-3's ability to extract a customer's sentiment towards a business.

Technical Challenge
Feedback systems, such as reviews on e-commerce websites, are valuable sources of truth to guide other customers in making satisfactory purchase choices. Reviews can be invaluable for sellers to understand their positioning in the market and to improve their offerings. However, the large number of reviews and the accuracy of the classification are key parameters to measure the success of the implementation.

Results
Sentiment Accuracy Score: 100%

Based on a test set of 100 amazon product reviews, aligning the sentiment with the Amazon Customer Stars Rating: Negative (1 & 2 stars), Neutral (3 stars), Positive (4 & 5 stars).

B. STRUCTURING DATA

The focus of this implementation is to unlock the potential of unstructured data, particularly emails, by utilizing GPT-3 to extract essential fields from the data.

Technical Challenge
Extracting relevant information from unstructured data can be a very time-consuming challenge and difficult to automate; as it would require the algorithm to understand the context and be able to adapt to the various ways in which the information can be expressed. Emails are an example of unstructured data, where it can be challenging to identify the sender and the single or multiple

receivers, and one would need to review the content of the email to identify them. Organisations have amassed large collections of unstructured data, such as emails, which can contain valuable information. GPT-3 can be used to query a business's unstructured text data with minimal labelled data due to its domain understanding and semantic search capabilities.

Results
Accuracy Score: 87%
Field identification Accuracy Score: 100%

Based on a test set of over 200 emails from the Enron email dataset (Kaggle), extracting the relevant fields from an unstructured email record.

C. EMAIL INFORMATION CAPTURING

GPT-3's summarization capabilities can be used to streamline project management processes. This application is designed to test GPT-3's ability to automate the creation of status reports.

Technical Challenge
In any project, most status information, such as urgent tasks, is typically exchanged via email. Project Managers tend to capture all relevant information to track progress and present it in status report sessions; however, this can be a time-consuming process and urgent emails may be left unread in inboxes for extended periods of time. The introduction of GPT-3 could optimise this process by extracting vital and time-sensitive information from emails in order to update reports and send additional notifications.

Results
Accuracy Score: 86%

Based on a test set of over 50 emails from the Enron email dataset (Kaggle), extracting the relevant actionable items mentioned.

The implementations leveraged the capabilities of GPT-3 for domain understanding and semantic search, allowing for high-quality summarisation of text data, reliable structuring of data, and accurate sentiment analysis. However, the interaction with an API was noted to increase latency, making GPT-3 less suitable for low-latency systems.

The most important development step focused on **prompt engineering**. GPT-3 displayed outstanding results with Few-Shot learning, being able to complete most tasks with One-Shot or Zero-Shot learning. All implementations used either Few-Shot or One-Shot learning, making the approach highly flexible and adaptable.

In conclusion, the technical implementations demonstrate that GPT-3 is a viable engine for each of the three use cases, highlighting its versatility.

IS GPT-3 ETHICAL?

It is essential to pause and reflect on the ethical implications of any technology before applying it to practical applications.

This section seeks to examine three critical implications of using GPT-3: model bias, misinformation proliferation, and environmental repercussions.

MODEL BIAS

GPT-3 has over 175 billion trainable parameters, thus requiring an extremely large training dataset.

The only place to find such volumes of natural language data is the internet, which, unfortunately, is rife with toxic biases, including gender, race, and religious prejudices, which could translate into the text produced by GPT-3.

OpenAI recognises the dangers of this bias and has invested heavily into anti-bias measures. An example is Process for Adapting Language Models to Society (PALMS), which fine-tunes models on a curated dataset of fewer than one hundred examples.

MISINFORMATION PROLIFERATION

When examining the remarkably lifelike material generated by GPT-3, it is easy to overlook that a human is not writing it. Models like GPT-3 lack the ability to discern if the material it creates is factually accurate or not; its default source is the Internet. This enables malicious actors to swiftly create believable misinformation, based on their prompts. This could be leveraged for nefarious activities such as spreading propaganda and committing exam fraud.

OpenAI is addressing this issue by providing a controlled release of GPT-3 to a select group of individuals. They may also impose strict API restrictions if they choose to launch a commercial product.

ENVIRONMENTAL IMPACT

The cost of training large models is known to increase exponentially with size. An estimate suggests that an energy input of 1.86^13 Watts would be needed to train a model of GPT-3's size. To put this into perspective, this would be equivalent to the energy used to power 1.72 million homes for a year, assuming each home requires an average of 900 kWh per month.

While this is a considerable energy usage, it has an advantageous side: the energy cost is a deterrent for malicious actors from training similar models to GPT-3, thus enhancing security. Additionally, it is important to consider that while training requires a lot of energy, generating predictions is very energy-efficient. For example, GPT-3 can generate 100 pages of text using only 0.4 kWh, indicating its potential for scalability once trained.

HOW CAN I USE IT?

GUIDELINES

GPT-3 is one of the largest and most advanced language models available, and can be used through a simple text interface. Even when faced with the most diverse tasks, we can reduce them to text form and explicitly ask GPT-3 for a possible solution. If we have a Dataset at our disposal, we can use it for fine-tuning the model and create an ad hoc process. If we have only a few examples, or no examples at all, techniques such as Few-shot Learning and Zero-shot Learning are available.

Zero Shot Learning - Predict unseen or new results without any training examples.
Few Shot Learning - Predict unseen or new results with few training examples.

PROMPT DESIGN

Prompt Design (Prompt Engineering) is the process of creating and defining the model input. In the case of GPT-3, the task description is embedded in the input, such as a question or sentence to be completed, rather than being given implicitly.

TASK - Find the subject of the sentence

PROMPT TEMPLATE - *What is the subject of the sentence "_____"?*

EXAMPLE - Alice was beginning to get very tired of sitting by her sister on the bank.

ASSESSING GPT-3

GPT-3 represents a new generation of NLP technologies that utilizes pre-trained models and fine-tuning for specific use cases. This approach streamlines model development, allowing for faster and more economic benefits.
We evaluated GPT-3 using Reply use case implementations, making a comparison of its performance against NLP benchmarks across four standard NLP tasks.

ENTITY RECOGNITION	In order to address unstructured data, there are various NLP models developed to identify and categorize named entities. However, due to the complexity of the task, traditional models necessitate large amounts of training. GPT-3 provides a useful shortcut for development, and more importantly, it is successful in data extraction.
TEXT CLASSIFICATION	When categorizing text into organised groups, GPT-3 matches the performance of state-of-the-art technologies specifically designed for that task. As long as the prompt provided to GPT-3 is configured correctly, the model can perform both zero-shot classification (without any prior fine-tuning) and few-shot classification (with contextual information).
TEXT GENERATION	The key characteristics that tend to identify computer-generated text are a lack of creativity, complexity, repetitions, and predictability. GPT-3 represents a significant breakthrough in the field of text generation. Its training dataset enables it to handle intricate conversations and tasks, producing text that is almost indistinguishable from text written by humans.
TEXT SUMMARISATION	NLP models have high requirements for understanding documents and identifying the most important sections. Likewise, GPT-3 can reduce the implementation time of these solutions and the benefits associated with entity identification. Its abstractive summarisation capability allows it to create its own summary rather than extracting key information. Moreover, the output can be adapted through rapid engineering to achieve the desired result.

USING GPT-3

GPT-3 was primarily created in Python, but thanks to OpenAI's GPT-3 API, it can be used with any programming language. This API runs on Azure, having received a billion-dollar investment from Microsoft to make it available globally, thereby offering more people access to language models.

The API also provides built-in support for all major programming languages: Python has the Chronology library for creating asynchronous flows and chaining prompts and responses (as it would be done in a chatbot implementation); in addition, there are multiple GitHub repositories with implementations in Go and Java.

In terms of pricing, OpenAI charges a flat usage fee based on the engine used (ranging from 0.0008 USD to 0.06 USD). Your usage is measured using tokens, which represent characters numerically (one token corresponds to roughly four characters or three-quarters of a word in English). To put this into perspective, the complete works of Shakespeare account for 900,000 words, translating to approximately 1.2 million tokens, costing 960.00 USD in the cheapest GPT-3 engine.

GPT-3'S LIMITATIONS

Although GPT-3 is remarkably large and powerful, there are several constraints that should be taken into account when incorporating it into a solution.

OpenAI uses tokens to estimate usage, both for billing purposes and to ensure API calls are fast and efficient. As a result, OpenAI imposes a limit of 2,048 tokens (approximately 1,500 words) for prompts and completions. Despite this limitation, GPT-3 still suffers from slow inference time, although it can still provide acceptable results in real-time implementations, such as chatbots.

Since GPT-3 is pre-trained, the major limitation lies in its training dataset:

- **No constant learning:** GPT-3 does not have long-term memory and cannot learn from interactions. However, OpenAI's latest model, ChatGPT, has introduced the capability to ask follow-up questions in a chatbot-like style.
- **ML bias:** being trained on internet data, GPT-3 exhibits biases that humans exhibit online. A research study has noted that GPT-3 is adept at producing radical text similar to conspiracy theories.

WHERE IS THE FUTURE HEADED?

GPT-3 has revolutionized Natural Language Processing, enabling the development of novel customer solutions that were unimaginable a few years ago.

Language Models bring the possibility of utilizing complex models for translation, question-answering, entity recognition, and more, paving the way for a future of intelligent systems that may support the analysis of text and cognitive processes.

Incorporating these tools into data analysis processes is becoming one of the most sought-after topics in the near future and has already demonstrated its potential in this early stage of adoption. Research in this field is continually progressing and increasing each day. For instance, OpenAI released the next generation of its GPT series, ChatGPT, in November 2022.

REPLY'S VALUE

Reply firmly believes in the potential of new technologies, so it is exploring the tools available in the commercial landscape, testing them, and experimenting through demos to gain a better understanding of the pros and cons of this new trend based on Artificial Intelligence and Machine Learning.

Reply is ready to review and analyze new tools that will come to the market, designing use cases for PoCs and creating projects for real business applications, with the main goal of providing cutting-edge and reliable solutions.

Going forward, Reply can help customers in several ways:

- Exploration of the main experiences and use cases is conducted through custom workshops.
- We aid customers in selecting the best platform or solution for their individual needs through surveys and assessment sessions.
- Analysing the business context to determine the adoption measures for the new GPT-3 and Natural Language Processing tools.