**LLMs:**

Before the introduction of Transformer models, the field of Natural Language Processing (NLP) was dominated by Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models. However, these models had certain limitations, especially when it came to handling long-range dependencies within the text.S

Transformers:

In 2017 google researchers introduced the transformer architecture in their paper “attention is all you need”. Instead of the sequential processing of inputs in RNNs and LSTMs (word by word), they opted for a parallel processing approach. In this case, the use of a parallel processing approach means that, instead of processing words one after another in a sentence, the model processes all words in the sentence at the same time. This allows the model to understand the relationships between all the words in a sentence, even those far apart, more efficiently.

The Transformer model’s architecture introduced the concept of attention mechanisms, which allows the model to weigh the contextual importance of different words and phrases. This approach significantly improved the model’s ability to handle long-range dependencies, resulting in improved results in tasks such as machine translation and text generation.

This new architecture. It served as the foundation for many incoming models and applications, including ChatGPT, which revolutionized the capabilities of language models.

Architecture:

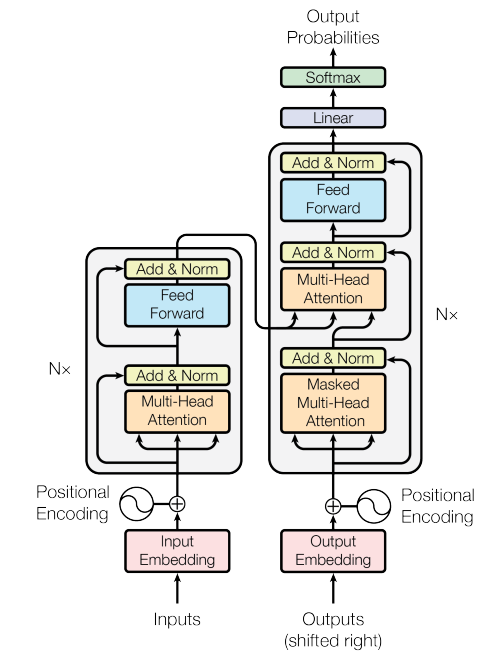


Figure 1 model architecture

In essence, the Transformer model is made up of two key parts, namely the encoder and the decoder.

**Encoder:** The encoder’s job is to turn the input data, like a sentence, into a format that’s easier for the model to understand. This is done using a series of N = 6 identical layers, all built in the same way. Each of these layers has two parts:

* **Self-Attention Layer.** mechanism is a key component in deep learning, particularly in the context of natural language processing (NLP) and computer vision tasks. It enables models to selectively focus on specific parts of the input data while processing information. The attention mechanism allows the model to give more importance to certain elements in the input sequence, leading to better performance in various sequential tasks. It checks how each word in a sentence relates to all the other words and uses this information to better understand the sentence as a whole. This part allows the model to create a context-aware representation of each word, getting both its meaning and its relationship to other words in the input.
* **Feed-Forward Neural Network:** This part is a simple, fully connected neural network that applies the same operation to each position in the sequence independently. In essence, it’s a network that learns from the data and makes predictions. It takes the output of the self-attention layer and tries to figure out more complex patterns. The FFN sub-layer in the Transformer is designed to add additional non-linearity and learn more complex representations from the input sequence. By applying a fully connected network at each position independently, the Transformer can effectively model dependencies within the sequence and capture meaningful relationships between different positions.

Decoders: Once the encoder has transformed the input data into a format the model can understand, the decoder steps in. Its job is to turn that format back into a form that’s useful for us, like a sentence. It does this through several layers, each made of three parts.

* Self-Attention Layer: Much like in the encoder, this layer allows the model to consider other words in the sentence when trying to understand one word. The difference here is that it is masked to only look at the words that have come before the current one, not those that come after.
* Cross-Attention Layer: This layer gives the decoder the ability to check the input sentence, meaning the output of the encoder, while generating the output. It’s a bit like looking back at your notes while doing homework. You want to keep looking back at the source to make sure what you’re writing makes sense.
* Feed-Forward Neural Network: Again, this part is very similar to the one in the encoder. It’s a mini-brain inside the model that learns from the data and makes predictions. It takes the output of the attention layers and tries to figure out more complex patterns.

These layers, similar to the layers in the encoder, have residual connections around them followed by layer normalization. The decoder also takes in a positional encoding of the input at the base of the stack to account for the order of the sequence.

Together, all these parts help the decoder turn the encoded input data into a useful output. This could be a translated sentence, a summary of a document, or an answer to a question. It all depends on what you want the model to do!

Large language model:

A large language model (LLM) is a type of machine learning model that can perform a variety of natural language processing (NLP) tasks, including generating and classifying text, answering questions in a conversational manner and translating text from one language to another.

LLMs are trained with immense amounts of data and use [self-supervised learning](https://www.techopedia.com/definition/34474/self-supervised-learning-ssl) to predict the next [token](https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html) in a sentence, given the surrounding context. The process is repeated over and over until the model reaches an acceptable level of accuracy.

Once an LLM has been trained, it can be fine-tuned for a wide range of NLP tasks, including:

* Building conversational chatbots like [ChatGPT](https://www.techopedia.com/definition/34933/chatgpt).
* [Generating text](https://www.techopedia.com/definition/34633/generative-ai) for product descriptions, blog posts and articles.
* Answering frequently asked questions (FAQs) and routing customer inquiries to the most appropriate human.
* Analyzing customer feedback from email, social media posts and product reviews.
* Translating business content into different languages.
* Classifying and categorizing large amounts of text data for more efficient processing and analysis.

Large language models typically have a transformer-based architecture. This type of AI architecture uses self-attention mechanisms to calculate a weighted sum for an input sequence and dynamically determine which tokens in the sequence are most relevant to each other.it allow a machine learning model to identify relationships between words in a sentence — regardless of their position in the text sequence — by using self-attention mechanisms.

The relationships between tokens in a sequence are calculated using attention scores that represent how import a token is in regards to the other tokens in the text sequence.

Different LLMs:

Some of the most popular large language models are:

GPT4:

Generative Pre-trained Transformer 4 (GPT-4) is a multimodal large language model created by OpenAI, and the fourth in its numbered "GPT-n" series of GPT foundation models. It is the newest version of OpenAI's language model systems. Its previous version, GPT 3.5, powered the company's wildly popular ChatGPT chatbot when it launched in November of 2022.

GPT-4 has 175 billion parameters, which is more than 10 times the number of parameters in GPT-3. This makes it capable of generating more complex and nuanced text, as well as performing more sophisticated tasks.

* **Features:**

1- Visual input option

Although it cannot [generate images](https://research.aimultiple.com/generative-ai-applications/#1-visual-applications) as outputs, it can understand and analyze image inputs. GPT-4 has the capability to accept both text and image inputs, allowing users to specify any task involving language or vision. It can generate various types of text outputs, such as natural language and code, when presented with inputs that include a mix of text and images.

2- Higher word limit

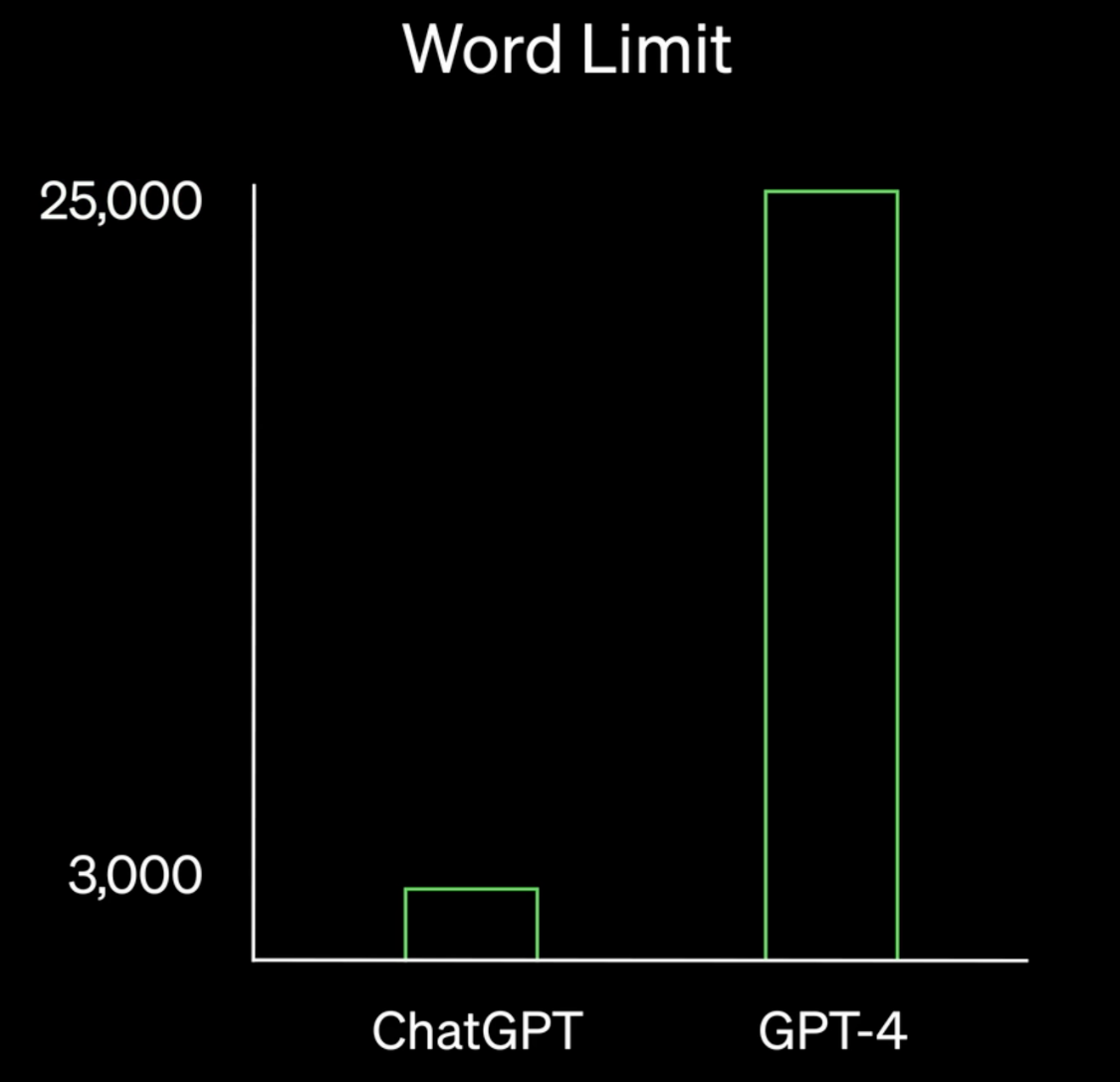


Figure 2 The comparison of ChatGPT with GPT-3.5 and GPT-4 in terms of word limit

GPT-4 has the ability to process more than 25,000 words of text (see Figure 5 above), making it suitable for a variety of use cases, such as:

* Creating long-form content
* Carrying out extended conversations
* Conducting document analysis and search tasks.

3- Advanced reasoning capability

GPT-4 is outstanding compared to the earlier versions with its natural language understanding ([NLU](https://research.aimultiple.com/nlu-vs-nlp/)) capabilities and problem-solving abilities. The difference may not be observable with a superficial trial, but the test and benchmark results show that it is superior to others in terms of more complex tasks.

As an example, OpenAI tested the large language models in a simulated bar exam. GPT-4’s bar exam results show that it scored in the top 10% of test-takers, while GPT-3.5’s score was in the bottom 10%. Overall, the performance of GPT-4 on various professional exams outperformed that of GPT-3.5 (Figure 7).

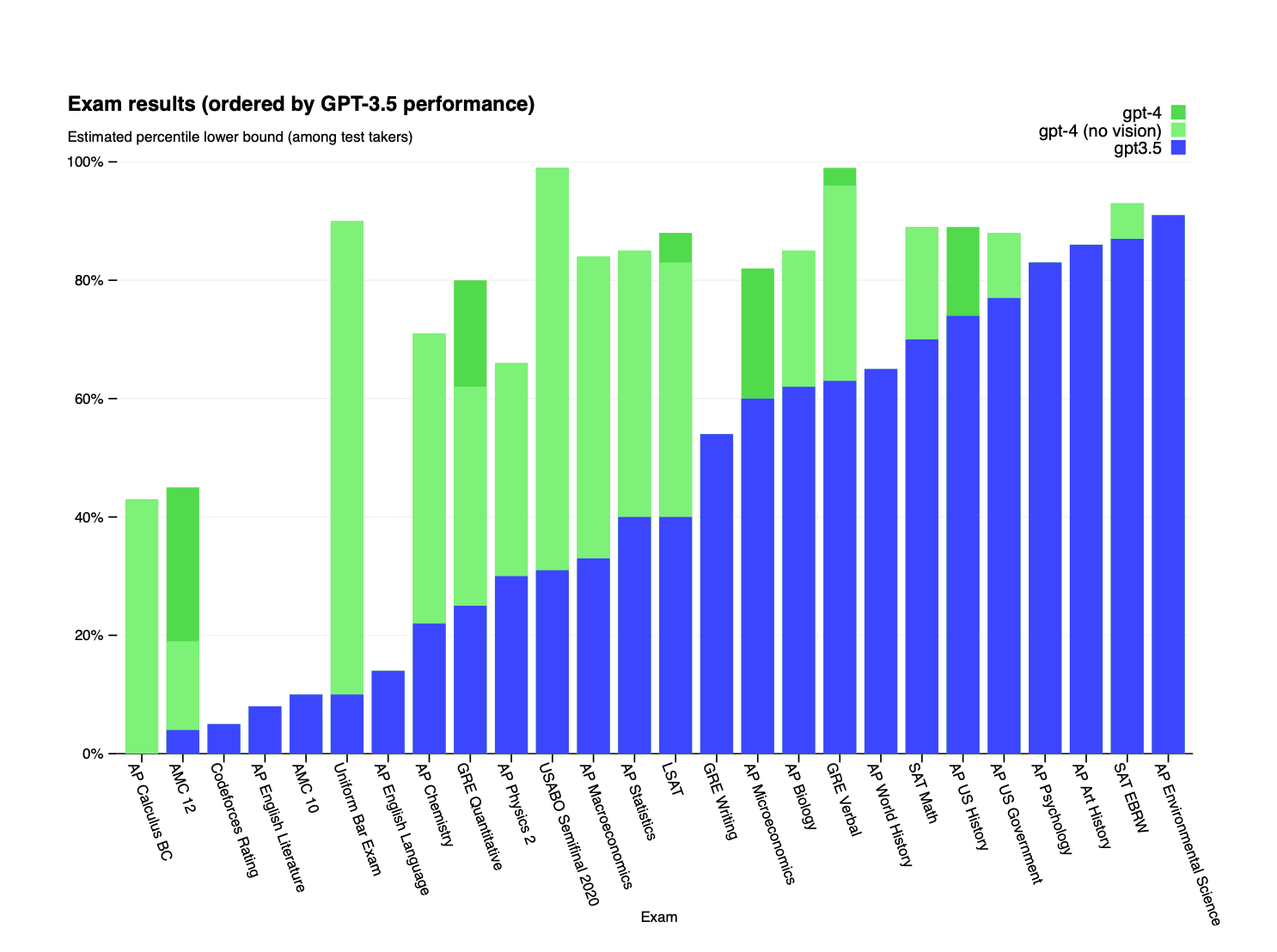
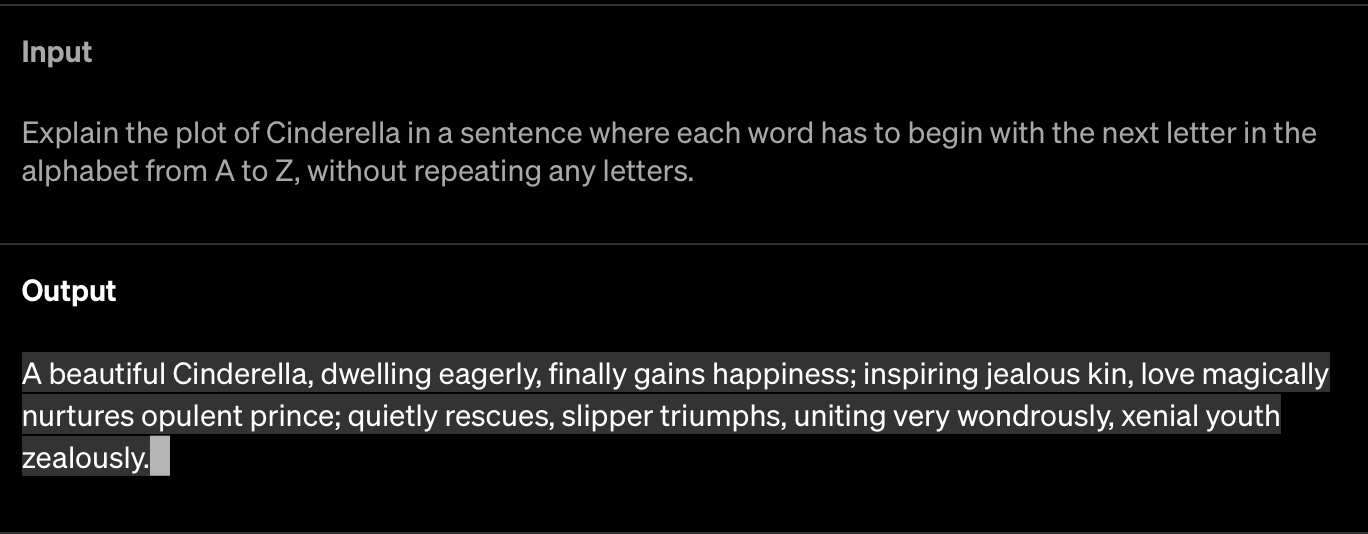


Figure 3 The comparative analysis of exam results of the three GPT models

4- Advanced creativity

As a result of its higher language capabilities, GPT-4 is advanced in creativity compared to earlier models (Figure 7). This can make the language model more adaptive to certain use cases that require creative writing skills, such as:

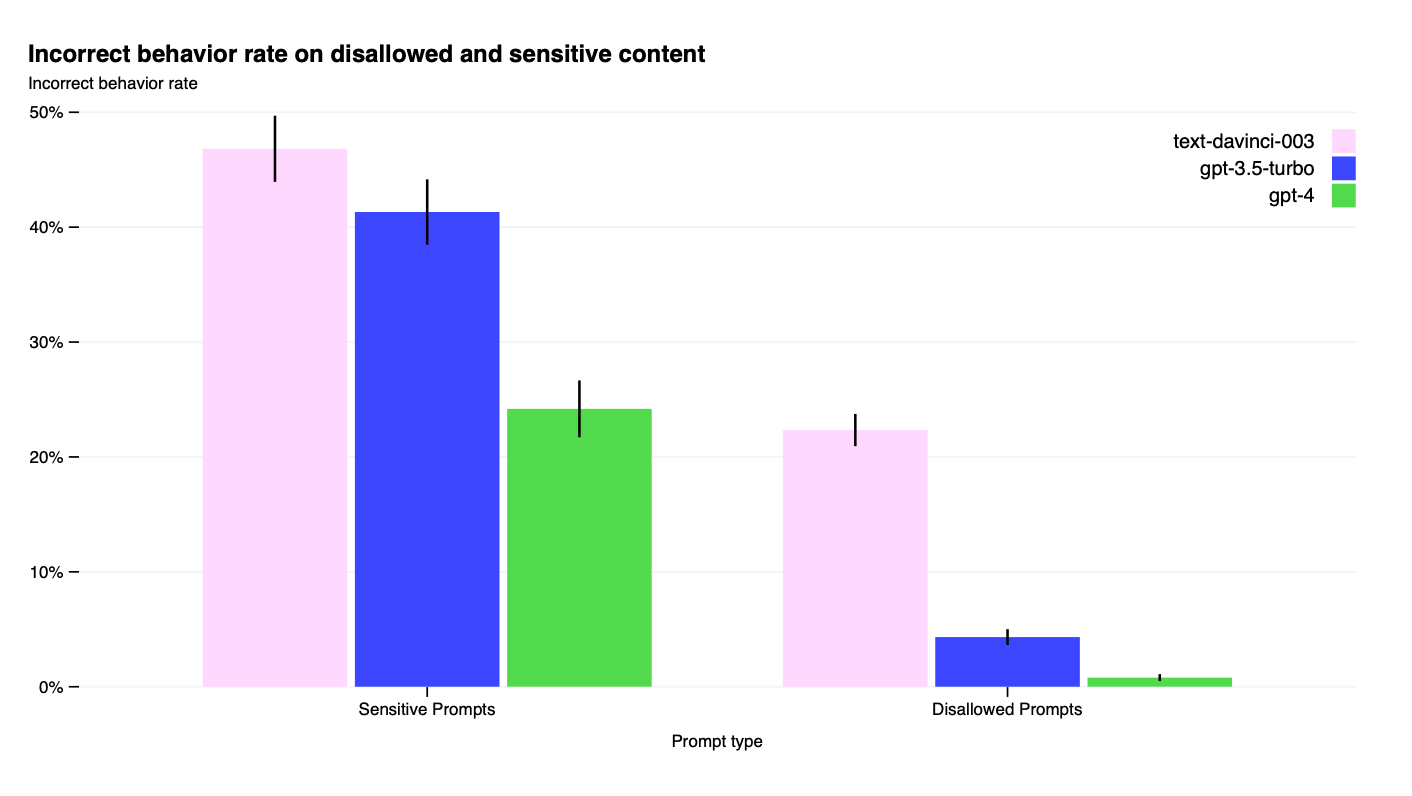
* Screenplay writing
* Blog post creation
* Essay writing



5- Adjustment for inappropriate requests

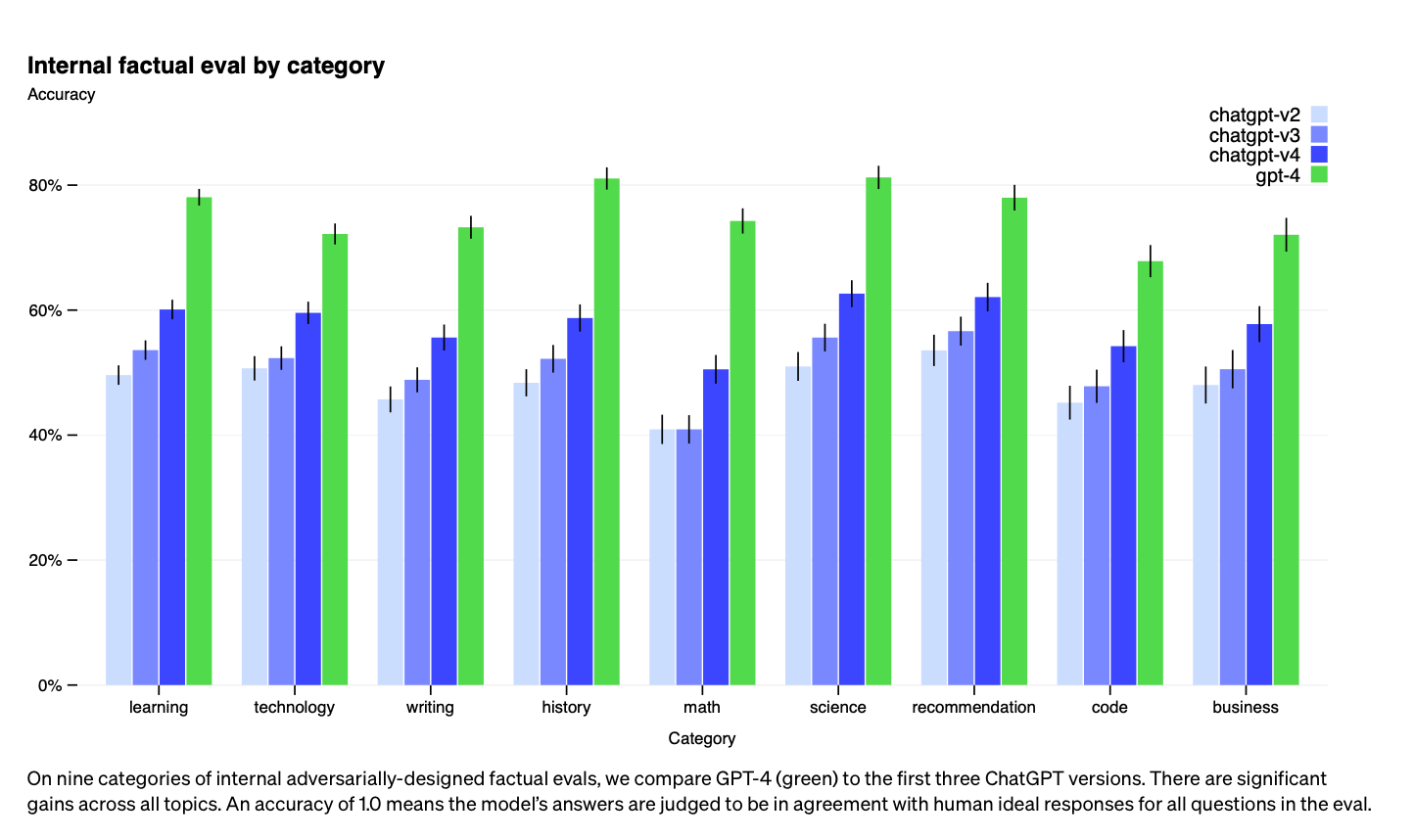
ChatGPT was criticized for its [handicap](https://research.aimultiple.com/generative-ai-ethics/#what-are-the-concerns-around-generative-ai-ethics) in terms of providing answers to inappropriate requests such as explaining how to make bombs at home, etc. OpenAI was working on this problem, and made some adjustments to prevent the language models from producing such content.

According to OpenAI, GPT-4 is 82% less likely to respond to requests for disallowed and sensitive content.



6- Increase in fact-based responses

Another limitation of the earlier GPT models was that their responses were not factually correct for a substantive number of cases. OpenAI announces that GPT-4 is 40% more likely to produce factual responses than GPT-3.5.



* **Limitations:**

Knowledge update limitation: Like previous GPT models, GPT-4 generally does not possess knowledge of events that have occurred after the vast majority of its training data was collected (i.e., before September 2021). Also, it does not have the ability to learn from its experiences.

ORCA:

Orca LLM is a new open-source language model developed by Microsoft Research. It is based on the Llama model, but it uses a technique called imitation learning to improve its reasoning capabilities. Imitation learning is a process where a model learns by observing and imitating the behavior of another model. In the case of Orca, the model that it is imitating is ChatGPT/GPT-4.

Orca has been shown to be effective at a variety of tasks, including:

**Generating text:** Orca can generate text that is both coherent and informative. It can be used to create realistic dialogue, write different kinds of creative content, and answer your questions in an informative way.

**Translating languages:** Orca can translate between different languages. It is more accurate than previous language models, and it can handle more complex translations.

**Answering questions:** Orca can answer your questions in an informative way. It can access and process information from the real world through Google Search, and it can provide you with summaries of factual topics.

**Following instructions:** Orca can follow complex instructions. It can be used to complete tasks such as writing code, generating images, and controlling robots.

Orca is still under development, but it has the potential to be a powerful tool for a variety of applications. It is a promising new model that could revolutionize the way we interact with computers.

* **Key features of Orca LLM:**
* It is a large language model with 13 billion parameters.
* It is trained on a massive dataset of text and code.
* It can generate text that is both coherent and informative.
* It can translate between different languages.
* It can answer your questions in an informative way.
* It can follow complex instructions.
* It uses imitation learning to improve its reasoning capabilities.

Orca is a powerful tool that has the potential to change the way we interact with computers. It is still under development, but it is already being used for a variety of tasks. As it continues to develop, we can expect to see even more innovative and groundbreaking applications.

* **Limitations: (**[**https://arxiv.org/pdf/2306.02707.pdf**](https://arxiv.org/pdf/2306.02707.pdf) **)**

Orca, built upon the LLaMA model family, retains many of its constraints, as well as the common limitations of other large language models, including:

**Data Biases:** Large language models, trained on extensive data, can inadvertently carry biases present in the source data. Consequently, the models may generate outputs that could be potentially biased or unfair.

**Lack of Contextual Understanding:** Despite their impressive capabilities in language understanding and generation, these models exhibit limited real-world understanding, resulting in potential inaccuracies or nonsensical responses.

**Lack of Transparency:** Due to the complexity and size, large language models can act as ‘black boxes,’ making it difficult to comprehend the rationale behind specific outputs or decisions. We recommend reviewing transparency notes from Azure for more information .

**Content Harms**: There are various types of content harms that large language models can cause. It is important to be aware of them when using these models, and to take actions to prevent them. It is recommended to leverage various content moderation services provided by different companies and institutions. On an important note, we hope for better regulations and standards from government and technology leaders around content harms for AI technologies in future. We value and acknowledge the important role that research and open source community can play in this direction.

**Hallucination:** It is important to be aware and cautious not to entirely rely on a given language model for critical decisions or information that might have deep impact as it is not obvious how to prevent these models to fabricate content. Moreover, it is not clear whether small model may more susceptible to hallucination in ungrounded generation use cases due to their smaller size and hence reduced memorization capacity. This is an active research topic and we hope there will be more rigorous measurement, understanding and mitigations around this topic.

**Potential for Misuse:** Without suitable safeguards, there is a risk that these models could be maliciously used for generating disinformation or harmful content.

Additionally, Orca’s performance is influenced by the data used for explanation tuning:

**Zero-Shot Settings:** Orca has been trained on data that simulate zero-shot setting with standard prompts. The model’s performance in other contexts such as multi-turn conversations, in-context-learning and few-shot learning, or advanced prompting techniques like chain-of-thought prompting remains untested.

**Data Distribution:** Orca’s performance is likely to correlate strongly with the distribution of the tuning data. This correlation might limit its accuracy in areas underrepresented in the training dataset such as math, coding, and reasoning.

**System messages:** Orca is trained with diverse system instructions to elicit different kinds of response. Additionally, the stochasticity introduced by the model size may lead to generation of non-deterministic responses to different system instructions.

**GPT-4 Behavior:** As Orca is trained to imitate GPT-4, it could inherit both the advantages and shortcomings of the teacher model. We posit that Orca benefits from the safety measures incorporated during GPT-4 training and safety guardrails (e.g., content filter) within the Azure OpenAI API. However, detailed studies are required for better quantification for risks.

This model is solely designed for research settings, and its testing has only been carried out in such environments. It should not be used in downstream applications, as additional analysis is needed to assess potential harm or bias in the proposed application.

LLAMA:

The Llama Model is an auto-regressive language model, based on the transformer architecture, developed by The FAIR team of **Meta AI**. The model was trained between December 2022 and February 2023, and it comes in different sizes: **7B**, **13B**, **33B**, and **65B** parameters. The primary intended users of the model are researchers in natural language processing, machine learning, and artificial intelligence. LLaMA is licensed for **non-commercial** use only. One of the most relevant factors for which model performance may vary in which language is used. The model has not been trained with human feedback, and can thus generate toxic or offensive content, incorrect information, or generally unhelpful answers The model was evaluated on several benchmarks and trained using various sources of data CCNet [67%], C4 [15%], GitHub [4.5%], Wikipedia [4.5%], Books [4.5%], ArXiv [2.5%], Stack Exchange [2%]. Llama is a powerful tool that can be used for a variety of tasks, including:

* Natural language understanding (NLU): Llama can be used to understand the meaning of text, identify entities and relationships, and answer questions.
* Natural language generation (NLG): Llama can be used to generate text, translate languages, write different kinds of creative content, and answer your questions in an informative way.
* Machine translation: Llama can be used to translate text from one language to another.
* Text summarization: Llama can be used to summarize text into a shorter, more concise version.
* Question answering: Llama can be used to answer your questions in an informative way.

Llama is still under development, but it has already shown great promise in a variety of tasks. It is likely to become an increasingly important tool in the years to come.

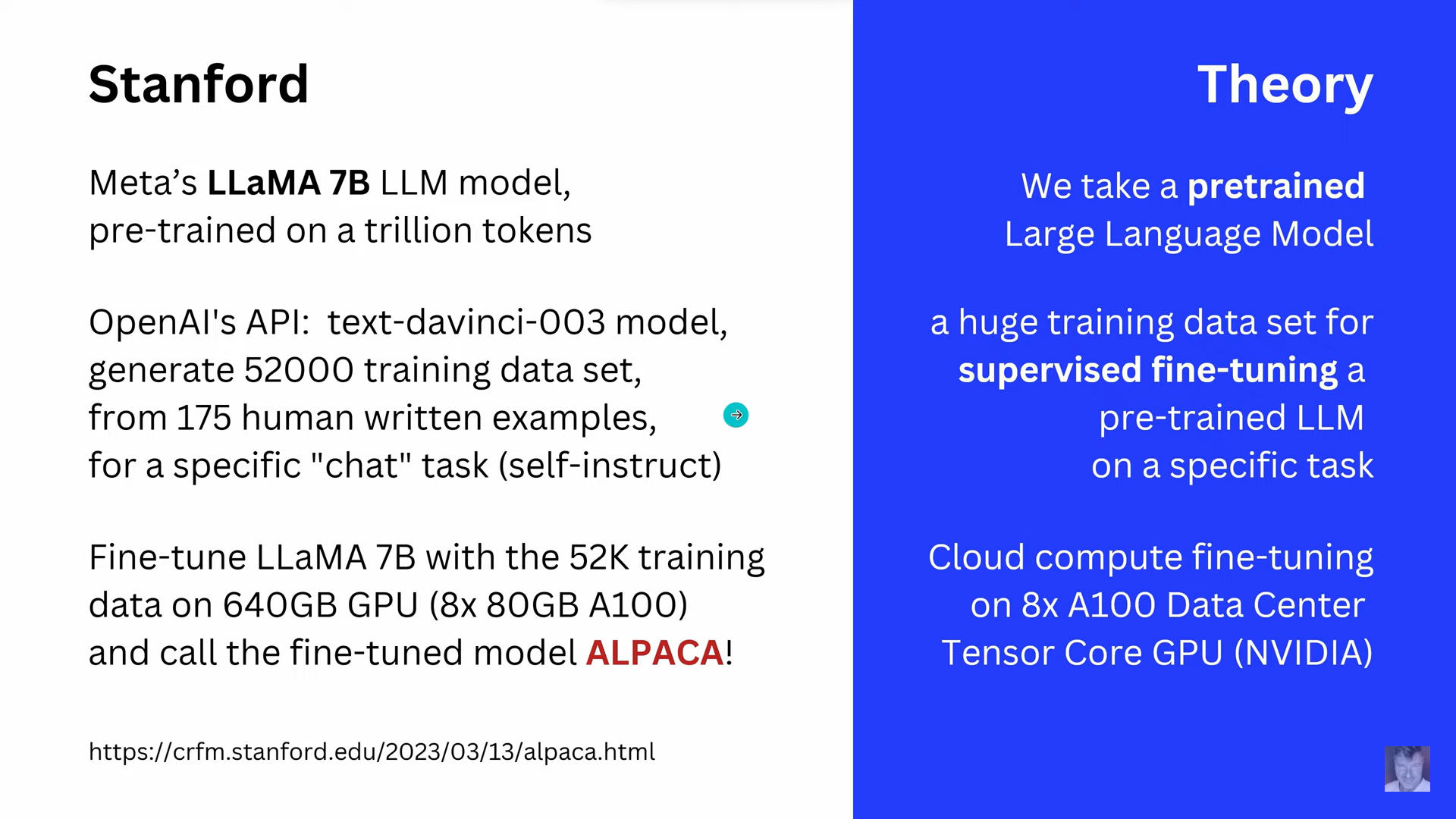
* the benefits of using Llama:
* It is a powerful tool that can be used for a variety of tasks.
* It is still under development, so it is constantly learning and improving.
* It is open source, so anyone can use it.
* the limitations of using Llama:
* It is a large language model, so it requires a lot of computing power to train and use.
* It is still under development, so it can sometimes make mistakes.
* It is not always able to understand the nuances of human language.
* Since the majority of dataset comprises English text, the performance of the model on languages other than English may be comparatively lower.
* It is not good at mathematical reasoning and domain knowledge.

Overall, Llama is a powerful tool that has the potential to be very useful in a variety of tasks. However, it is important to be aware of its limitations before using it.

In July 2023, Meta AI and Microsoft announced Llama 2, the next generation of Llama. Llama 2 is trained on a larger dataset and has up to 70 billion parameters. It is also able to perform a wider range of tasks, including **code generation, creative writing, and question answering**.

Llama 2 is still under development, but it has the potential to be even more powerful and versatile than Llama. It is likely to become an important tool for a variety of applications, including artificial intelligence, machine learning, and natural language processing.

ALPACA:





ALPACA is a large language model (LLM) developed by Stanford University's Center for Research on Foundation Models. It is a smaller and more affordable version of LLMs like GPT-3 and Jurassic-1 Jumbo, but it is still able to perform a variety of tasks, including:

* **Natural language understanding (NLU):** ALPACA can be used to understand the meaning of text, identify entities and relationships, and answer questions.
* **Natural language generation (NLG):** ALPACA can be used to generate text, translate languages, write different kinds of creative content, and answer your questions in an informative way.
* **Machine translation:** ALPACA can be used to translate text from one language to another.
* **Text summarization:** ALPACA can be used to summarize text into a shorter, more concise version.
* **Question answering:** ALPACA can be used to answer your questions in an informative way.
* **the benefits of using ALPACA:**
* It is a smaller and more affordable version of LLMs like GPT-3 and Jurassic-1 Jumbo.
* It is still able to perform a variety of tasks.
* It is still under development, so it is constantly learning and improving.
* **the limitations of using ALPACA:**
* It is a smaller model, so it may not be as powerful as some other LLMs.
* It is still under development, so it may sometimes make mistakes.
* It is not always able to understand the nuances of human language.

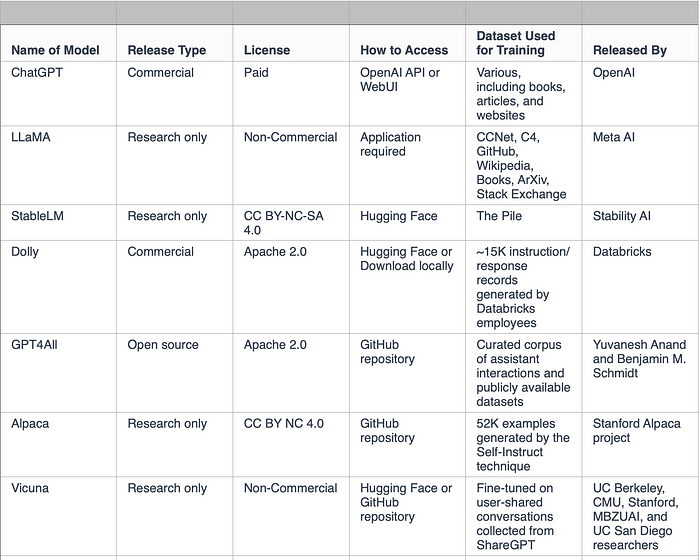
the differences between ALPACA and Llama:

* ALPACA is a smaller model, with 7B parameters compared to Llama’s 70B parameters.
* ALPACA is trained on a different dataset, which includes instruction-following demonstrations.
* ALPACA is fine-tuned using supervised learning, while Llama is fine-tuned using reinforcement learning.

**Biggest weakness of large language models**

AI language models, whether it is [ChatGPT](https://aibusiness.com/search?q=ChatGPT), [Llama](https://aibusiness.com/meta/meta-s-llama-language-model-outperforms-openai-s-gpt-3), [Alpaca](https://aibusiness.com/nlp/meet-alpaca-the-open-source-chatgpt-made-for-less-than-600) or [GPT-4](https://aibusiness.com/nlp/openai-unveils-gpt-4-and-plans-to-add-it-to-chatgpt-), share one major weakness: Hallucinations.

Hallucinations occur when a conversational AI application powered by a large language model generates a response to a prompt that is either false or irrelevant to the original request.



BIBLIOGRAPHY:

Illustrated transformers: <http://jalammar.github.io/illustrated-transformer/>

“Attention is all you need”: <https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>

<https://dagshub.com/blog/large-language-models-explained/>

<https://blog.gopenai.com/how-transformers-and-large-language-models-llms-work-3f20bb41c1ff>

<https://github.com/Hannibal046/Awesome-LLM#tools-for-deploying-llm>

<https://www.borealisai.com/research-blogs/a-high-level-overview-of-large-language-models/>

<https://plainenglish.io/blog/detailed-comparison-of-the-latest-large-language-models>

<https://www.projectpro.io/article/llama-vs-alpaca-models/866>

<https://crfm.stanford.edu/2023/03/13/alpaca.html>

<https://research.aimultiple.com/meta-llama/>

<https://www.computerworld.com/article/3697649/what-are-large-language-models-and-how-are-they-used-in-generative-ai.html>