

WHITE PAPER

THE RISE OF AI ASSISTANTS: FROM CONCEPT TO PRACTICAL APPLICATIONS

Authors: Viktor Seredovych, Taras Khapko



Introduction 4

A Glimpse into the Past 5

How do Large Language Models Work? 7

Key Idea

Autoencoders

Transformers

How to Train Large Language Models? 16

Pretraining Phase

Fine-tuning Phase

Custom Data Assistants 19

Grounding

Retrieval Augmented Generation (RAG)

Custom Data Fine-tuning

| | |
|---|----|
| GPT and LLaMa Comparison | 26 |
| Accuracy of Responses | |
| Grounding: Ease and Accuracy | |
| Deployment: Local vs Online Deployment | |
| Maintenance: Price and Complexity | |
| Security and Privacy | |
| <hr/> | |
| LLMs in Action | 34 |
| <hr/> | |
| Benefits for Businesses: Why Choose LLMs? | 40 |
| <hr/> | |
| Conclusion | 43 |
| <hr/> | |
| References | 44 |

Introduction

In an era where digital communication has become the cornerstone of modern business and personal interactions, AI assistants have emerged as a powerful tool for redefining how we engage with technology. Risen from rule-based chatbots to Large Language Models (LLMs) with transformer architecture, these virtual assistants have directed a new age of human-computer interchange. They offer businesses and individuals unique opportunities to enhance efficiency, foster innovation, and many other benefits.

However, with such a great variety of virtual assistants available today in the AI landscape, there is still a lack of understanding of how these solutions work, differ, evolve, and update to provide their myriad benefits. At **tsukat**, we recognize the importance of staying at the edge of technological progress. We provide our clients with the most advanced solutions while supplying all the essential information to make informed decisions.

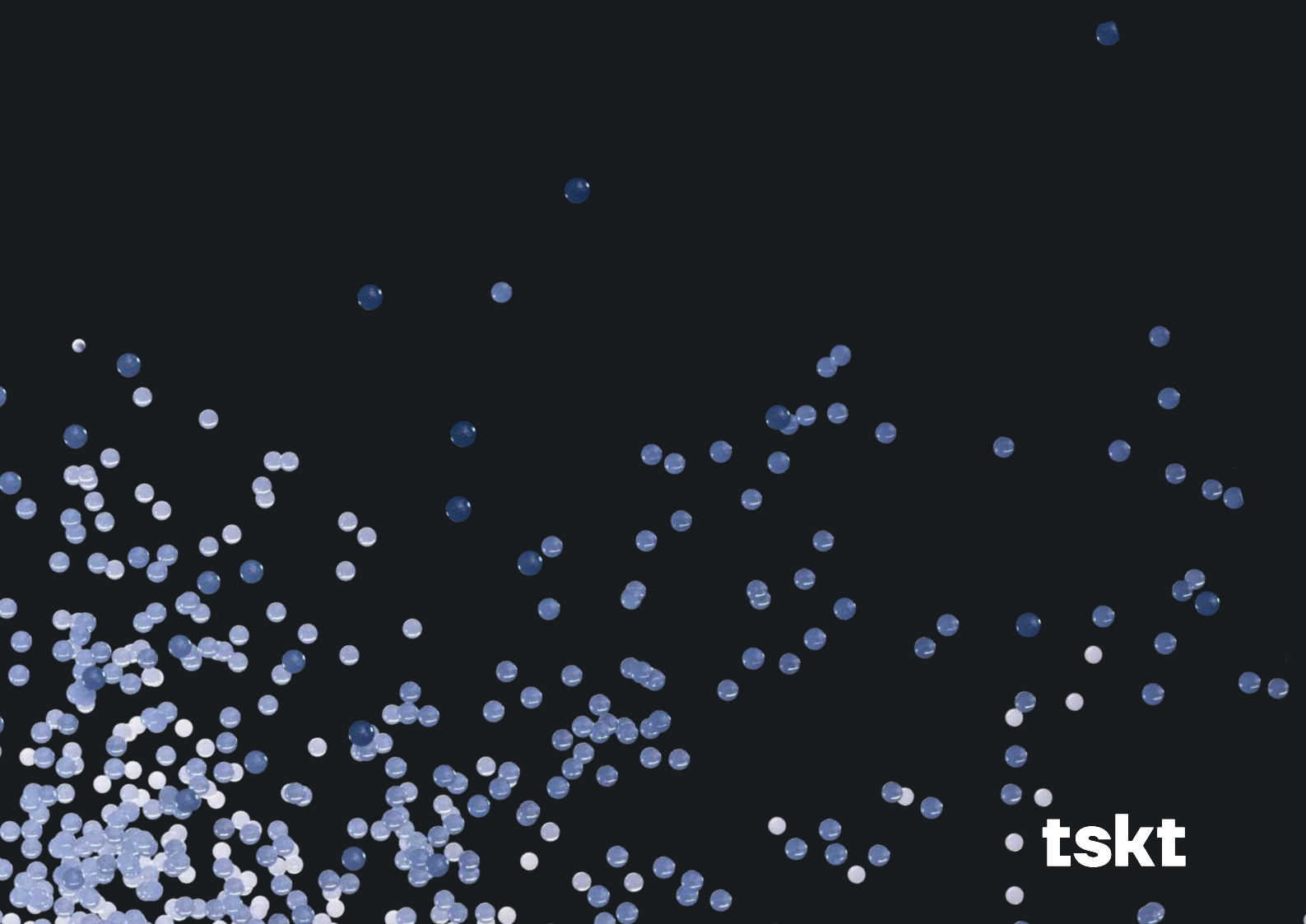
This whitepaper aims to provide a detailed and accurate representation of AI assistants, the training methodology, a comparative analysis of the dominant LLMs, and their potential for business use.

Ultimately, leveraging the power of Large Language Models (LLMs), which are constantly evolving, businesses can enhance their operations and services and gain significant benefits: automation, scalability, customization, efficient data analysis, and many others.

“ Virtual assistants have directed a new age of human-computer interchange

CHAPTER 1

A GLIMPSE INTO THE PAST



tskt

The story of chatbots dates to the early days of computing, but only in recent years have they truly come into their own. Chatbots are software applications designed to simulate human-like conversations with users. Their evolution is intertwined with the development of Natural Language Processing (NLP) and machine learning technologies, especially LLMs.

The early 2000s saw the rise of rule-based chatbots. These chatbots operated on predefined rules and scripts. You ask a specific question, and you get a specific answer. Though limited in functionality, these rudimentary chatbots paved the way for what was to come.

LLMs are not just another incremental step in the evolution of chatbots. They signify a paradigm shift. By leveraging vast amounts of data and cutting-edge neural network architectures, they are capable of understanding context, nuance, and even emotions to a certain extent. Their ability to understand and generate language at a level comparable to humans has elevated these systems beyond traditional chatbots. Recognizing their enhanced capabilities and sophistication, it is more fitting to refer to them as 'assistants' rather than mere chatbots. This terminology shift reflects their evolution from simple message responders to versatile, intelligent systems capable of a wide range of sophisticated interactions.

The first breakthrough came with the integration of artificial intelligence (AI) and NLP techniques. AI-powered chatbots started to understand and generate human-like text, making conversations with them more natural and intuitive. However, these early AI chatbots still had their limitations, often struggling to handle complex inquiries.

The true game-changer in the evolution of chatbots and NLP was the advent of Large Language Models (LLMs). These sophisticated models, such as [OpenAI's GPT-3](#) and their successors, have revolutionized the field by bringing human-level language understanding and generation capabilities to machines.

CHAPTER 2

HOW DO LARGE LANGUAGE MODELS WORK?

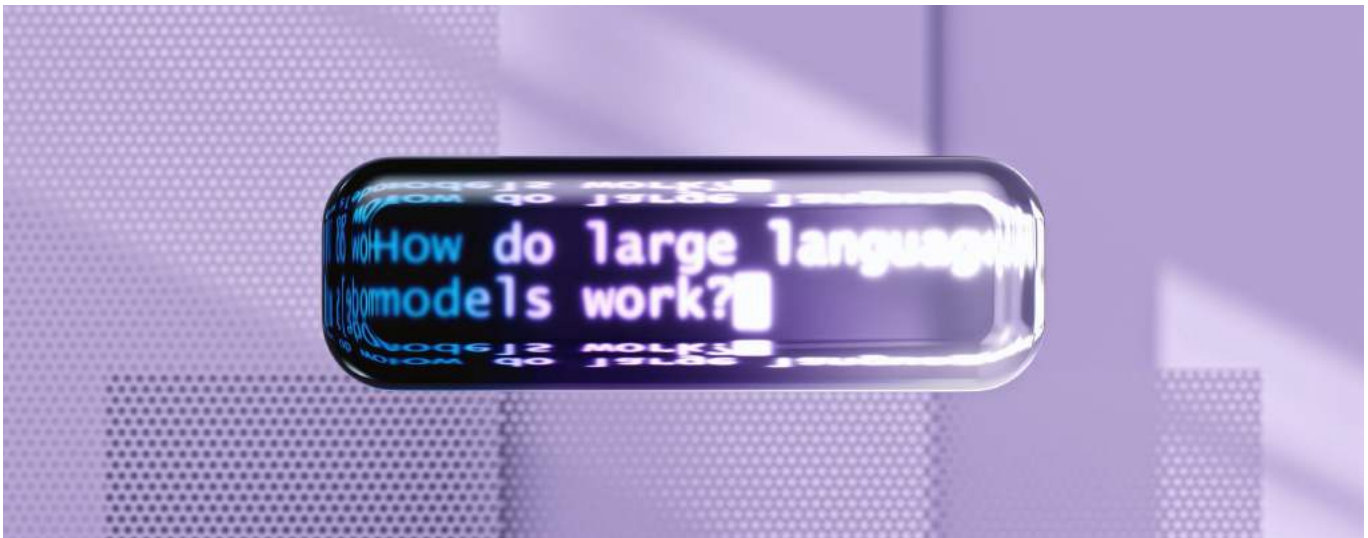
A decorative background consisting of numerous small, semi-transparent blue and white dots scattered across the dark background, primarily concentrated in the lower half of the page.

tskt

A language model can be thought of as a highly skilled interpreter of text. When given a sequence of words, it anticipates the next word based on patterns it's familiar with. Essentially, it's a tool that comprehends and crafts text that mirrors human communication, underpinning applications from chatbots to content generation.

The latest language models use neural networks to analyze large amounts of data for training purposes. A neural network mimics the workings of a human brain within a digital realm. It learns from data, much like our brains learn from experiences. This network consists of interconnected digital "neurons" arranged in layers.

As data flows through these layers, the network establishes links and makes decisions, much like the neuron interactions in our brain. With continued training and sample data input, these decisions become more nuanced and accurate, allowing the model to generate coherent and contextually relevant text.



“ “ The latest language models use neural networks to analyze large amounts of data for training purposes.

In its essence, upon processing a sequence of words, the network aims to accurately predict the subsequent word that best completes the sentence. This process is connected with the problem of finding the probability of the next word, expressed as follows

$$P(\text{word}_n | \text{word}_1, \text{word}_2, \dots, \text{word}_{n-1})$$

The model evaluates the likelihood of every word in its extensive dictionary being the appropriate next word in the sequence. The word with the highest probability is then selected as the next word.

Ideally, a well-functioning model will consistently select words that are not only probable but also contextually relevant, enhancing the coherence and fluency of the generated text.



However, even assuming we use only 10,000 English words and want to create a network that would make prediction based on one word only, a direct connection between every input neuron to every output neuron results in $10,000 \times 10,000 = 100$ million connections. The complexity magnifies with larger vocabulary and processing more than just one word as input, which calls for a smart solution to condensing the model. This is done by the use of encoders.

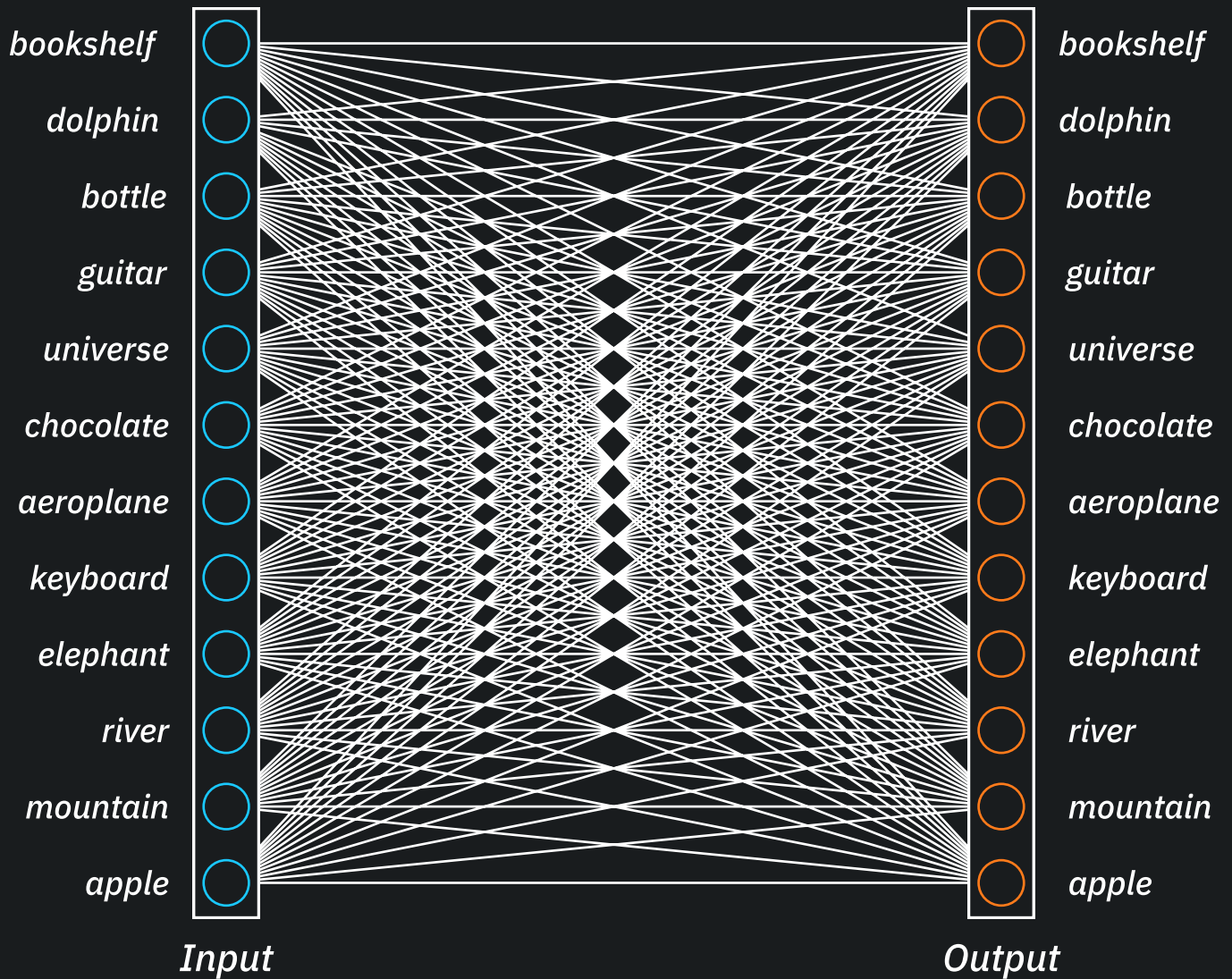


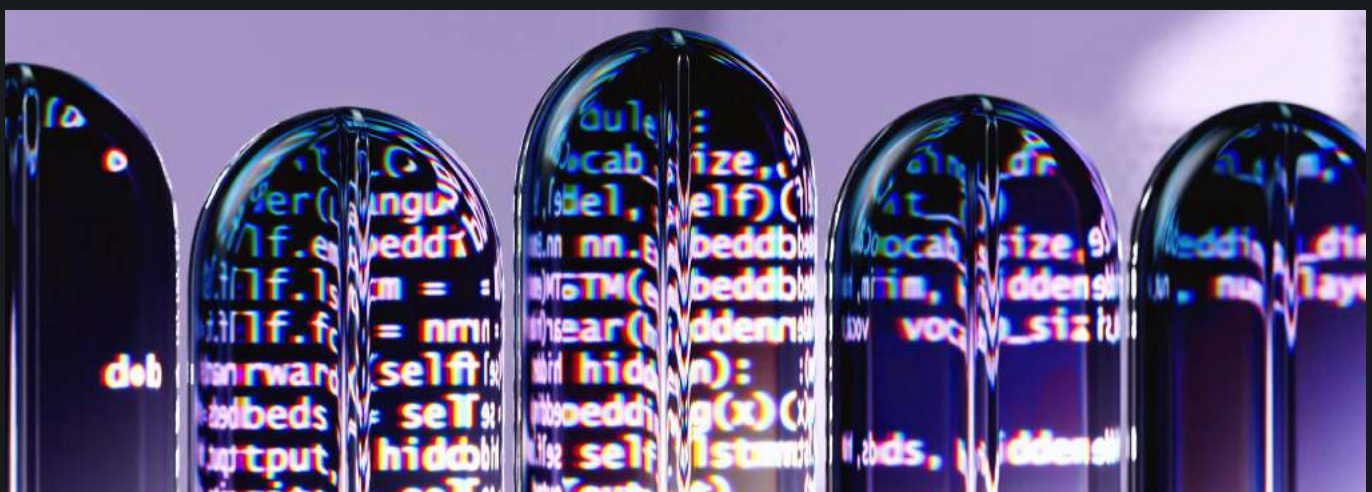
Figure 1: Every circle or neuron represents one word. Input layer takes the last word of the sentence, and as an output we obtain the consequent word.

In any language, there are words that are either synonyms (words with similar meanings) or semantically related (words connected in meaning). Consider the following example:

“The mist gathers over the ____”

The term “mist” can be interchangeably used with words such as “fog” or “haze”. These words share similar meanings, and substituting one for another will not significantly alter the context of a sentence. Rather than considering each word as entirely distinct, we can group them together and encode this group into a single, fictional term that collectively represents all of them. For instance, the words “mist”, “fog”, and “haze” might be encapsulated by a singular, hypothetical term that approximates their collective meaning.

This approach allows the model to interpret a cluster of words as conveying a related concept, thus eliminating the need to learn and memorize each word individually. Consequently, this strategy substantially reduces the complexity of the language model. Instead of requiring unique representations for each of the 10,000 words, the model can efficiently operate with a smaller, more manageable set of encoded groups.



Consider an encoder setup that processes 10,000 words and maps them to a more compact set of outputs, for example, 256 numerical values ranging between 0 and 1.

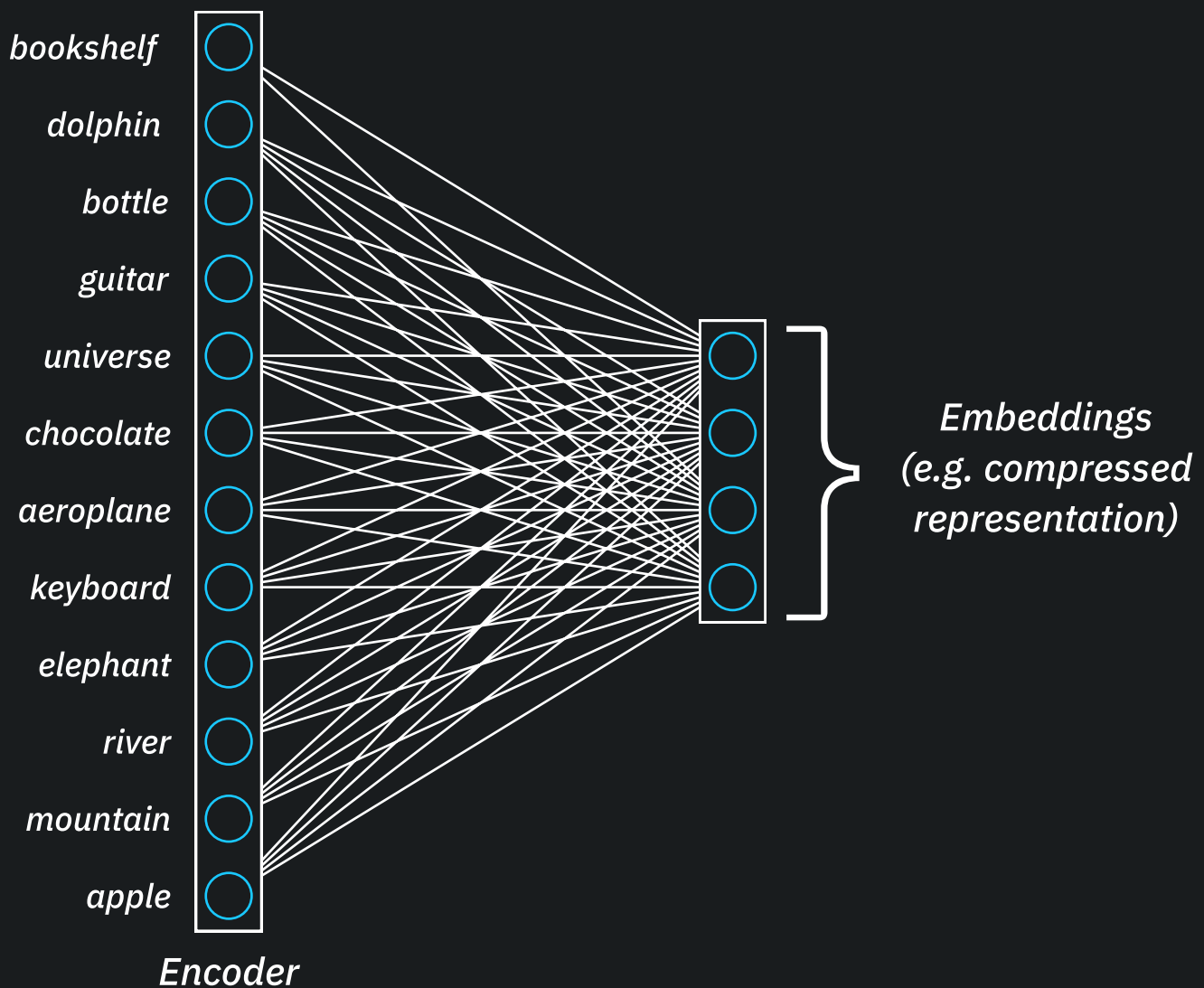


Figure 2: Encoder model that takes a word and outputs its embeddings vector.

These new representations of words, typically referred to as “embeddings”, “encodings”, or “latent space”, are essentially vectors of numbers. The degree of similarity between these vectors increases as the meanings of the words they represent become more closely aligned.

We can now introduce the concept of the decoder network. The primary function of this network is to take the embeddings generated by the encoder and map them back to the original 10,000 words. The encoder and decoder can be combined to create an autoencoder neural network architecture.

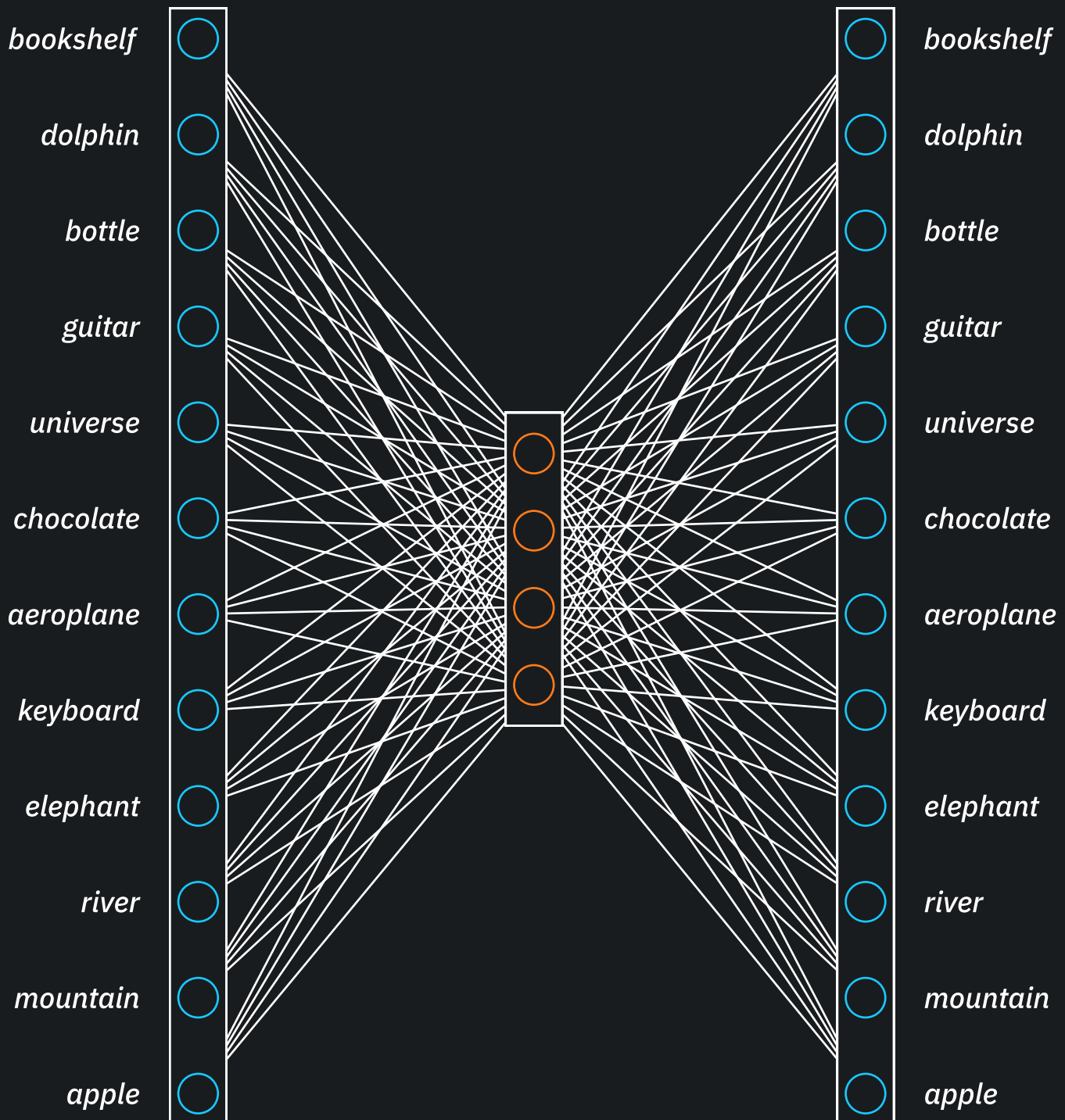
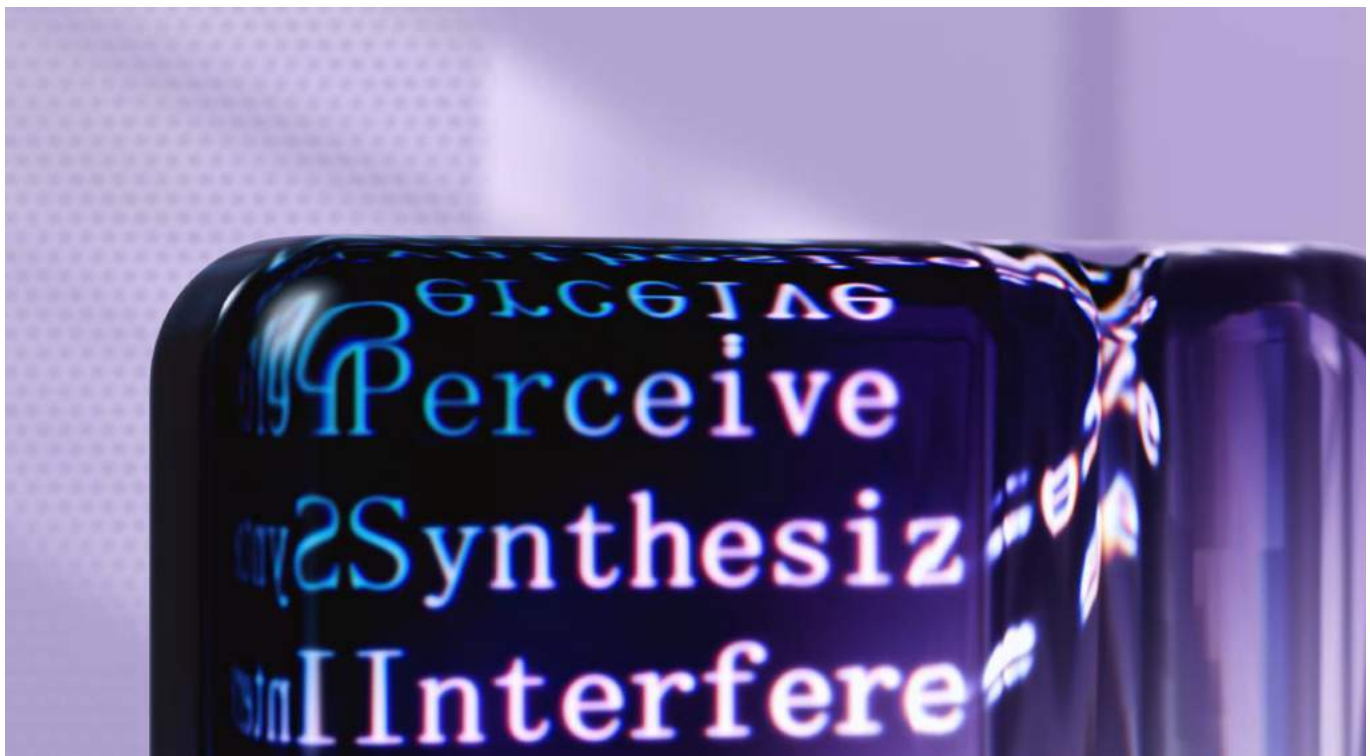


Figure 3: Autoencoder model that combines the encoder and decoder into a single neural network.

In essence, the encoder compresses input data into a dense, encoded representation, capturing the essence and context of the words. The decoder then takes this processed information and reconstructs it, generating the most likely subsequent word in the sequence. This collaborative functioning between the encoder and the decoder can be regarded as a form of lossy compression.

By condensing 10,000 words into a smaller set of values, the network is compelled to make compromises and grouping words together, so the original breadth of the language is simplified, leading to some loss of detail.

However, this strategy is crucial for reducing the complexity of the neural network, allowing for more efficient processing while still maintaining the capability to produce meaningful and contextually appropriate text.



The real game-changer in the LLM space is the transformer architecture, the backbone of models like GPT. At its core lies the classical encoder-decoder network that we briefly described earlier. However, the main power of transformers comes from a novel component known as self-attention.



In the **beginning** the **Universe** was **created**.

Figure 4: The self-attention allows the transformer model to understand the context of a word better based on its relation to other words. Many state-of-the-art LLMs are based on transformers.

Self-attention [1] is a mechanism that allows the model to consider the entire context of a sentence or a sequence of words all at once, rather than one at a time. This means that for any given word, the model can assess which other words in the sentence have the most relevance to it and adjust its understanding of that word accordingly. As a result, transformers are particularly good at handling tasks that require an understanding of the relationship between all parts of the input data, such as predicting a word that has been intentionally omitted from a sentence.

Transformers have proven highly effective for a variety of tasks due to their ability to process and generate contextually relevant and coherent text.

CHAPTER 3

HOW TO TRAIN LARGE LANGUAGE MODELS?



tskt

Training Large Language Models (LLMs) is a complex and resource-intensive endeavor that constitutes a significant portion of the associated engineering challenge. To better understand this process, we'll explore the training methodology of the LLaMA2 70B model – an open-source model with a wealth of research information available about it [2].

The LLaMA2 70B model is constructed of two primary components: the parameters that weight approximately 140GB, and the code that executes the LLaMA architecture using these parameters. The model's training can be broken down into two primary stages: pretraining and fine-tuning.

Pretraining Phase

01

Pretraining is an unsupervised training stage that involves processing an extensive dataset, such as a 10TB collection comprising various internet texts, including articles, technical documentation, code samples, and social network exchanges. Usually, this data is gathered automatically all over the internet, and thus may be of a low quality. This vast dataset of internet knowledge is lossy compressed into the model effectively reducing its overall size to 140GB. This stage requires substantial computational resources:

6000 GPUs

12 million

12 days

While the pretraining phase equips Large Language Models (LLMs) with the foundational skill of predicting subsequent words in a sequence, this alone doesn't fully harness their potential. For instance, in the context of AI assistants, this fundamental skill is not enough. AI assistants require a more nuanced understanding and response mechanism to respond to diverse and complex user queries. This is why the supervised fine-tuning phase becomes essential.

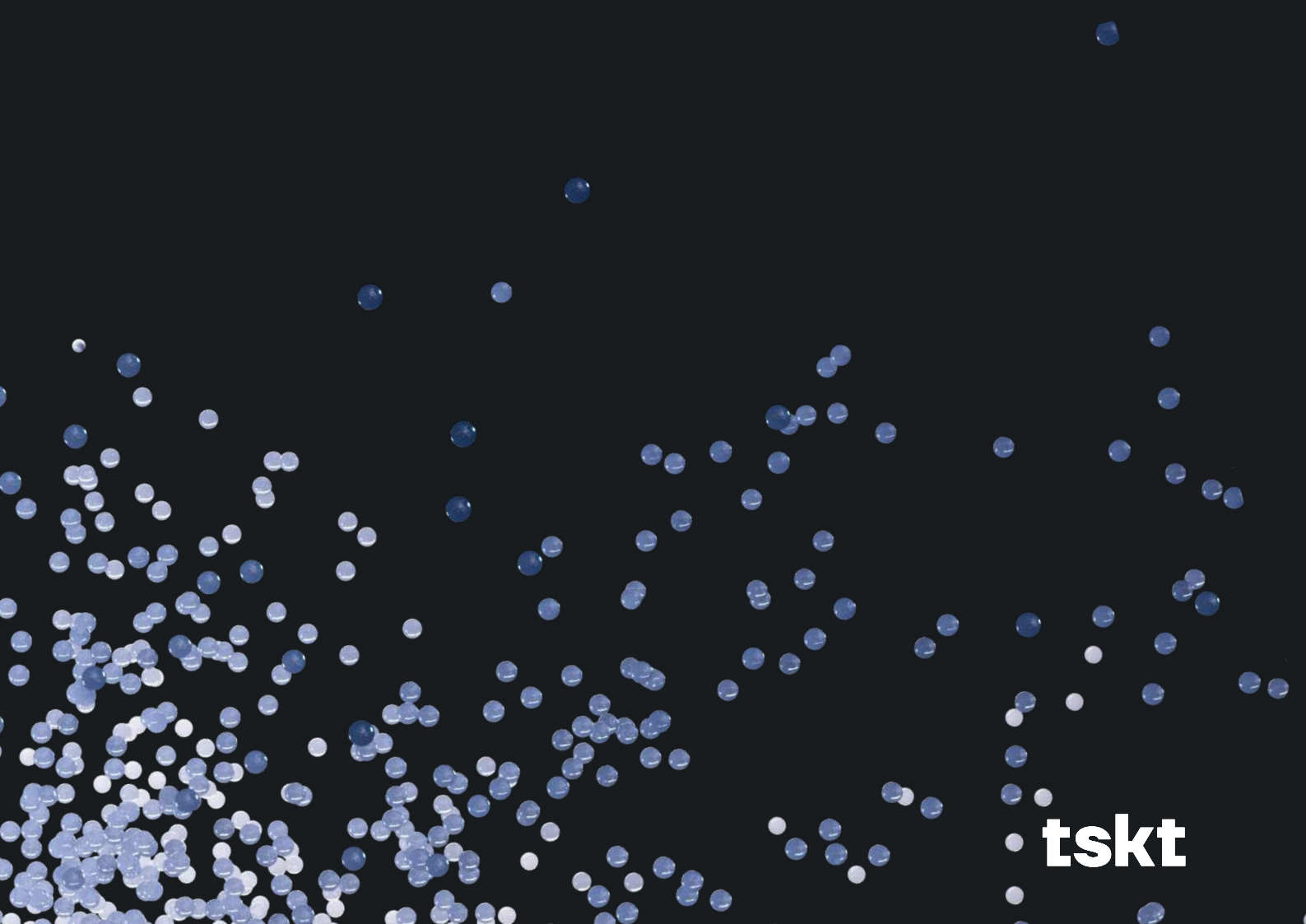
To transform the base LLM generative model to an interactive assistant, we retain the base model weights, and introduce a different set of training data. This new dataset is comprised of user prompts and their corresponding "ideal" answers to questions. The process of creating such pairs of prompts and responses is known as labeling data. Consequently, the model, which has already been exposed to a broad 10TB dataset during pretraining, undergoes a fine-tuning process with this more specialized conversational data.

Unlike the diverse and extensive pretraining data, the conversational dataset is typically smaller, manually curated, and of superior quality. In this phase, the emphasis is on the quality of data over the sheer amount. Typically, this involves refining the model's responses with around 100,000 high-quality conversations, ensuring it can provide helpful and contextually appropriate answers.

It's important to highlight that as Large Language Models (LLMs) become more sophisticated, the process of labeling data gets progressively easier. With more capable models, the initial drafts of responses can be generated by the LLMs themselves, leaving to the humans only the workload of reviewing and refining these suggestions. Consequently, the more advanced the model, the less human effort is required for labeling data. This efficiency gain not only speeds up the process of improving the model but also enhances the accuracy of the data used for further training and fine-tuning.

CHAPTER 4

CUSTOM DATA ASSISTANTS



tskt

Although the LLMs possess a wealth of knowledge, they are not databases and can't be simply updated with custom information. While they have been trained on a vast corpus, providing a broad understanding of language and reasoning, their knowledge is limited to publicly available information, lacking in private, corporate, or use-case specific data.

Also, even though the models are continuously updated, they can't be updated that often because it is too resource demanding task. So, for example the most recent GPT models contain data up to April 2023.

Grounding

01

Grounding is the process of using Large Language Models (LLMs) with information that is use-case specific, relevant, and not available as part of the LLM's trained knowledge [3]. There are two main methods for grounding.

Custom Data Fine-tuning

Custom Data Fine-tuning, on the other hand, involves an additional training step on the custom datasets to create a new version of the model. Through this process the model becomes more proficient in generating relevant responses in particular contexts (healthcare, finance, or anything else).

Retrieval Augmented Generation (RAG)

Retrieval Augmented Generation (RAG) is a process for retrieving information relevant to a task, providing it to the language model along with a prompt, and relying on the model to use this specific information when responding.

Retrieval Augmented Generation (RAG)

02

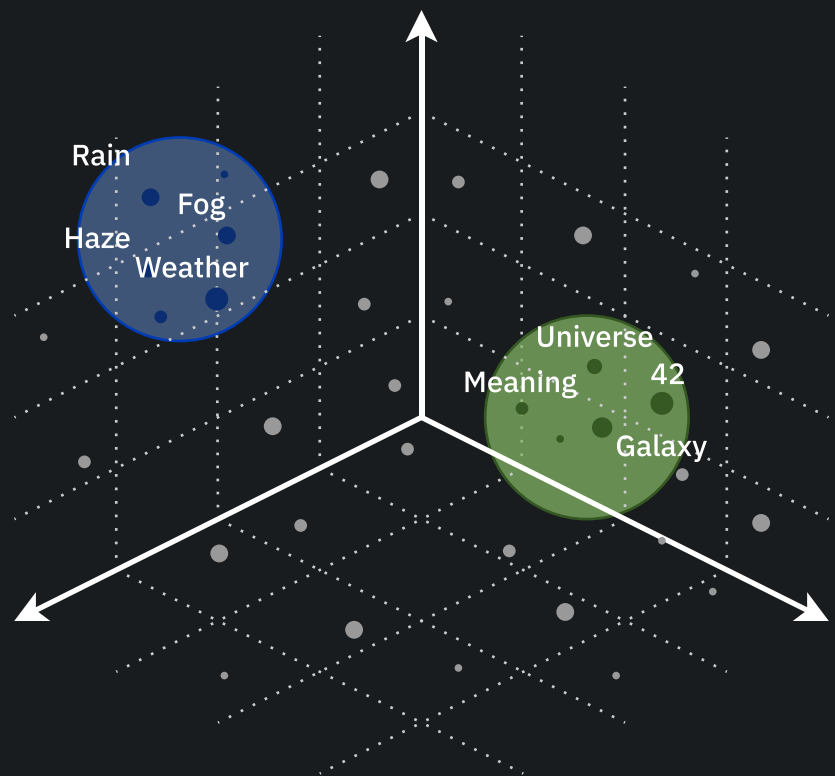
Retrieval Augmented Generation (RAG) for Large Language Models (LLMs) starts with a user query, which triggers a retrieval process. Relevant information is fetched from a database and combined with the initial query, giving it an additional context. Then, this enriched input is used by the LLM to generate a response, making it more accurate and relevant.

In essence, RAG extends the capabilities of the language model by enriching their responses with information sourced from external databases, thereby improving the relevance and accuracy of the model's output in various applications. This makes RAG a powerful tool for grounding LLMs in specific information and contexts.

Semantic Search

Semantic search has emerged as a key technique in Retrieval Augmented Generation (RAG), although it's not the only method used. This approach indexes documents or document fragments by converting their semantic content into internal model representation – embeddings mentioned above.

Embeddings are essentially vector representations of data. In the context of large language models, this data is textual. Imagine a 3D space where related concepts, such as "fog" "rain" and "haze" are positioned close to each other.



While our example simplifies embeddings to a three-dimensional space for ease of understanding, in reality, these embeddings are far more complex, usually consisting of hundreds or even thousands of numerical values.

The primary purpose of organizing text into this vector space is to facilitate efficient and accurate semantic searches. When a user enters a query, the system employs these embeddings to locate texts that have semantically similar content.

Vector Databases

Vector databases and indexes are modern tools that have become increasingly important in fields like natural language processing and data management. These systems are specifically designed to store documents and index them in the form of vector representations or embedding, which allows making efficient similarity searches and document retrieval.

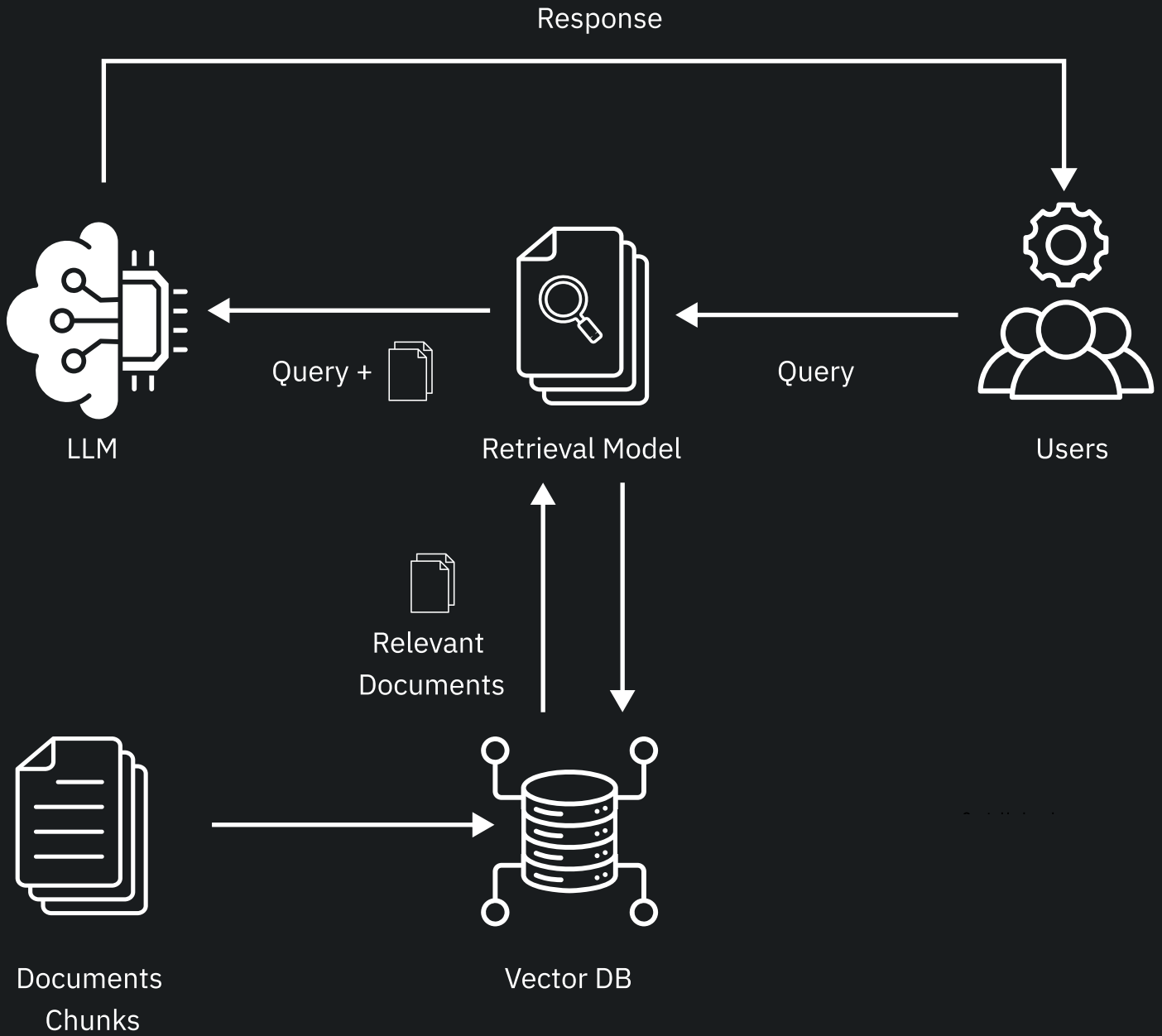
There is a wide variety of vector databases and products available, each offering unique features and capabilities. Among the popular choices, Chroma DB stands out as a light and open-source solution for vector databases, making it an ideal choice for smaller-scale projects.

This process involves performing a similarity search, where the semantic representation of the user's query is compared against the available documents in the vector database to identify the most relevant ones. By leveraging this method, we can effectively bridge the gap between a query and the vast information stored within the database, ensuring quick and relevant results.

As for cloud-based solutions, Azure Cognitive Search is a noteworthy option, offering seamless integration with the existing Microsoft Cloud infrastructure.

When selecting a vector database or index, it's essential to consider factors such as scalability, performance, ease of use, and compatibility with existing infrastructure. The choice largely depends on the scale of the data as well as the specific requirements of the project.

Altogether, the process of Retrieval Augmented Generation (RAG) looks as follows



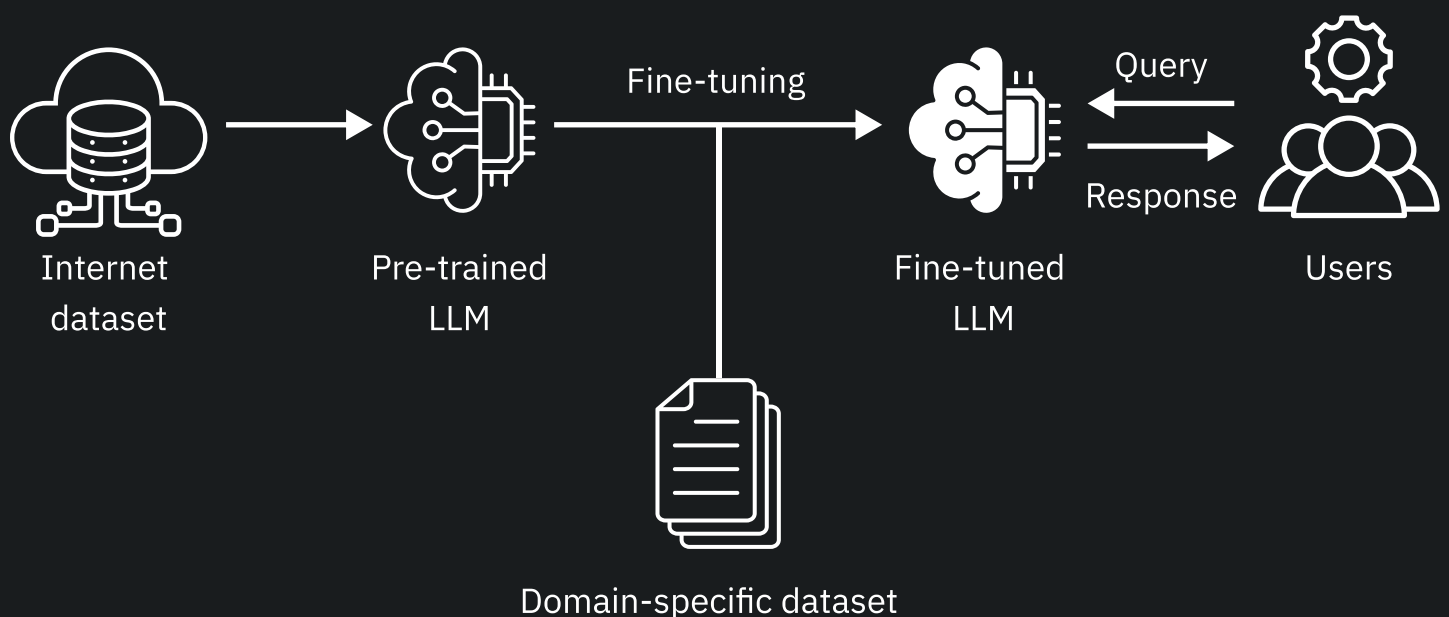
Custom Data Fine-tuning 03

Custom data fine-tuning is a crucial phase in the development of large language models, typically employed to achieve specific behaviors such as those required for a chatbot assistant. This approach can also be effectively utilized for grounding, a process that allows the model to integrate and utilize external information or databases. By adjusting certain parameters, fine-tuning optimizes the model's performance for specific tasks or a range of tasks.

Full fine-tuning

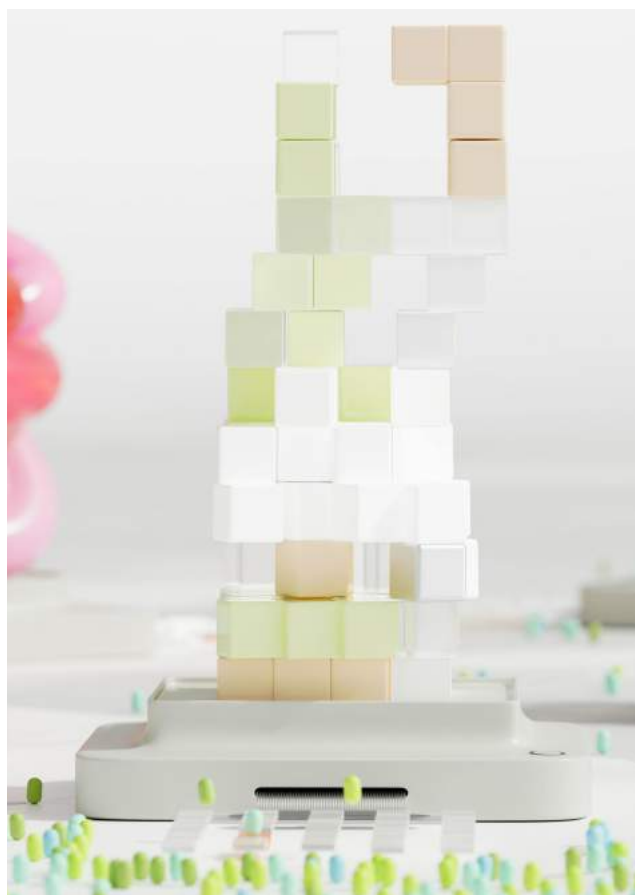
This method involves updating all parameters of a large model. Due to the complexity and size of modern LLMs, which may have tens or hundreds of billions of parameters, even a relatively small fine-tuning dataset can result in significant computational costs.

Full fine-tuning



Parameter-Efficient Fine-Tuning (PEFT)

Parameter-Efficient Fine-Tuning (PEFT), on the other hand, involves updating only a subset of the total parameters of a pre-trained model. This strategy can be applied to both pre-trained and previously fine-tuned models to further adapt them for specific tasks. As these models already encompass a broad spectrum of knowledge, it is often unnecessary and inefficient to adjust the entire model for specific tasks. Therefore, PEFT strategically focuses on fine-tuning a limited set of parameters, making it a more resource-efficient option.



Ultimately, the choice of approach for grounding a large language model depends on the specific use case and requirements. Retrieval-augmented generation should be considered initially, as it can frequently minimize the need for extensive fine-tuning by supplementing the model with external knowledge. When fine-tuning is indeed necessary, the PEFT approach is generally sufficient for most tasks, since it efficiently updates only a subset of the model's parameters, making it a practical and resource-conscious choice for fine-tuning needs.

CHAPTER 5

GPT AND LLAMA COMPARISON



tskt

In the landscape of Large Language Models (LLMs), Generative Pre-trained Transformer (GPT), developed by OpenAI and LLaMA (Large Language Model Meta Artificial Intelligence) represent two prominent approaches. GPT stands out as the most advanced and accurate model among proprietary offerings, known for its sophisticated algorithmic capabilities and extensive dataset training.

In the open-source sector, LLaMA has established itself as a fundamental framework, serving as the backbone for numerous models in this space. Our aim is to provide a detailed understanding of where each model excels and the potential trade-offs involved, offering valuable insights into the current state of LLM technology.

Accuracy of Responses

01

When comparing the effectiveness of GPT and LLaMA, a key factor to consider is the accuracy of their responses. GPT models, particularly the latest iterations like GPT-4, are known for their highly accurate and contextually relevant responses. They are trained on vast and diverse datasets, enabling them to handle a wide range of topics with a high degree of proficiency.

LLaMA, on the other hand, is an open-source alternative that has shown promising results in terms of accuracy. While it may not have the same level of data access as GPT, it's designed to be adaptable and efficient, making it a strong competitor in terms of response quality.

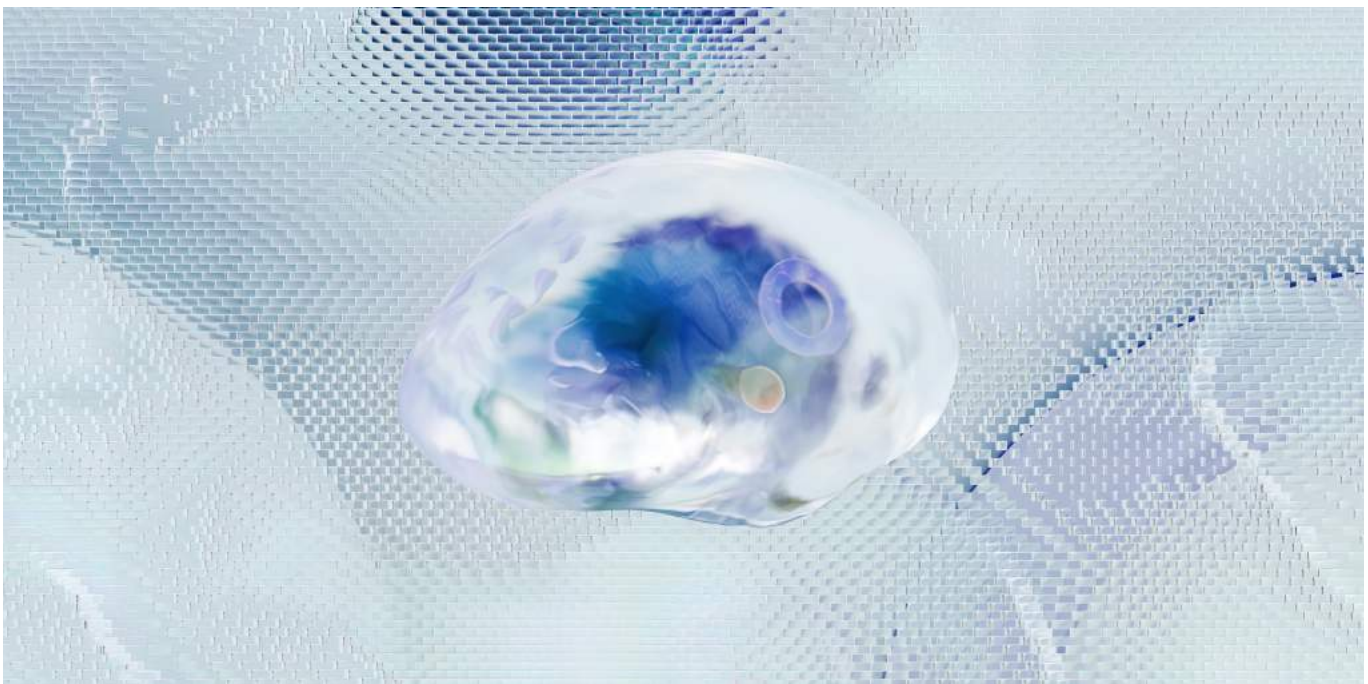
However, the actual effectiveness can vary based on the specific model of LLaMA used and the nature of the task it is applied to.

Grounding: Ease and Accuracy

02

The grounding process for Generative Pre-trained Transformer (GPT) models is generally straightforward, particularly the retrieval-augmented generation approach. GPT models, due to their large size and expansive context windows, are highly suitable for complex RAG applications, demonstrating high accuracy in such scenarios. The ease of integrating RAG with GPT models further enhances their applicability in a variety of contexts, allowing them to effectively pull in and synthesize information from external sources.

LLaMA also supports RAG effectively, particularly in its larger models. These versions handle RAG well, maintaining effectiveness in applications that require integrating additional context. However, the smaller LLaMA models face challenges in processing requests with extensive external context. The effectiveness of RAG in these smaller models is less consistent, presenting potential difficulties in scenarios requiring detailed and context-heavy information retrieval.

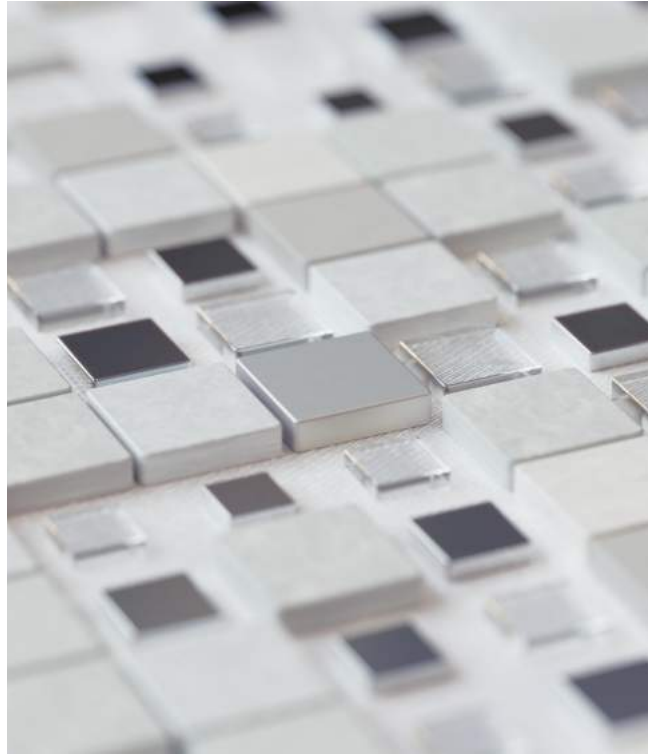


Deployment: Local vs Online Deployment

03

Deployment and maintenance are crucial aspects of any LLM. GPT, typically accessed through OpenAI's API, is an online, cloud-based service. This means it requires a continuous internet connection and relies on OpenAI's infrastructure for updates and maintenance. The benefits include ease of access, regular updates, and minimal local infrastructure requirements.

LLaMA, in contrast, offers more flexibility in deployment. It can be deployed locally, which allows for greater control over data and privacy.



However, local deployment comes with its own set of challenges, including the need for substantial computational resources, expertise in model deployment and maintenance, and responsibility for updates and security.

“ Deployment and maintenance are crucial aspects of any LLM

Maintenance: Price and Complexity

04

The cost of maintaining a GPT-based chatbot largely depends on usage, as OpenAI typically charges based on the number of tokens processed. For high-volume or frequent use, this can add up to a significant cost. However, the trade-off is access to a state-of-the-art model without the overhead of managing infrastructure which makes it an easy and straightforward approach.

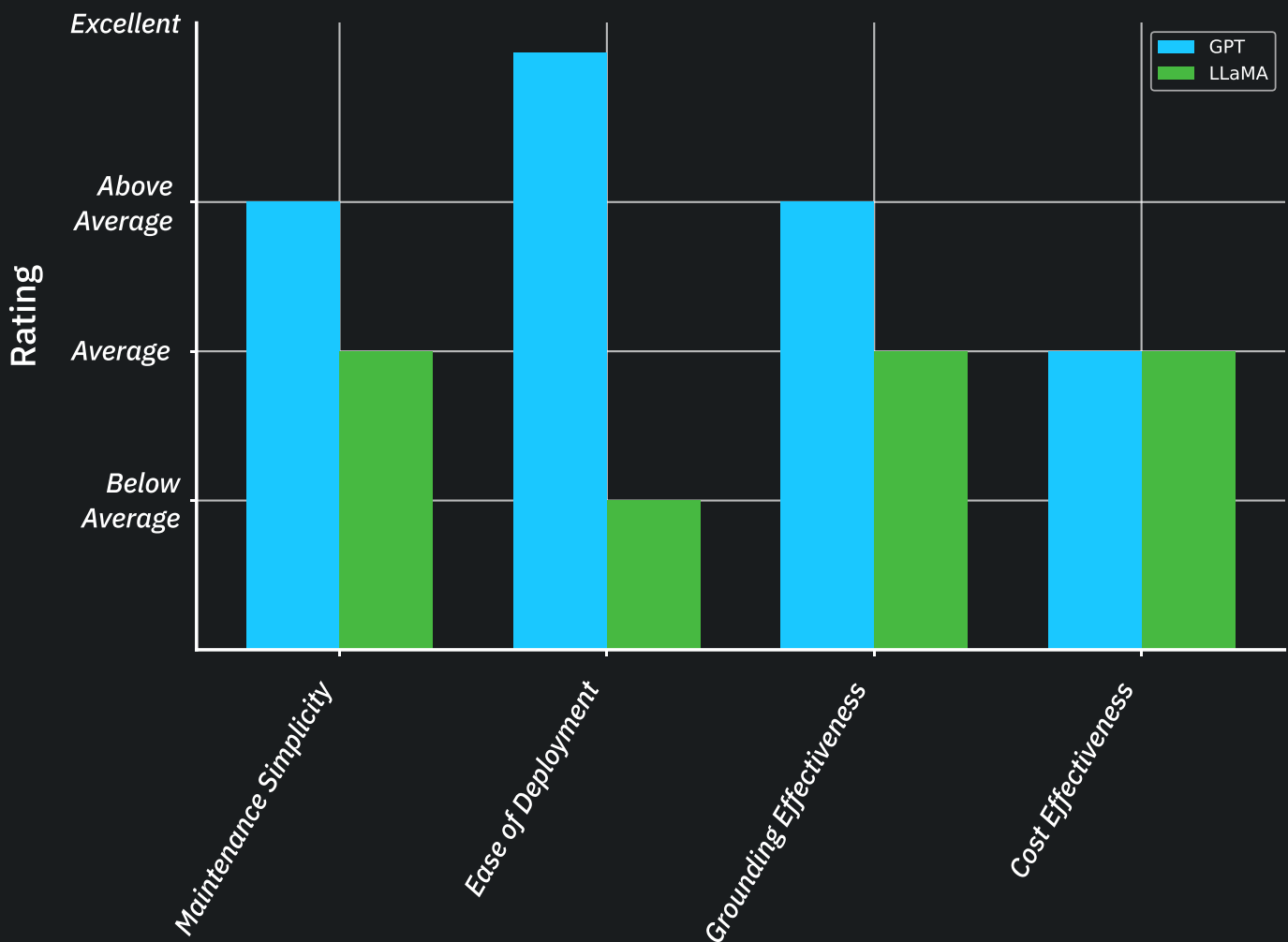
Maintaining a LLaMA-based chatbot can potentially be more cost-effective, especially if deployed locally, as there are no direct costs associated with API calls. Although, the initial setup costs, including hardware and any required development or model customization, can be substantial. Additionally, ongoing costs related to infrastructure and maintenance should also be considered.

Generally, hosting a self-managed model presents a more complex undertaking, demanding careful consideration of various aspects.



The following graph represents the high-level comparison of the aforementioned properties in the average case. It's important to note that cost-effectiveness largely depends on the specific use case. Depending on the requirements, one approach may be significantly more affordable than the other.

LLMs Comprasion



In terms of security and privacy there are differences, depending on the LLM and the hosting type. Here is the brief description focusing on OpenAI, Azure OpenAI, that is an alternative service that enables hosting OpenAI models within Azure Infrastructure, and self-hosted models like LLaMA.



OpenAI

Compliant with GDPR OpenAI's framework includes SOC 2 Type 2 certification and annual third-party penetration tests. It processes user data in the U.S., adhering to legal obligations and privacy policies. Data is used for service administration and legal compliance, but not for model training.



Azure OpenAI Service

This service provides robust data privacy and security. It is hosted entirely within Microsoft's Azure environment, without interaction with OpenAI's services, and ensures customer data isn't shared or used for model improvement. Features include stateless models, content filtering, abuse monitoring, and customer-managed encryption keys.



Self-Hosted LLMs

Suitable for offline use, these models offer control over the deployment environment and reduce network-based threats. They require comprehensive local security measures, including physical access control and internal network security.

Each option should be chosen based on the organization's data privacy requirements, legal compliance needs, and specific LLM use-case.

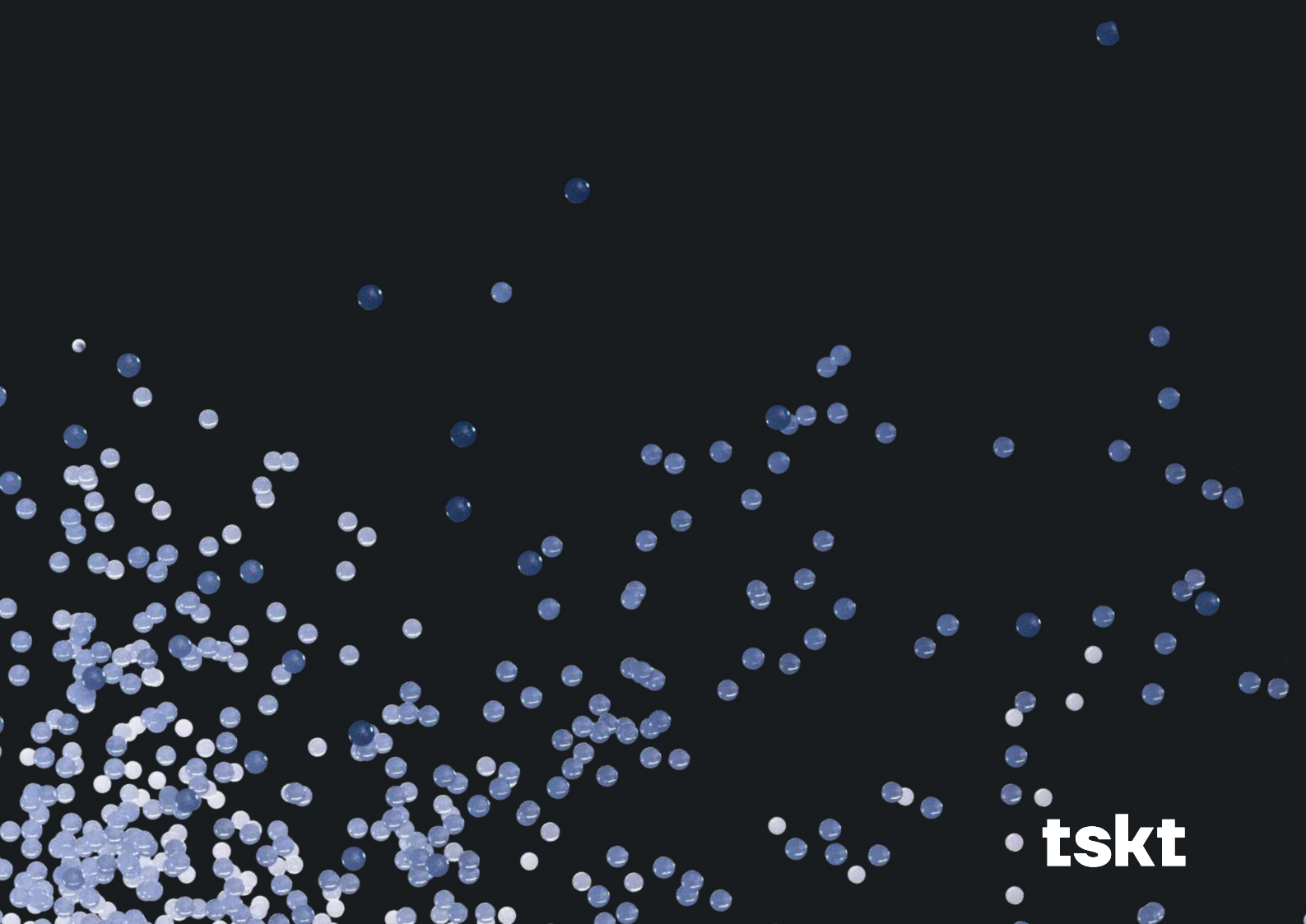
In conclusion, both GPT and LLaMA have their strengths and considerations. GPT stands out for its ease of use, regular updates, and high-quality responses, though at a potentially higher operational cost. LLaMA offers flexibility, cost-effectiveness, and customization opportunities, but with additional requirements in terms of technical expertise and infrastructure. The choice between the two depends on the specific needs, resources, and goals of the business.



“ The choice depends on the specific needs, resources, and goals of the business.

CHAPTER 6

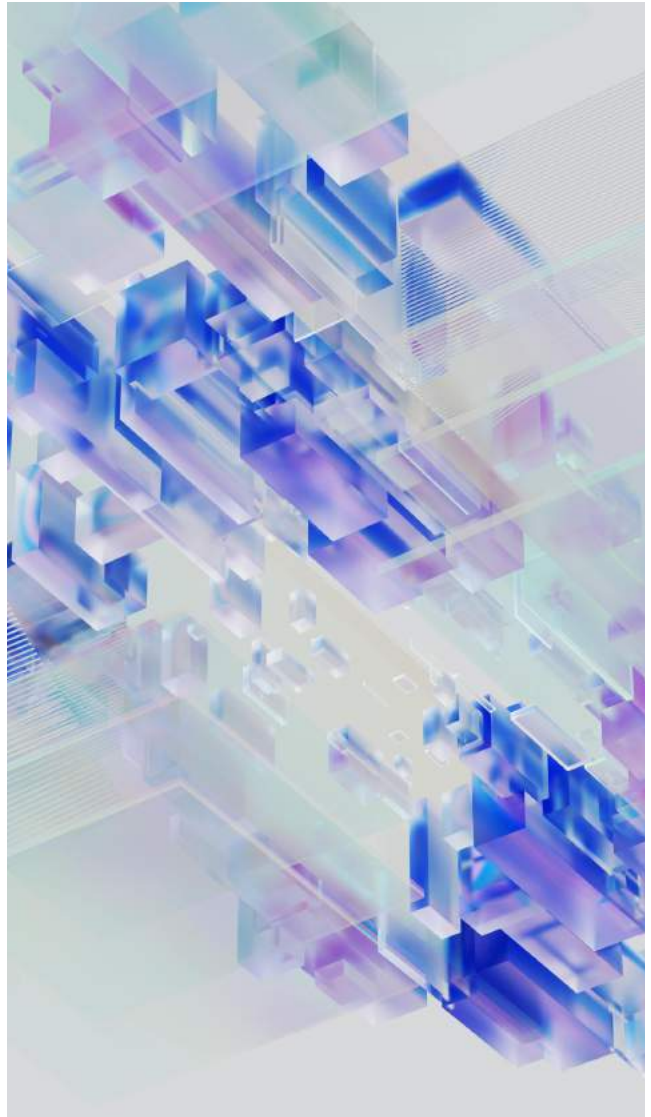
LLMS IN ACTION



tskt

In our research, we focused on comparing two lower-tier language models: GPT-3.5 and LLAMA2 13B. These models were selected for their balance of speed and accuracy, making them suitable candidates for real-time web assistant applications. Our goal was to evaluate their performance in handling company-specific inquiries, particularly for our organization, **tsukat**. To facilitate a comprehensive comparison, we examined the models' behavior both before and after grounding via retrieval-augmented generation.

We used Chroma vector database to store details about the company. This database was then utilized in conjunction with the RAG framework, enabling the language models to access and incorporate this specific company information into their responses.



We used similar initial prompts for both models with a difference of the model specific syntax: “You are an AI assistant of tsukat company. You should be friendly answering the question regarding tsukat company to the best of your knowledge. If you need additional clarification, you should ask about it. Don't make up the answer if you are not sure it is correct.”

The following is the example of the conversation with the models before and after implementing RAG:

| | LLaMA2 13B | GPT-3.5-Turbo |
|---|--|---|
| Before grounding | | |
| Q1: "Tell me about tsukat." | Sure, I'd be happy to help! Tsukat is a leading provider of innovative technology solutions and services. <Hallucinations> | Tsukat is a company that specializes in [Please provide more specific information or clarify your question]. I'd be happy to provide you with more details, but I need a bit more information about what aspect of Tsukat you're interested in. Are you looking for information about its products, services, history, or something else? Please let me know, and I'll provide you with a more specific answer. |
| Q2: "Tell me about your VR projects". | Sure, I'd be happy to tell you about our VR projects! At Tsukat, we have been exploring the potential of virtual reality (VR) technology in various industries. <Hallucinations> | I apologize for any confusion, but as of my last knowledge update in January 2022, I don't have specific information about Tsukat's VR projects or any updates beyond that date. |
| Q3: "Do you work with AI?" | Yes, we do work with AI! At Tsukat, we have been exploring the potential of artificial intelligence (AI) in various industries. Here are some examples of our AI projects: <Hallucinations> | To provide you with accurate information, I would need more specific details about Tsukat's involvement with AI. If you could provide more context or specify the aspect of AI you're interested in regarding Tsukat, I'll do my best to provide relevant information based on my knowledge up to January 2022. |
| Q4: "Can you create a VR training project?" | Yes, we can definitely create a VR training project! At Tsukat, we have experience in developing immersive VR experiences for various industries. <Hallucinations> | While I don't have specific information about Tsukat's current projects or capabilities, if Tsukat is a company that specializes in VR development, it is possible that they have the expertise and resources to create VR training projects. If you are interested in Tsukat's potential involvement in creating a VR training project, I recommend reaching out to Tsukat directly to discuss your specific requirements and see if they have the expertise and interest in undertaking such a project. |

After grounding

| | | |
|--|---|---|
| <p>Q1: "Tell me about tsukat."</p> | <p>Sure! Tsukat is a virtual reality, augmented reality, and artificial intelligence company that specializes in developing solutions for businesses. We have been in this industry for almost seven years and currently have a team of around 80 people with production offices located in Portugal, Poland, Italy, and Lviv, Ukraine. From a legal standpoint, we are a US-registered company but have multiple production offices worldwide. Our expertise includes developing digital twins for cities and creating immersive virtual reality experiences for various industries such as real estate, education, and healthcare. We also offer AI and machine learning solutions to help businesses automate their processes and make data-driven decisions.</p> | <p>I'd like to introduce our studio, Tsukat. We specialize in designing and developing virtual reality, augmented reality, and more recently, artificial intelligence and machine learning solutions for businesses. We've been in this industry for almost seven years. Currently, we have a team of about 80 people, with production offices located in Portugal, Poland, Italy, and Lviv, Ukraine. From a legal standpoint, we are a US-registered company, but we have multiple production offices worldwide. Is there anything specific you would like to know about our studio?</p> |
| <p>Q2: "Tell me about your VR projects".</p> | <p>Sure! Tsukat has an extensive collection of Virtual Reality (VR) training projects that have gained traction due to their efficacy. Our studies suggest that our VR trainings are at least 40% more efficient than traditional methods, with some results indicating a higher percentage of efficiency. We develop digital environments primarily for business applications, where clients can conduct various activities such as meetings, presentations, and sales activities. One notable example of our Metaverse and Digital Twins expertise is a project where we developed a digital twin for a section of a city in the US. City officials use this digital twin to conduct business meetings, narrate stories, and highlight opportunities available for investing within the city.</p> | <p>Our VR projects fall under our Interactive XR Solutions category. We have developed applications with augmented reality portals that allow users to step inside and view a 360-degree video of a brand. These applications are designed for both online and offline events. Additionally, we collaborate with clients to create more expansive products, such as a mobile app that lets users scan their face to generate a textured 3D model for crafting custom-fit frames. This product has been refined over the past four years.</p> |

| | | |
|--|---|--|
| <p>Q3: “Do you work with AI?”</p> | <p>Yes! Tsukat works extensively with artificial intelligence (AI) and machine learning (ML). Our technical director, a former Microsoft employee who developed computer vision modules for products like HoloLens and Kinect Windows, leads our AI and ML solutions. We have a robust team of AI engineers who tackle complex R&D tasks, including natural language processing systems trained on client-specific data, generative image systems, and computer vision projects such as facial and object tracking. Our AI and ML solutions help businesses automate their processes and make data-driven decisions.</p> | <p>Yes, we do work with AI. Our technical director, a former Microsoft employee, leads our AI and ML Solutions team. They have developed computer vision modules for products like HoloLens and Kinect Windows. Our team of AI engineers tackle complex R&D tasks, including natural language processing systems trained on client-specific data, generative image systems, and various computer vision projects such as facial and object tracking.</p> |
| <p>Q4: “Can you create a VR training project?”</p> | <p>Yes! Tsukat has experience creating virtual reality (VR) training projects. We have developed several successful VR trainings with feedback indicating an efficiency close to 60 or 70% when compared to conventional training methods. Our studio crafts immersive and realistic VR experiences that simulate various scenarios, allowing trainees to practice and hone their skills in a safe and controlled environment. We can create customized VR training projects tailored to the specific needs of your organization. Please let us know if you have any further questions or would like to discuss this further.</p> | <p>Yes, we do work with AI. Our technical director, a former Microsoft employee, leads our AI and ML Solutions team. They have developed computer vision modules for products like HoloLens and Kinect Windows. Our team of AI engineers tackle complex R&D tasks, including natural language processing systems trained on client-specific data, generative image systems, and various computer vision projects such as facial and object tracking.</p> |

In our observations, the LLaMA model demonstrated a higher tendency to produce hallucinations, particularly in scenarios where it did not possess sufficient information to provide an accurate answer. In the context of language models, hallucinations refer to outputs that are coherent and fluent in structure but contain factually incorrect information. This occurred despite explicit instructions in the prompt to refrain from responding if the model was uncertain or lacked knowledge on the subject. To prevent the dissemination of incorrect information, such instances in the model's responses are labeled as <Hallucinations>. In contrast, GPT-3.5 consistently maintains accuracy in its responses, effectively avoiding the provision of misleading information.

Following the RAG integration, we observed noticeable improvements in the models' responses. They correctly answered the question using data from the vector database and gave a helpful and accurate response.

In terms of response speed, it is highly dependent on the hardware used. In our setup with RTX 3070 Ti GPU the LLaMA model was quite slow and was considerably outperformed by the GPT-3.5 model.

The accuracy of responses in the specific context we tested was quite similar between LLaMA and GPT-3.5, despite GPT's broader knowledge base. This indicates that RAG integration helps narrow the accuracy gap in specific contexts, making both models comparably reliable in their responses.

CHAPTER 7

BENEFITS FOR BUSINESSES: WHY CHOOSE LLMS?



tskt

In today's fast-changing digital business world, companies are always looking for ways to stay ahead of the competition. Large Language Models (LLMs) are becoming a popular choice because they offer many benefits for different business needs. Among them are:

Scalability

LLMs offer significant advantages to businesses by assisting both external customers and internal employees. Capable of managing a multitude of interactions, from responding to customer inquiries to processing orders, these models maintain a high level of consistency and quality, regardless of the volume. Internally, AI assistants prove invaluable in simplifying interactions with company documentation, thereby boosting employee productivity. This capability not only diminishes the need for extensive customer service teams but also optimizes internal processes, leading to substantial cost savings and enhanced operational efficiency.

Customization

Each business is unique, with its own set of needs, challenges, and customer expectations. LLMs offer the flexibility of grounding the models, allowing companies to tailor the responses to align with the brand specific products, or targeted customer needs. This ensures that every interaction resonates with the brand's identity and values.

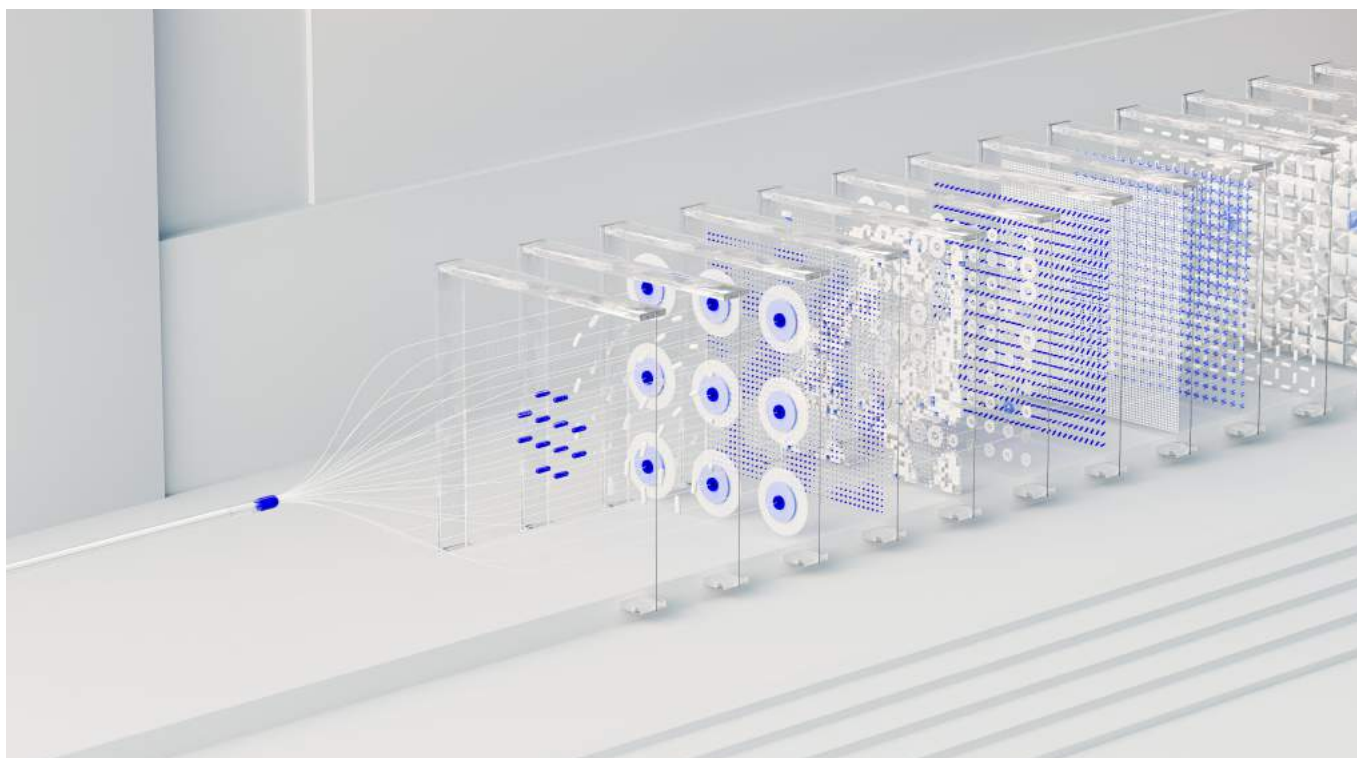


Facilitating Translation and Localization

LLMs extend their utility by offering automatic localization capabilities, making them essential for businesses with a global presence. They enable region-specific analysis and insights, ensuring culturally relevant and applicable interpretations of data. This is particularly advantageous for companies operating across multiple countries or regions, as it not only provides accurate translations but also saves on the costs associated with professional translation services. This feature marks a significant stride in making LLMs invaluable for international business operations.

Efficient Data Processing and Analysis

Businesses inundated with extensive data sets can benefit greatly from LLMs for rapid and proficient data analysis and interpretation, crucial for informed decision-making. These models are particularly skilled in tasks such as market trend analysis, decoding customer feedback, and executing competitive research, while also excelling in document processing. The efficiency of LLMs in handling such diverse data processing tasks makes them invaluable tools for businesses seeking to leverage data for strategic insights.



Conclusion

The realm of Large Language Models (LLMs) is constantly improving, with new enhanced model versions being regularly announced. This continuous evolution means that language models are becoming increasingly adept at handling complex tasks, offering scalable and efficient solutions for both customer-facing and internal processes.

Customization is a key feature of LLMs, allowing these models to align closely with your brand identity and precisely cater to specific customer needs. Additionally, they can be configured to address industry-specific queries, provide customized product recommendations, or handle unique customer service scenarios, making each interaction highly relevant and specific to your business's offerings.

If you are looking for an AI-driven solution, we invite you to explore the potential of LLMs for your business. Whether you aim to enhance customer engagement, streamline internal processes, or gain deeper insights from your data, our LLM solutions are designed to meet these needs.

Looking ahead, the trajectory of LLM technology promises even more sophisticated and versatile applications. As these models continue to evolve, they will offer businesses unprecedented opportunities to innovate, improve customer experiences, and stay competitive.

tsukat tech leads have worked with cutting-edge AI technologies for over ten years, ensuring high-quality and efficient products for various industries, from healthcare and manufacturing to retail and automotive.

References

[1] Ashish Vaswani et al., 2023, "Attention Is All You Need".
arXiv:1706.03762 [cs.CL]

[2] Hugo Touvron et al., 2023, "LLaMA: Open and Efficient Foundation Language Models". arXiv:2302.13971 [cs.CL]

[3] Eleanor Berger, 2023, Grounding LLMs, Microsoft Blog

About tsukat

We are a multidisciplinary team of XR experts, passionate about cutting-edge interactive, VR, AR, AI technologies and their massive impact on the modern world. The practical experience of our tech leads in this field exceeds 10 years, and during this time, they have been diligently working with projects of various complexity and duration. This has ultimately allowed us to permanently involve and use next-generation technologies for the needs of our clients.

Clutch

CLUTCH'S TOP 10 AR/VR COMPANIES

10+

years of commercial
experience define
each of our tech leads

80+

in-house talents

90%

of clients highlight our
transparent cooperation
model and communication

81%

NPS (Net Promoter Score)

How we different:

IDEA IMPLEMENTATION

No matter if you have just an idea or a detailed plan we can implement it for you

NARROW FOCUS

Focused solely on VR, AR, AI technologies

TOP 10 ON CLUTCH

Top 10 Virtual Reality Companies and Augmented Reality Developers by Clutch

KEEP DEADLINES

Proven accuracy in estimation

PROFICIENCY AND TRANSPARENCY

By utilizing agile methodologies, the engineering process is fully transparent. This includes visibility from the release roadmap down to daily activities.

READY TO TURN YOUR IDEAS INTO POWERFUL TOOLS THAT ENHANCE YOUR BUSINESS?

Connect with tsukat to unlock the potential of immersive technologies and discover how they can transform your business today.

CONTACT US

hello@tsukat.com

+17866001713

4300 Biscayne blvd, Suite 203, Miami,
Florida 33137, USA.

www.tsukat.com

tskt