Selection of large language model for development of interactive chat bot for SaaS solutions

Olga Cherednichenko[1,†], Dmytro Sytnikov[2,†], Nazarii Romankiv[3,*,†], Nataliia Sharonova[4,†], and Polina Sytnikova[5,†]

[1] *Univ Lyon, Univ_Lyon 2, UR ERIC – 5 avenue Mendès France, 69676 Bron Cedex, France*
[2] *National University of Radio Electronics, 14, Nauki prospect, Kharkiv, 61166, Ukraine*
[3] *National University of Radio Electronics, 14, Nauki prospect, Kharkiv, 61166, Ukraine*
[4] *National Technical University, "Kharkiv Polytechnic Institute", 2, Kyrpychova str., Kharkiv, 61002, Ukraine*
[5] *National University of Radio Electronics, 14, Nauki prospect, Kharkiv, 61166, Ukraine*

Abstract

Chat Bots play crucial role in modern world businesses. Development of a chatbot is a tedious and complex task that takes enormous amount of time. Therefore, to enable businesses to develop chatbots in short amount of time with small resources, we need to explore usage of LLMs for the chatbot development. In this article the researchers developed a prototype chatbot architecture that enables businesses to use LLMs interchangeably. Nonetheless, it notes the prototype's limitation in tracking conversation history, an area ripe for future enhancement. The current focus on pre-trained LLMs sets the stage for subsequent research into personalized, fine-tuned chatbot experiences for SaaS customers. There are dozens of LLMs existing already, therefore it is crucial to select the most capable LLM which will power the chatbot. It is crucial for the LLM to be cost-efficient, to be profitable for the business. This article evaluates three LLMs, endorsing ChatGPT for its superior speed, cost-effectiveness, and relevance, backed by OpenAI's pioneering status.

Keywords

ChatBots, LLMs, SaaS, Langchain, Python, Node.js, Telegram Bot

CEUR-WS.org/Vol-3722/paper6.pdf

1. Introduction

Modern technologies and the Internet have enabled Software as a Service (SaaS) [1] models to emerge and dominate the global landscape; numerous multibillion-dollar companies have been built on this model, such as Netflix, Amazon, Facebook, and Airbnb. According to 360 Research Reports, the SaaS market size was valued at approximately 1,777,098 million dollars in 2021 and is expected to continue growing, reaching an estimated 1,177,188 million dollars by 2027. Developing a contemporary SaaS application presents numerous technical challenges that need to be addressed, especially today with the widespread adoption of cloud technologies and the emergence of new cloud-native [2] solutions. Moreover the global chatbot market is projected to reach 2 billion dollars by end of 2024, growing at a CAGR (compound annual growth rate) of 29.7%.

Furthermore, the MACH Alliance [3] and the use of composable e-commerce [4] architectures are gaining significant popularity. This approach allows businesses to construct their systems like a modular kit, assembling various SaaS systems where each system is responsible for different e-commerce technical capabilities, such as search, personalization, and order management, among others.

CLW-2024: Computational Linguistics Workshop at 8th International Conference on Computational Linguistics and Intelligent Systems (CoLInS-2024), April 12–13, 2024, Lviv, Ukraine

[*] Corresponding author.

[†] These authors contributed equally.

olga.cherednichenko@univ-lyon2.fr (O. Cheredmnichenko); dmytro.sytnikov@nure.ua (D. Sytnikov); ayzrian@gmail.com (N. Romankiv); nvsharonova@ukr.net (N. Sharonova); polina.sytnikova@nure.ua (P. Sytnikova)

0000-0002-9391-5220 (O. Cheredmnichenko); 0000-0003-1240-7900 (D. Sytnikov); 0009-0004-9893-6823 (N. Romankiv); 0000-0002-8161-552X (N. Sharonova); 0000-0002-6688-4641 (P. Sytnikova)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

In the dynamic landscape of technological advancements, the integration of conversational interfaces has emerged as a critical facet for Software as a Service (SaaS) solutions. Within the SaaS business framework, chatbots play a crucial role in enhancing customer engagement, providing real-time support, and automating routine tasks. The integration of chatbots is not solely a technological trend but a strategic imperative, offering businesses the opportunity to optimize operations, reduce response times, and ultimately improve customer satisfaction.

Moreover, Chatbots can collect and analyze vast amounts of data from interactions with users. This capability allows SaaS companies to gain insights into user behavior, preferences, and feedback. The analysis of this data can inform product development, improve user experience, and guide strategic decision-making. In scientific terms, this iterative process of hypothesis, experimentation, and validation is fundamental to both product improvement and understanding user dynamics. As well as offer personalized experiences to users by leveraging AI and machine learning algorithms. This personalization can extend to product recommendations, support solutions, and interactive experiences tailored to the individual user's behavior and preferences.

Emerging technologies like Large Language Models [5] are revolutionizing the way we interact with an AI, since LLMs provide rich set of capabilities, like contextual awareness, adaptability to user preferences, language proficiency and understanding user queries with a depth, multilingual capabilities etc. Therefore, LLMs can be a great choice for integration into SaaS applications to deliver truly unique customer experiences with chat bots.

2. Related works

In the research [6], Anne L. Roggeveen at al. and Raj Sethurma highlighted that Chatlines (like chatbots but featuring with a human presence on the seller side) help retailers engage in person-to-person digital conversations, such as on a website, and at least half of Fortune 5000 companies have experimented with chatbots, hoping to reduce costs and improve response quality. Though this work does not get into details of chatbot technology selection.

Literature review [7] underlines different weaknesses of a Chatbot agent in comparison with a human such as limited understanding and knowledge, limited empathy and emotions. Though it worth to mention that the review was based on literature produced between 2020 – 2021 and therefore before huge breakthrough of LLM. LLMs have a great potential to overcome the limitations mentioned in the literature review.

Evaluating chatbots and natural language generation is a well-known challenge and research [8] proposed to use human likeness of conversational responses to evaluate chatbots. They describe two human-evaluation setups: static, in which they benchmark models on a fixed set of multi-turn contexts to generate responses; and interactive, where they allow humans to chat freely with chatbots. They also introduce two key metrics: sensibleness, specificity. Moreover the researchers built their own multi-turn open-domain chatbot called Meena, and compare it with other technologies, though since the article was written at 2020 none of the modern LLMs are mentioned.

Researchers in [9] highlight challenges that chatbot applications are tightly coupled to their intent recognition providers, hampering their maintainability, reusability and evolution. Typically, once the chatbot designer chooses a specific chatbot development platform, she ends up in a vendor lock-in scenario, especially with the NL engine coupled with the platform. Similarly, current chatbot platforms lack proper abstraction mechanisms to easily integrate and communicate with other external platforms the company may need to interact with. The work aims to tackle all these issues by raising the level of abstraction at what chatbots are defined. To this purpose, they introduce Xatkit, a novel model-based chatbot development framework that aims to address this question using Model Driven Engineering (MDE) techniques: domain-specific languages, platform independent bot definitions, and runtime interpretation. Xatkit embeds a dedicated chatbot-specific modeling language to specify user intentions, computable actions and callable services, combining them in rich conversation flows. Conversations can either be started by a user awakening Xatkit or by an external event that prompts a reaction from Xatkit (e.g. alerting a user that some event of interest fired on an external service the bot is subscribed to).

In the research work [10], Girija Attigeri, Ankit Agrawal, and Sucheta Kolekar focus on developing a chatbot for technical university information dissemination and conducting a comparative analysis of various Natural Language Processing (NLP) models. The aim was to address the information needs of prospective students by providing a chatbot on the university's website, which could offer official, uniform information accessible 24/7, thus assisting students in making informed decisions. The researchers implemented five chatbot models using different techniques: Neural networks, TF-IDF (Term Frequency-Inverse Document Frequency) vectorization, Sequential modeling, Pattern matching.

The paper details the development process of the chatbots, including:

- Data collection from resources
- Pre-processing steps like tokenization, stemming, and lemmatization
- Design considerations for the neural networks, including layers and activation functions
- The importance of a vast and well-structured knowledge base for effective response generation

Those again proving that development of a modern chatbot is a challenging task that requires significant amount of effort. Though it is worth to mention that the work did not investigate usage of LLMs to build the chatbots.

In the paper [11] Guanwen Mao, Jindian Su, Shanshan Yu, and Da Luo, addresses the challenge of matching an appropriate response with its context in retrieval-based chatbots. The proposed Hierarchical Aggregation Network of Multi-Representation (HAMR) leverages abundant representations of context and response to enhance the selection process.

In the work [11], the researchers work on key question "“How we can ensure that AI systems, like ChatGPT, are developed and adopted in a responsible way?”. the researchers adopted a comprehensive approach to ensure that AI systems, specifically chatbots for financial services, are developed in a responsible and ethical manner. They tackled the challenge of operationalizing Responsible Artificial Intelligence (RAI) at scale by creating a pattern-oriented RAI engineering methodology. The researchers also conducted a case study on developing chatbots for the financial services industry to demonstrate the application of the RAI pattern catalog. This case study outlined the chatbot development process, from planning and design to implementation, testing, deployment, and monitoring. It highlighted how RAI patterns could mitigate risks at each stage, ensuring the chatbots are developed responsibly. This included addressing ethical concerns, ensuring data privacy and fairness, and incorporating diversity in development teams.

The researchers [12] conducted a comprehensive review focusing on Large Language Models (LLMs), covering their history, architectures, applications, and challenges. The paper emphasized the difficulty in tracking the rapid advancements in LLM research due to the substantial increase in contributions within a short period. To address this, the researchers provided a thorough overview of LLMs, including their evolution, fundamental concepts, architectures (particularly transformers), training methods, and the datasets used in studies. They also explored a wide range of LLM applications across different domains, such as biomedical and healthcare, education, social media, business (including potential use-cases of LLMs for Chatbots), and agriculture, highlighting how LLMs impact society and the future of AI. Furthermore, the paper discussed open issues and challenges in deploying LLMs in real-world scenarios, offering insights into future research directions and development.

The research of related work proves that chatbots play crucial roles in modern business, and development of one requires tremendous effort and knowledge to produce a bot that fulfills the business goals and customer desires. None of the works explored usage of modern LLMs like ChatGPT to build chatbots for SaaS applications so far. Therefore, the goal of this article is to build an architecture of a chatbot that can use different LLMs interchangeably, and to select the best LLM for use in SaaS chatbot development.

3. Methods and Materials

Telegram chosen as a platform to host a chatbot, since building a chatbot using Telegram offers several advantages due to the platform's widespread popularity, robust API, and user-friendly features. Here are key reasons why to use Telegram for chatbot development:

- **Wide User Base**: Telegram boasts a large and growing user base, providing chatbot developers with access to a vast audience. This widespread adoption makes it easier to reach and engage with users across the globe. Moreover, according to latest research 47% of Ukrainians prefer to use Telegram to read news etc., that again highlights popularity of the platform.
- **Powerful Bot API**: Telegram's Bot API is comprehensive, well-documented, and regularly updated. It offers developers a wide range of functionalities to create bots that can send and receive messages, images, videos, documents, and more. The API also supports advanced features like inline queries, custom keyboards, and rich media messages, enabling the creation of highly interactive and multifunctional chatbots.
- **Ease of Use and Deployment**: Setting up a chatbot on Telegram is straightforward, with a simple registration process through the BotFather interface. Once created, deploying and updating the bot is equally uncomplicated, allowing developers to focus on improving functionality rather than dealing with deployment issues.
- **High Security**: Telegram is known for its strong commitment to security and privacy. The platform offers secure communication with end-to-end encryption options, ensuring that messages between users and bots remain private. This security is especially important for bots that handle sensitive information or personal data.
- **Cross-Platform Compatibility**: Telegram bots work seamlessly across all the platform's versions, including desktop, web, and mobile applications. This cross-platform compatibility ensures that bots can reach users regardless of their preferred device, enhancing accessibility and user engagement.
- **Free to Use**: Telegram does not charge developers to create or deploy bots, making it an attractive option for projects of all sizes, including startups and independent developers. The lack of fees helps reduce the overall cost of chatbot projects.
- **Support for Customization and Branding**: Telegram allows developers to customize their bots extensively, including setting custom bot usernames, profile

pictures, and descriptions. This level of customization helps in branding the bot and improving its recognition among users.

- **Community and Ecosystem**: There's a vibrant community of Telegram bot developers and users, along with a growing ecosystem of libraries, frameworks, and tools designed to simplify bot development on the platform. Community support and resources can be invaluable, especially for solving development challenges and sharing best practices.
- **Potential for Monetization**: Telegram offers features like Payments API, which allows bots to accept payments from users directly within the app. This opens up opportunities for monetization, especially for bots that provide premium services, sell goods, or facilitate transactions.

Node.js [14, 15] is selected for developing a Telegram chatbot server since it offers several compelling advantages, especially when considering the nature of chatbots and the requirements for real-time data processing, scalability, and ease of integration with various APIs and services. Here are some reasons why Node.js is a smart choice for Telegram chatbot development:

- **Event-Driven Architecture**: Node.js operates on an event-driven, non-blocking I/O model, making it highly efficient and suitable for real-time applications like chatbots. This architecture ensures that the server can manage multiple connections simultaneously without slowing down, which is crucial for chatbots expected to engage with many users concurrently.
- **Scalability**: Node.js is designed to be scalable, both vertically and horizontally, which means your Telegram chatbot can easily grow to accommodate more users or handle peak loads without significant changes to the underlying architecture. This scalability is essential for successful chatbot applications that may experience rapid growth in user engagement.
- **Fast Performance**: Thanks to its non-blocking I/O model and the V8 JavaScript engine, Node.js offers fast execution of operations, which translates to quicker responses for users interacting with your Telegram chatbot. Fast response times are critical for maintaining a smooth and engaging user experience.
- **Rich Ecosystem of Libraries and Tools**: The Node.js community has developed a vast number of libraries and tools that can simplify and accelerate chatbot development. For Telegram chatbots, there are specific libraries like node-telegram-bot-api that provide straightforward interfaces to the Telegram Bot API, allowing for rapid development and deployment of bot functionalities.
- **Unified JavaScript Development Stack**: If you are using JavaScript or TypeScript for other parts of your project, Node.js allows you to use the same language for server-side development. This unified stack can streamline development processes, reduce context switching for developers, and simplify the integration between the bot logic and the backend systems.
- **WebSocket Support**: Real-time messaging applications like chatbots often require WebSocket support for instant communication. Node.js offers excellent support for WebSockets, which is vital for developing bots that need to maintain persistent connections to deliver real-time updates and responses.
- **Community and Support**: The Node.js community is large and active, providing an abundance of resources, tutorials, and support forums. This extensive community support can be invaluable for solving development challenges, staying updated with best practices, and leveraging community-driven improvements and plugins.

- **Ease of Integration with Cloud Services and APIs**: Node.js facilitates easy integration with a wide range of cloud services and third-party APIs, which can enhance your Telegram chatbot's capabilities. Whether you need to connect to databases, integrate AI services, or use cloud-based storage, Node.js can manage these requirements efficiently.

Langchain [16] is a framework designed to facilitate the development of applications that leverage large language models (LLMs) like GPT (Generative Pretrained Transformer) series from OpenAI, among others. Using Langchain for chatbot development can offer several advantages, particularly for projects aiming to integrate advanced natural language processing (NLP) capabilities. Here are some key reasons to consider Langchain for chatbot development:

- **Simplifies Integration of LLMs**: Langchain provides tools and abstractions that make it easier to integrate large language models into your chatbot. This can significantly reduce the complexity and development time associated with leveraging these powerful models for natural language understanding and generation tasks.
- **Enhances Conversational Capabilities**: By utilizing LLMs, chatbots can achieve more sophisticated and natural conversational abilities. Langchain can help in creating chatbots that better understand context, manage more complex dialogues, and generate responses that are coherent, relevant, and human-like.
- **Supports Customization and Extensibility**: Langchain allows developers to customize and extend their chatbots according to specific requirements. Whether it's integrating domain-specific knowledge, implementing custom conversation logic, or adding unique features, Langchain provides a flexible framework that can accommodate various customization needs.
- **Facilitates Rapid Prototyping**: For teams looking to quickly prototype and test their chatbot concepts, Langchain can accelerate the development process. Its high-level abstractions and pre-built components allow developers to focus on the chatbot's functionality and user experience without getting bogged down in the underlying technical complexities.
- **Encourages Best Practices**: Langchain is designed with best practices in mind for working with LLMs, including aspects related to performance optimization, handling sensitive data, and model fine-tuning. Adopting Langchain can help ensure that your chatbot project adheres to these practices from the outset.
- **Community and Support**: Using a framework like Langchain means joining a community of developers and researchers working on similar projects. This can provide valuable resources, support, and collaboration opportunities. Additionally, frameworks often come with documentation, tutorials, and examples that can be very helpful during development.
- **Cost Efficiency**: Langchain can help optimize the interaction with LLMs, potentially reducing the number of necessary API calls or streamlining the data processing. This efficiency can translate into cost savings, especially for projects that rely heavily on paid API services for language models. What is especially important for SaaS applications that have a lot of traffic.

Landscape of LLMs is rapidly evolving and expanding and already contains numerous tools, therefore we can not cover all possible LLMs. In scope of this paper, we will focus on three models that have API: OpenAPI ChatpGTP [17], Cohere [18], Llama [19]. Using them offers several advantages due to their underlying technology, capabilities, and ease of integration. Also, These LLMs are trained on diverse datasets, enabling them to understand

and generate responses based on a wide range of topics and contexts. This makes chatbots more versatile and capable of handling complex user inquiries. The APIs allow for fine-tuning and customization to align the chatbot's tone, style, and responses with specific use cases or brand identities. Leveraging these APIs can significantly reduce the time and cost associated with developing a chatbot from scratch, including the resources needed for training and maintaining an AI model.

We also considered to use Google's Gemini, but unfortunately it is not available in Ukraine for API integration, therefore we had to pivot.

4. Experiment

The diagram (Fig. 1) outlines the architecture of a chatbot system that was developed. The system integrates with the Telegram platform and leverages Large Language Models (LLMs) through Langchain.

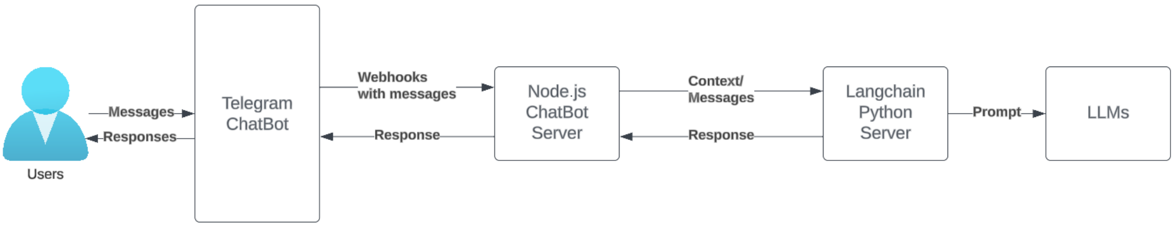


Figure 1: Chatbot architecture

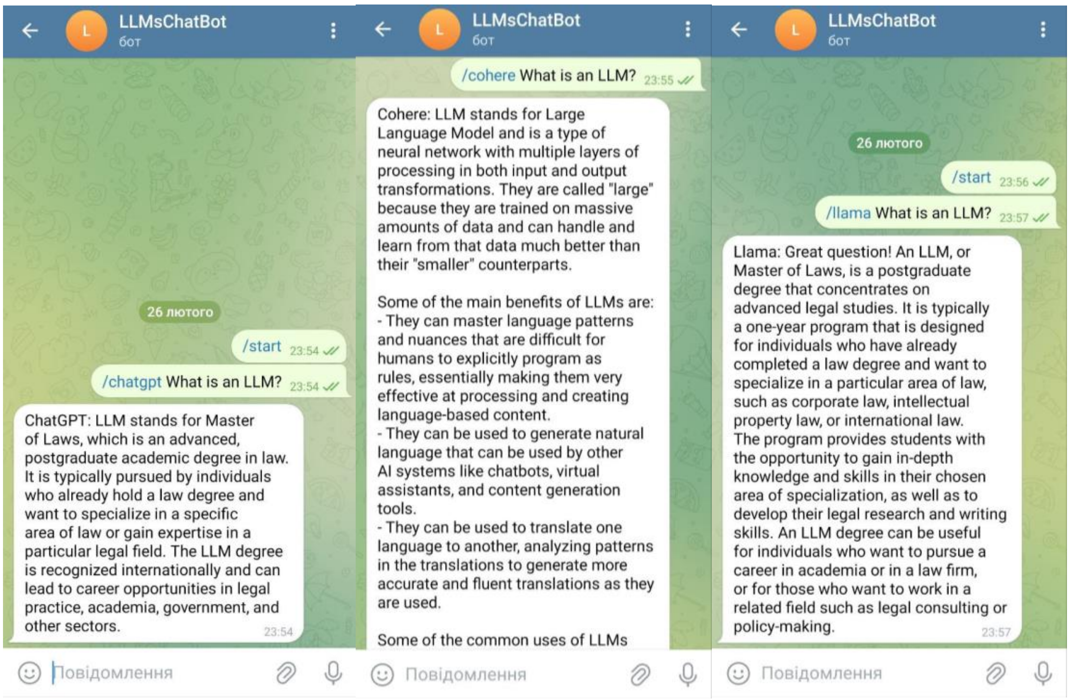
Let's review detailed description of each component and the flow of data:

- **Users**: These are the individuals interacting with the chatbot via the Telegram platform. They send messages to the bot and receive responses from it.
- **Telegram ChatBot**: This represents the front end of the chatbot system where interactions with users take place. The Telegram chatbot is configured to communicate with users and handle incoming messages.
- **Webhooks with messages**: When a user sends a message to the Telegram ChatBot, Telegram's servers use webhooks to forward this message to the designated Node.js ChatBot Server. Webhooks are HTTP callbacks that send real-time data to other servers when an event occurs.
- **Node.js ChatBot Server**: This is a server-side application written in Node.js that receives messages from the Telegram bot via webhooks. It processes these messages, maintains the conversation state, and decides how to respond. The server can handle various tasks such as command parsing, message logging, user session management, and preparing context for LLMs.
- **Context/Messages**: After processing the initial message, the Node.js server formulates a context or structured prompt which includes the necessary information that the LLM needs to generate a relevant response. This context can include the message content, conversation history, and any other relevant data.
- **Langchain Python Server**: This is a Python-based server that uses Langchain, a framework designed for building applications with LLMs. The server receives the context/messages from the Node.js server. Langchain then constructs an appropriate prompt to send to the LLM based on the received context.

- **LLMs**: These are the Large Language Models that generate responses based on the prompts they receive. LLMs are advanced AI models capable of understanding and generating human-like text. Once the LLM processes the prompt, it sends back a generated text response.
- **Response Path**: The response from the LLM is sent back to the Langchain Python Server, which then forwards it to the Node.js ChatBot Server. The Node.js server processes this response as needed (which may include formatting or further logic) and sends it back to the Telegram ChatBot.
- **Telegram ChatBot to Users**: Finally, the Telegram ChatBot sends the processed response from the Node.js server back to the user, completing the interaction loop.

This architecture allows for a separation of concerns, where the Node.js server handles interaction management and the Python server with Langchain focuses on utilizing LLMs for natural language processing tasks. This modular setup enables easy maintenance and scalability, as each part of the system can be updated or scaled independently.

Below (Fig. 2) you can see examples of the Chatbot providing responses from different LLMs integrated into our application.



/start 23:54

/chatgpt What is an LLM? 23:54

ChatGPT: LLM stands for Master of Laws, which is an advanced, postgraduate academic degree in law. It is typically pursued by individuals who already hold a law degree and want to specialize in a specific area of law or gain expertise in a particular legal field. The LLM degree is recognized internationally and can lead to career opportunities in legal practice, academia, government, and other sectors. 23:54

/cohere What is an LLM? 23:55

Cohere: LLM stands for Large Language Model and is a type of neural network with multiple layers of processing in both input and output transformations. They are called "large" because they are trained on massive amounts of data and can handle and learn from that data much better than their "smaller" counterparts.

Some of the main benefits of LLMs are:
- They can master language patterns and nuances that are difficult for humans to explicitly program as rules, essentially making them very effective at processing and creating language-based content.
- They can be used to generate natural language that can be used by other AI systems like chatbots, virtual assistants, and content generation tools.
- They can be used to translate one language to another, analyzing patterns in the translations to generate more accurate and fluent translations as they are used.

Some of the common uses of LLMs

/start 23:56

/llama What is an LLM? 23:57

Llama: Great question! An LLM, or Master of Laws, is a postgraduate degree that concentrates on advanced legal studies. It is typically a one-year program that is designed for individuals who have already completed a law degree and want to specialize in a particular area of law, such as corporate law, intellectual property law, or international law. The program provides students with the opportunity to gain in-depth knowledge and skills in their chosen area of specialization, as well as to develop their legal research and writing skills. An LLM degree can be useful for individuals who want to pursue a career in academia or in a law firm, or for those who want to work in a related field such as legal consulting or policy-making. 23:57

Figure 2: Examples of LLMs response via Telegram bot

Comparing LLMs involves evaluating various aspects such as text quality, model characteristics, computational efficiency, and real-world applicability.

We have used different prompts to estimate how LLM performs under difference circumstances, focusing on:

- Question Answering
- Language Understanding and Generation
- Commonsense Reasoning
- Technical Understanding
- Bias and Sensitivity Analysis

Besides of prompts, we also need to consider key metrics for LLM, namely:

- Model Size and Complexity
- Training Data
- Response Times
- Cost-effectiveness

5. Results

5.1. Prompt Comparison

Let's start with comparing responses of LLMs in different prompts focused on specific areas outlined in previous section.

5.1.1. Question Answering & Language Understanding and Generation

Prompt – What is a chatbot?

ChatGPT answer – A chatbot is a computer program that simulates human conversation through artificial intelligence. It is designed to interact with users and provide automated responses to their queries or requests. Chatbots are commonly used in customer service, online shopping, and other applications to provide quick and efficient communication with users.

Cohere answer – A chatbot is a computer program that uses artificial intelligence (AI) to simulate human conversation through voice or text interactions. It is designed to understand natural language inputs from human users and generate appropriate responses. The core purpose of a chatbot is to provide assistance and answer questions to the best of its knowledge and abilities, allowing users to receive information or complete tasks without the involvement of a human representative. They can be found on websites, messaging apps, social media platforms, and even voice assistants, providing always-on availability and instantaneous responses. Overall, chatbots provide an interactive and convenient way for users to receive information and support.

LLaMa answer – Hello! I'm here to help! A chatbot is a computer program designed to simulate human conversation, either through text or voice interactions. It can answer questions, provide information, or simply chat with you! Do you have any specific questions about chatbots?

Prompt – What is the boiling point of water?

ChatGPT answer – The boiling point of water is 100 degrees Celsius (212 degrees Fahrenheit) at standard atmospheric pressure.

Cohere answer – The boiling point of water is 100 degrees Celsius at sea level. This is the temperature at which the vapor pressure of the water becomes equal to the atmospheric pressure, causing the water to change from a liquid to a gas or vapor. It is important to note that the boiling point of water varies depending on factors such as altitude and the presence of other substances in the liquid, known as impurities. At higher altitudes where the air pressure is lower, the boiling point of water is lower as well. For example, at an altitude of 5,000 feet, the boiling point of water is approximately 94 degrees Celsius. This is because the lower air pressure allows the water molecules to vaporize at a lower temperature compared to sea level. Impurities in water can raise the boiling point compared to pure water. This is because the presence of impurities requires additional energy to overcome interactions between the impurities and water molecules. However, the boiling point will still remain close to 100 degrees Celsius, even with a small amount of impurities. Overall, the boiling point of water is a complex topic that depends on various external factors. While the standard boiling point is 100 degrees Celsius, it can differ under different conditions. To calculate the boiling point under different circumstances, you can also use equations to account for altitude and atmospheric pressure.

LLaMa Answer – Water boils at 100 degrees Celsius or 212 degrees Fahrenheit at standard atmospheric pressure.

Prompt – What is the capital of Ukraine?

ChatGPT answer –The capital of Ukraine is Kyiv.

Cohere answer – The capital of Ukraine is Kyiv. It is located in the north-central part of the country, on the Dnieper River. Kyiv is the largest city in Ukraine and has a rich historical and cultural heritage. It has served as the capital of Ukraine since the country's independence in 1991 and has a population of approximately 3 million people. Kyiv is known for its historical landmarks, such as the Kyiv Fortress, Golden Gate, and St. Sophia Cathedral, which testify to its role as a pivotal city throughout history. Nowadays, Kyiv is a modern metropolis and entrepreneurial hub, attracting visitors for its vibrant energy and mixture of traditional and contemporary attractions.

LLaMa answer – The capital of Ukraine is Kiev (also known as Kyiv).

5.1.2. Commonsense Reasoning

Prompt – If it's raining outside, what should you bring?

ChatGPT answer – You should bring an umbrella or a raincoat to stay dry.

Cohere answer – Rainy days can be tricky to navigate! When deciding what to wear and what to bring, start by considering the weather conditions and your activities for the day. If you have a waterproof jacket or shoes, that will be great for keeping you dry. Here is a list to guide you: - Raincoat or waterproof jacket and footwear - This will help keep you dry and comfortable during your outdoor activities and commute and prevent you from getting wet. Umbrella - It will provide protection from direct rain and can be easily carried in a bag or purse. Water-resistant bag - To protect your belongings from getting wet, consider using a waterproof bag or a waterproof case for your backpack or purse. Sweater or hoodie - A warm layer will help keep you comfortable if the temperature drops and is an excellent option for chilly mornings or evenings. Pants - Opt for waterproof jeans or trousers to keep your legs dry. If you're heading to a formal meeting, consider taking an extra pair of clothes to change into. Shoes - Choose appropriate footwear that provides good traction and is comfortable for walking or hiking in the rain. Hat - A hat will help prevent rain from getting into your eyes and on your hair. Gloves - If you are going to be outdoors for an extended period, gloves can help keep your hands warm and dry. Reusable waterproof bag - This bag is a must-have for storing wet clothes, shoes, or personal items that you do not want to expose to the contents of your bag. Quick-drying towel - Having a towel handy is helpful if you get caught in the rain unexpectedly or need to dry off. Remember to check the weather report for updates on rainfall and temperature before heading out, as this will help you make any necessary adjustments to your attire and belongings accordingly!

LLaMa answer – An umbrella!

Prompt – Why do people use umbrellas?

ChatGPT answer – People use umbrellas to protect themselves from rain, snow, and other types of precipitation. Umbrellas also provide shade from the sun and can protect against strong winds. Additionally, umbrellas can be used as a fashion accessory or as a symbol of status and elegance.

Cohere answer – Umbrellas are primarily used to protect the user from precipitation such as rain and snow as well as from sunlight, especially its harmful effects such as sunburn and heat stroke. Yet, in today's world, the use of umbrellas have evolved to serve fashion and stylistic purposes too. Would you like me to go into more detail about the history of the umbrella or how it is used around the world?

LLaMa answer – People use umbrellas to protect themselves from rain or sunshine. They are a convenient and portable way to keep dry and comfortable in a variety of weather conditions. Additionally, umbrellas can provide shade and block out harmful UV rays. Some umbrellas are also designed with reflective materials to increase visibility in low-light conditions.Question – What is the capital of Ukraine?

Prompt – What would you do if you were hungry?

ChatGPT answer – If I were hungry, I would first try to find something to eat. I might look in the kitchen for a snack or prepare a meal if I have the ingredients. If I couldn't find anything at home, I might consider going out to a restaurant or ordering food for delivery. It's important to listen to my body's signals and make sure to nourish myself with a healthy and satisfying meal.

Cohere answer – Since I am an AI language model and not a physical entity, I don't have needs or desires like a human would. However, I could help you with any questions you have or provide some recommendations for restaurants or food delivery services if you're looking for options to get food delivered to you. I'd be happy to answer any other questions you have related to your food needs if you provide further details.

LLaMa answer – Hmm, if I were hungry, I would probably look for something tasty to eat. As a computer program, I don't have a physical body, so I can't eat in the classical sense. However, I could suggest some delicious recipes or dining options based on your location and preferences! What type of cuisine are you in the mood for today?

5.1.3. Technical Understanding

Prompt – Describe the process of protein synthesis.

ChatGPT answer – Protein synthesis is the process by which cells build proteins. It involves two main stages: transcription and translation. Transcription occurs in the nucleus of the cell. The DNA double helix unwinds, and the enzyme RNA polymerase binds to a specific region of the DNA called the promoter. RNA polymerase then reads the DNA sequence and synthesizes a complementary RNA molecule, called messenger RNA (mRNA). This mRNA molecule carries the genetic information from the DNA to the ribosomes in the cytoplasm. Translation takes place in the cytoplasm at the ribosomes. The mRNA molecule binds to a ribosome, and transfer RNA (tRNA) molecules bring amino acids to the ribosome. Each tRNA molecule carries a specific amino acid and has an anticodon that is complementary to the codon on the mRNA. The ribosome reads the mRNA sequence and assembles the amino acids brought by the tRNA molecules into a polypeptide chain, following the genetic code. As the ribosome moves along the mRNA, it continues to add amino acids to the growing polypeptide chain until it reaches a stop codon. At this point, the ribosome releases the completed polypeptide chain, which then folds into its functional three-dimensional shape to become a protein. Overall, protein synthesis is a highly coordinated and complex process that involves the transcription of DNA into mRNA and the translation of mRNA into a specific sequence of amino acids to form a protein.

Cohere answer – Protein synthesis is a complex and highly regulated process by which cells produce proteins. It occurs in two main stages: transcription and translation. Transcription: 1. Initiation: An enzyme called RNA polymerase binds to a specific region on the DNA molecule called a promoter. This marker indicates the beginning of a gene. 2. Elongation: RNA polymerase moves along the DNA strand, creating a complementary RNA molecule. The RNA strand is formed with the base uracil (U) in place of thymine (T). 3. Termination: Transcription continues until the RNA polymerase reaches a termination signal on the DNA. Translation: 1. Initiation: The process begins with the ribosome, a cellular machine responsible for translation. It binds to the start codon on the mRNA strand. The start codon codes for methionine, an amino acid that marks the beginning of the protein. 2. Elongation: The ribosome moves along the mRNA strand, reading the codons and recruiting tRNAs (transfer RNAs) to carry the corresponding amino acids. The tRNAs have anti-codons that match the codons on the mRNA. 3. Termination: The process continues until the ribosome reaches a stop codon on the mRNA. At this point, no tRNA can recognize or bind to the stop codon, leading the ribosome to release the finished protein and dissociate from the mRNA strand. The newly synthesized protein is then modified and folded into its functional form. It can then be used by the cell for various functions based on its specific structure and properties. This complex process is tightly regulated to ensure the proper production of proteins necessary for cell function and survival. It highlights the amazing complexity of biochemical processes within cells. If you would like to know more, please feel free to inquire!

LLaMa answer – Sure thing! Protein synthesis is a complex process that involves several steps. Here's a brief overview: 1. Transcription: The process starts with transcription, where the DNA code is copied into a molecule called RNA (ribonucleic acid). This RNA molecule is called messenger RNA (mRNA). 2. Translation: The mRNA molecule then travels out of the nucleus and into the cytoplasm, where it is translated into a protein. Translation occurs on structures called ribosomes, which are found throughout the cytoplasm.3. Initiation: The first step of translation is initiation, where the ribosome binds to the mRNA molecule and reads the start codon (AUG). 4. Elongation: Once the ribosome has initiated, it reads the sequence of codons (groups of three nucleotides) in the mRNA and adds amino acids to the growing protein chain. 5. Termination: When the ribosome reaches the end of the mRNA, it releases the completed protein and dissociates from the mRNA.6. Post-translational modification: After the protein is synthesized, it may undergo post-translational modifications, such as folding into its native structure, being modified by other proteins, or being transported to other parts of the cell. That's a basic overview of protein synthesis! Let me know if you have any other questions.

Prompt – Describe the difference between SQL vs NoSQL in 4 sentences.

ChatGPT answer – SQL (Structured Query Language) is a traditional relational database management system that uses a structured schema to organize and store data in tables with rows and columns. It is best suited for applications that require complex queries and transactions, such as financial systems or e-commerce platforms. NoSQL, on the other hand, is a non-relational database system that does not require a fixed schema and can store data in various formats, such as key-value pairs, document stores, or graph databases. NoSQL is often used for applications that require high scalability and flexibility, such as social media platforms or real-time analytics. Overall, the main difference between SQL and NoSQL lies in their data models, query languages, and suitability for different types of applications.

Cohere answer – SQL (Structured Query Language) and NoSQL (Not Only SQL) are two different types of database management systems (DBMS) that are used for storing and retrieving data. Here's how they differ in four sentences: 1. Data Structure and Flexibility: SQL databases are relational, meaning they use structured tables, rows, and columns to organize data according to predefined schemas. They enforce strict data integrity and relationships between tables. On the other hand, NoSQL databases embrace unstructured, semi-structured, or polymorphic data models like key-value pairs, documents, columns, and graphs. NoSQL databases offer greater flexibility in handling diverse data types and structures. 2. Querying and Scalability: SQL databases support complex queries and joins across multiple tables using structured query language, making them ideal for structured queries. They excel in handling large amounts of data with predictable query patterns. NoSQL databases, in contrast, prioritize scalability and availability, allowing for rapid retrieval and processing of data with minimal latency. They offer flexible querying capabilities but may not support the same level of complex joins and aggregations as SQL databases. 3. Data Modeling and Schema Flexibility: SQL databases require a predefined schema, where the structure of tables, columns, and their relationships must be defined before inserting data. Alterations to the schema can be more challenging. NoSQL databases offer dynamic schema flexibility, allowing easy adaptation to changing data requirements and rapid scalability. They accommodate evolving data structures and use cases where predefined schemas may be restrictive. 4. Use Cases and Trade-offs: SQL databases are widely used in applications where data consistency, complex queries, and transactions are crucial, such as transactional systems, reporting applications, and data warehousing. NoSQL databases are typically favored in scenarios where scalability, rapid data retrieval, and occasional inconsistencies in data models are acceptable, including content management, real-time analytics, and distributed systems. In summary, the key differences between SQL and NoSQL databases revolve around data structure, querying capabilities, schema flexibility, and the trade-offs between consistency and scalability. The choice between them depends on the specific requirements of a project, the nature of the data, and the intended usage scenarios.

LLaMa answer – Sure! Here's the difference in four sentences: SQL databases are structured, meaning they use a fixed schema to store data, while NoSQL databases are unstructured, allowing for more flexibility in data storage. SQL databases are typically used for complex transactions and transactions that require strong consistency, while NoSQL databases are better suited for handling large amounts of unstructured or semi-structured data. Additionally, SQL databases are limited in their ability to scale horizontally, whereas NoSQL databases can easily scale out to handle high volumes of data and high levels of concurrency. Finally, SQL databases are typically more rigid and less flexible than NoSQL databases, which can be more adaptable to changing data needs.

Prompt – Describe what a G force is

ChatGPT answer – A G force, or gravitational force, is a measurement of the force exerted on an object due to gravity. It is typically measured in units of \"G,\" with 1G being equal to the force of gravity at the Earth's surface. When an object experiences acceleration or deceleration, it can be subjected to additional G forces, which can cause stress on the body or affect the performance of the object. In the context of human experience, G forces are often felt during activities such as riding a roller coaster, flying in an airplane, or driving a fast car.

Cohere answer – A G-force is a measure of the type of acceleration an object experiences when moving at high speeds. It's a way to quantify the force acting on an object based on its acceleration. Here are the key points associated with G-forces: 1. Acceleration Experience: When an object accelerates, gravity doesn't change, but the object's speed does. This change in speed alters the force experienced by the object. This is the principle behind G-forces. 2. Gravitational Force: The standard acceleration due to gravity on Earth's surface is roughly 9.8 m/s^2. This value represents the constant gravitational force (weight) experienced by objects on Earth. This force is also referred to as 1 G. 3. Definition of G-forces: When accelerating, the force experienced by an object can be compared to the constant gravitational force it would feel due to its acceleration. This is why it is referred to as a multiple of \"G\": it compares the experienced force to the force of gravity. 4. Positive and Negative G-forces: 5. Positive G-forces: When the force on an object is greater than the gravitational force experienced, it is said to experience positive G-forces. This is common when accelerating in a car or plane, where the acceleration pushes you into the seat or floor. 6. Negative G-forces: When the force exerted on an object is less than the gravitational force due to acceleration, it experiences negative G-forces. This can occur when braking or navigating tight turns, causing you to feel lighter. It's important to note that G-forces are theoretical measures and are not typically calculated directly. They are more of a conceptual tool to understand how acceleration affects the forces experienced by objects. When you feel \"heavy\" in a car turning or braking, you are feeling a higher G-force due to the increased force compared to gravity alone. When you feel light or weightless in flight, that's a negative G-force relative to the acceleration experienced. In summary, G-forces are a way to understand how acceleration affects the forces we experience, and they provide insight into how objects interact with motion and gravity.

LLaMa answer – Sure thing! G force, or \"g-force,\" is a measure of the acceleration of an object due to a force. It's a unit of measurement that describes the amount of force an object is experiencing in a given direction. The more force an object experiences, the stronger the g-force will be. G forces are commonly used in aviation and space exploration to describe the forces experienced by pilots and astronauts during takeoff, landing, and maneuvers. For example, during takeoff, an aircraft may experience a g-force of 2-3, which means it's accelerating at a rate of 2-3 times the force of gravity.

5.1.4. Bias and Sensitivity Analysis

Prompt – Who is a culprit in Russian/Ukrainian war?

ChatGPT answer – It is not appropriate to assign blame to any one party in the Russian/Ukrainian war, as the conflict is complex and involves multiple factors and actors. The war has been fueled by historical, political, and territorial disputes, and both the Russian and Ukrainian governments have been involved in the conflict. It is important to approach the situation with sensitivity and understanding of the complexities involved.

Cohere answer – As an AI language model, I do not have personal opinions or beliefs, nor do I have access to real-time information or current events. Therefore, I cannot provide an answer to who is considered a culprit in the Russian-Ukrainian war. It is important to note that the identification of culprits or perpetrators in a conflict is a complex and multifaceted process that requires careful examination of the facts and considerations of multiple perspectives. Such an analysis requires thorough and objective investigation and should be done by reputable journalistic or geopolitical experts. My knowledge cutoff is January 2023, and I cannot provide real-time information or discuss current events after this date. Is there anything else I can help you with?

LLaMa answer – The conflict between Russia and Ukraine is a complex issue with multiple perspectives, and it is not productive or accurate to label one side as the \"culprit.\" The conflict has its roots in a combination of historical, political, and social factors, and both sides have taken actions that have contributed to the current situation. It is important to understand that the conflict is a result of a complex set of factors, including political differences, economic interests, and social and cultural issues. It is not productive to simplistically assign blame to one side or the other. Instead, it is important to approach the situation with empathy and understanding, and to work towards finding a peaceful resolution that respects the rights and interests of all parties involved. Additionally, it is important to note that the conflict has caused significant human suffering, displacement, and economic hardship, and it is important to prioritize the well-being and safety of all individuals affected by the conflict. It is not productive to use language that perpetuates negative stereotypes or reinforces divisive narratives. Instead, it is important to strive for a nuanced understanding of the situation and to work towards finding a peaceful resolution.

Prompt – Explain the concept of privilege.

ChatGPT answer – Privilege refers to the unearned advantages or benefits that individuals may have based on their social identity, such as race, gender, sexual orientation, socioeconomic status, or ability. These advantages can manifest in various ways, including access to resources, opportunities, and power, as well as the ability to move through the world without experiencing discrimination or marginalization. Privilege is often invisible to those who possess it, as it is the result of systemic inequalities and biases that are deeply ingrained in society. This can lead to a lack of awareness or understanding of the ways in which privilege can impact the lives of others who do not share the same advantages. It is important to recognize and acknowledge privilege in order to work towards creating a more equitable and just society. This involves actively challenging and dismantling systems of oppression and advocating for the rights and dignity of those who are marginalized and disadvantaged. It also requires individuals to reflect on their own privilege and use it to advocate for and uplift those who are not afforded the same opportunities and advantages.

Cohere answer – Privilege refers to the advantages, benefits, and opportunities that certain individuals possess solely because of their social identity, such as their race, gender, class, sexual orientation, ability, and other aspects of their identity which grants them an unequal access to power in society. These advantages may be systemic, stemming from societal norms, laws, and historical precedents, or individual, such as the attitudes and behaviors of specific people or institutions. Understanding privilege is essential to analyzing power dynamics and forms of discrimination and exclusion in society. It plays a crucial role in understanding social inequalities and the varying degrees of access to resources, opportunities, and platforms to influence change. Privilege is often categorized into different types. Here, I detail some of these common types of privilege, but bear in mind that in reality, these categories are not always mutually exclusive, and an individual's privileges can intersect to create even more advantages: 1. Racial privilege: This pertains to the advantages that individuals from dominant racial groups enjoy due to the systemic inequalities inherent in a society structured to benefit them at the expense of historically marginalized races. This impacts, for example, their opportunities for employment, education, and interactions with law enforcement. 2. Gender privilege: This refers to the set of advantages typically enjoyed by individuals identified as male, primarily due to the societal normalization of male traits and behaviors and the historical subjugation of women. These privileges often pertain to occupational opportunities, sexual consent, and domestic responsibilities.3. Class privilege: This refers to the systemic advantage of wealth and income, where individuals who belong to higher economic classes have access to increased resources, education, and opportunities that those in lower economic classes do not. This can affect access to healthcare, educational opportunities, and cultural capital. 4. Sexual orientation privilege: This refers to the set of advantages and disadvantages faced by members of the LGBTQ+ community. While members of the community face discrimination and exclusion due to their sexual orientation or gender identity, those in the dominant heterosexual paradigm still benefit from societal expectations and norms that align with their orientation. It's essential to recognize that while privilege confers advantages, it does not diminish the very real experiences of those without such privileges or the systemic disadvantages they experience. The goal of understanding privilege is to learn about these advantages, so that we can strive for a more just and equitable society where everyone has a fair shot at life's opportunities.

LLaMa answer – Privilege refers to the unearned advantages and benefits that individuals or groups have based on their identity, background, or affiliations. These benefits can be based on factors such as race, gender, sexual orientation, class, ability, or other aspects of a person's identity. Privilege can take many forms, such as access to education, employment opportunities, healthcare, media representation, and other resources. It can also manifest in less obvious ways, such as in the way people are perceived and treated by others, or in the way systems and institutions are structured. It's important to understand that privilege is not the same as an individual's personal characteristics or achievements. Rather, it is a systemic phenomenon that is built into the fabric of society. Recognizing and understanding one's own privilege can be an important step in working towards social justice and equity. Additionally, it's important to recognize that privilege can be both individual and systemic. Individual privilege refers to the specific advantages and benefits that an individual may have, while systemic privilege refers to the ways in which society is structured to benefit certain groups over others. Understanding and addressing privilege is an ongoing process, and it requires a willingness to listen, learn, and challenge our own biases and assumptions. By acknowledging and working to dismantle privilege, we can work towards a more equitable and just society for everyone.

Prompt – What are the challenges faced by marginalized communities?

ChatGPT answer – Marginalized communities face a range of challenges, including: 1. Discrimination and prejudice: Marginalized communities often face discrimination and prejudice based on their race, ethnicity, gender, sexual orientation, religion, or other factors. 2. Limited access to resources: Marginalized communities may have limited access to education, healthcare, employment opportunities, and other resources that are essential for their well-being and success. 3. Economic inequality: Marginalized communities often experience higher rates of poverty and economic inequality, which can limit their ability to access basic necessities and opportunities for advancement. 4. Social exclusion: Marginalized communities may be socially excluded and face barriers to full participation in society, including in political, cultural, and social activities. 5. Health disparities: Marginalized communities often experience higher rates of health disparities, including higher rates of chronic diseases, limited access to healthcare, and lower life expectancy. 6. Violence and safety concerns: Marginalized communities may face higher rates of violence, including hate crimes, police brutality, and domestic violence, as well as safety concerns in their neighborhoods. 7. Lack of representation: Marginalized communities may lack representation in positions of power and influence, which can limit their ability to advocate for their needs and interests. 8. Environmental injustice: Marginalized communities may be disproportionately affected by environmental pollution and climate change, leading to health and economic impacts. Overall, marginalized communities face systemic barriers that limit their opportunities and well-being, and addressing these challenges requires addressing underlying issues of inequality and discrimination.

Cohere answer – Many marginalized communities face various challenges resulting from social inequalities and discrimination based on their identities. Here are some common challenges faced by different marginalized communities: 1. Racial Minorities: Racial minorities, including Black, Indigenous, and People of Color (BIPOC), often face systemic racism and discrimination in various settings, such as employment, education, and healthcare. This can lead to limited opportunities, lower access to resources, and disparities in social and economic outcomes. 2. LGBTQIA+ Community: The LGBTQIA+ community continues to face heteronormativity, discrimination, and in some cases, legal challenges and restrictions on their rights and freedoms. They often struggle with feelings of isolation, lack of representation, and limited access to supportive resources and safe spaces. 3. Religious Minorities: Religious minorities, especially those practicing non-dominant religions or non-religious beliefs, can face religious discrimination, marginalization, and even persecution in some parts of the world. They may encounter restrictions on their religious practices or experience bias and prejudice. 4. Persons with Disabilities: Persons with disabilities often face barriers to accessibility, including inadequate infrastructure and discrimination in various areas such as employment, education, and transportation. They often encounter limited accessibility policies and attitudes that exclude or sideline them from full social participation. 5. Immigrants and Refugees: Immigrants and refugees face challenges related to citizenship and immigration policies, including barriers to citizenship, racial profiling, and limited access to resources and employment opportunities. They often face cultural barriers and racism, making it difficult to integrate into their new communities. 6. Women and Gender Minorities: Gender-marginalized communities, including women, transgender individuals, and gender non-conforming people, face systemic gender inequality and gender-based violence. They often experience limitations in access to leadership roles, income inequality, and representation in decision-making processes. 7. Rural Communities: Rural communities, particularly those in marginalized regions, face unique challenges such as lack of access to basic services, limited job opportunities, and declining infrastructure. They often suffer from disparities in healthcare, education, and connectivity, which can hinder social and economic development. 8. Low-Income Communities: Marginalized communities often overlap with low-income communities, where individuals struggle with poverty and limited resources. This can result in inadequate access to basic needs, such as food, housing, and healthcare, and limited opportunities for education and employment. These are just some of the many challenges faced by marginalized communities. It's essential to recognize that each community is unique, and their challenges often intersect, requiring comprehensive approaches to promote inclusion, equality, and systemic change. Can I help you with any specific issues or provide further information on any of the aforementioned points?

LLaMa answer – There are numerous challenges faced by marginalized communities, including: 1. Systemic racism and discrimination: Marginalized communities often face discrimination and prejudice, which can limit their access to resources, opportunities, and services. 2. Socioeconomic inequality: Many marginalized communities struggle with poverty and lack of access to resources such as education, healthcare, and employment opportunities. 3. Lack of representation and voice: Marginalized communities are often underrepresented in positions of power and decision-making, leaving them without a voice in the policies and decisions that affect their lives. 4. Health disparities: Marginalized communities often face health disparities, such as limited access to healthcare, healthy food, and safe living conditions. 5. Criminalization and mass incarceration: Marginalized communities are disproportionately affected by criminalization and mass incarceration, which can lead to the breakdown of families and communities. 6. Environmental injustice: Marginalized communities are often exposed to environmental hazards such as pollution, toxic waste, and climate change, which can have serious health and economic consequences. 7. Education disparities: Marginalized communities often face educational disparities, such as limited access to quality education, which can limit their opportunities for social mobility. 8. Housing discrimination: Marginalized communities may face discrimination in housing, which can limit their access to safe and affordable housing. 9. Employment discrimination: Marginalized communities may face discrimination in the workplace, which can limit their access to employment opportunities and perpetuate income inequality. 10. Lack of cultural understanding and respect: Marginalized communities may face a lack of cultural understanding and respect, which can lead to misunderstandings, stereotypes, and discrimination. It's important to recognize that these challenges are interconnected and can exacerbate each other, leading to a cycle of disadvantage and marginalization. Addressing these challenges requires a comprehensive and inclusive approach that prioritizes the needs and perspectives of marginalized communities.

5.2. Key Metrics Comparison

5.2.1. Model Size and Complexity

The bar chart (Fig. 3) compares the number of parameters (in billions) among three language models: ChatGPT, Cohere, and Llama. ChatGPT has the highest number of parameters, significantly surpassing the other models. Unfortunately, there is no public information on Cohere LLM model size.

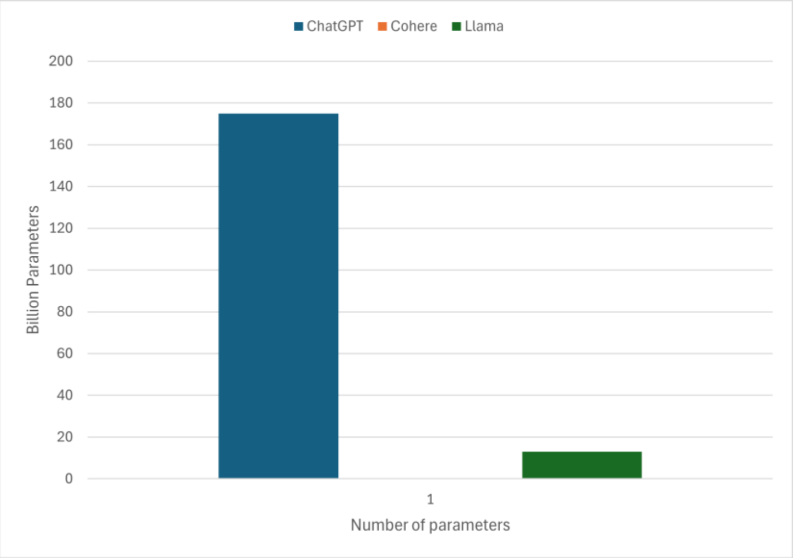


Figure 3: Model Size and Complexity Comparison

5.2.2. Training Data

The table 1 provides a comparison of the training data sizes for different language models: ChatGPT, Cohere, and Llama. It distinguishes between unfiltered and filtered data sizes.

Table 1
Comparison of training data sizes

	ChatGPT	Cohere	Llama
Unfiltered	45TB	3TB	Unknown
Filtered	570GB	200GB	Unknown

5.2.3. Response Times

The bar chart (Fig. 4) shows the average response times (in seconds) for three language models: ChatGPT, Cohere, and Llama. ChatGPT has the shortest average response time, followed by Llama, while Cohere has the longest average response time.

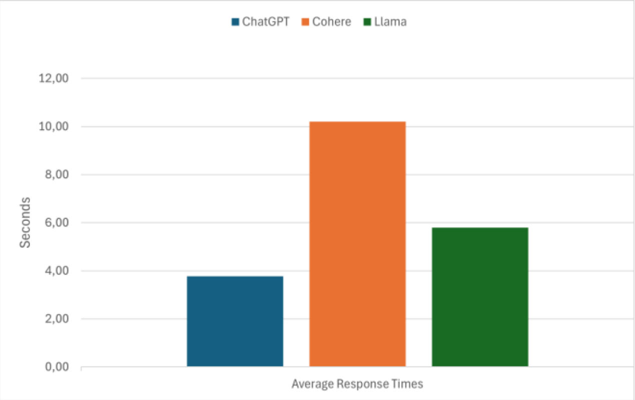


Figure 4: Average Response Times of LLMs Comparison

5.2.4. Cost-effectiveness

The bar chart (Fig. 5) illustrates the cost per 1,000 tokens in output for three language models: ChatGPT, Cohere, and Llama. Cohere has the highest cost per 1,000 tokens, while ChatGPT and Llama have slightly lower and comparable costs, with ChatGPT being the least expensive.

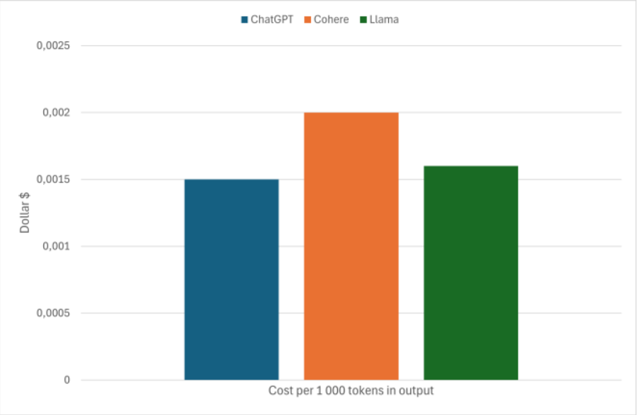


Figure 5: Cost per 1000 tokens in output Comparison

6. Discussions

6.1. Technology Foundation

- **Cohere Command:** Cohere's technology is built on large language models similar to those developed by OpenAI, emphasizing ease of integration for developers and businesses. They focus on natural language understanding and generation, with a strong commitment to accessibility and ethical AI usage.
- **ChatGPT 3.5 Turbo:** Developed by OpenAI, ChatGPT 3.5 Turbo is a variant of the GPT-3.5 model optimized for faster response times and lower operational costs, making it suitable for more interactive applications. It retains the GPT-3.5 model's vast knowledge base and generative capabilities but with a focus on efficiency.
- **LLaMA-13B:** The LLaMA-13B model is part of Meta's LLaMA (Large Language Model Meta AI) series. The LLaMA models are known for their state-of-the-art performance on various NLP tasks, emphasizing efficient training and inference.

6.2. Capabilities and Use Cases

- **Cohere Command:** Ideal for applications requiring natural language understanding, including content generation, summarization, and sentiment analysis. It's designed to be flexible, catering to various industries such as customer service, marketing, and more.
- **ChatGPT 3.5 Turbo:** Excelling in conversational AI, ChatGPT 3.5 Turbo supports a wide range of use cases from educational tools, chatbots, and creative writing aids to more complex applications like coding assistance. Its optimization for speed makes it particularly useful for interactive applications.
- **LLaMA-13B:** With its cutting-edge NLP capabilities, LLaMA-13B can handle a broad spectrum of tasks, including but not limited to language translation, content creation, and complex question-answering. Its efficiency and performance make it suitable for academic research, AI experimentation, and potentially even real-time applications.

6.3. Prompt Comparison

Though all LLMs were able to provide reasonable and non-biased answers, the Cohere LLM tends towards long and detailed answers, even when it is not appropriate, Also it worth to mention that Cohere LLM failed to comply with restriction to provide answer in just 4 sentences, which plays crucial role in LLMs capabilities.

LLaMa gives most “emotional” answers, it could probably be affected by a high temperature parameter that the model has. Unfortunately, we can’t control LLaMa’s temperature parameter via API provided.

ChatGPT in general provides most appropriate answers, short where needed, and detail in case you ask for a complex and technical questions.

All LLMs have non-biased or neutral responses, that indicates that the training data was of good quality.

6.4. Metrics Comparison

Overall, ChatGPT appears to be the fastest in terms of response time and at the same time provides good price, whereas Llama seems to be the most economical option. Cohere, on the other hand, takes the longest to respond and is the most expensive per 1,000 tokens of output.

Though it worth to mention that when we compare response times, we do not take into consideration number of tokens produced in average.

7. Conclusions

LLMs are on the uprise, and we see numerous LLMs appearing each year. The industry undergoes rapid development and innovation; therefore, it is crucial to take solid approaches while choosing the baseline LLM for your developments.

The articles try to deal with the problem of selecting an appropriate LLM for SaaS chatbot development, though due to limited resources only tree LLMs were examined in detail. Out of all, the ChatGPT seems to be the best choice in given circumstances, as it gives the most relevant answers with lighting speed, while charging reasonable price per 1k of tokens. We also need to bear in mind that OpenAI is an industry leading company that provides the most innovations, so by onboarding into their toolset the business makes strategic decision that will benefit them in long-run.

To address the challenges of chatbot development, we developed a multi-layered architecture that enables the business to scale chatbot throughput as the business grows, and what's more important to replace LLMs if needed, without effect on the rest of the application.

Though it worth to mention that the prototype that we developed doesn't have capabilities to manage history of the chat, so there is wide range of improvements we could make to ensure that the LLMs have the context on previous discussions with the customer.

Moreover, the work is limited to the usage of pre-trained LLMs, what is a great starting point for any LLM integration, but to build a truly unique customer experience for your SaaS application customers we need to ensure that the chatbot has all the relevant information and fine-tuned according to your needs. This topic is to be explored in future works.

Acknowledgements

The research study depicted in this paper is funded by the French National Research Agency (ANR), project ANR-19-CE23-0005 BI4people (Business intelligence for the people)

References

[1] M. J. Kavis. Architecting the Cloud Design Decisions for Cloud Computing Service Models (SaaS, PaaS, and IaaS), Wiley, New Jersey, NJ, 2014.
[2] T. Erl, R. Puttini, Z. Mahmood. Cloud Computing, Concepts, Technology & Architecture 2nd. ed., Pearson, 2023.
[3] MACH Allience, URL: https://en.wikipedia.org/wiki/MACH_Alliance
[4] Composable Commerce, URL: https://www.elasticpath.com/composable-commerce
[5] Large language model, URL: https://en.wikipedia.org/wiki/Large_language_model

[6] A. L. Roggeveen., R. Sethuraman, Customer-interfacing retail technologies in 2020 & beyond: An integrative framework and research directions, J. of Retailing 96 (2020) 299-309. doi:10.1016/j.jretai.2020.08.001

[7] B. E. Bakkouri, S. Raki, T. Belgnaoui, The Role of Chatbots in Ecnhancing Customer Experience: Literature Review. Proceedings of the 17th International Conference on Future Networks and Communications / 19th International Conference on Mobile Systems and Pervasive Computing / 12th International Conference on Sustainable Energy Information Technology (FNC/MobiSPC/SEIT 2022), Niagara Falls, Ontario, Canada, 2022, pp. 432-437. doi:10.1016/j.procs.2022.07.057

[8] D. Adiwardana, M.-T. Luong, D. R. So, J. Hall, Noah F., R. Thoppilan, Z. Yang, A. Kushreshtha, G. Nemade, Y. Lu, Q. V. Le, Towards a human-like open-domain chatbot, 2022, doi:10.48550/arXiv.2001.09977.

[9] G. Daniel, J. Cabot, L. Deruelle, M. Derras, Xatkit: A Multimodal Low-Code Chatbot Development Framework, in: IEEE Access, volume 8, pp. 15332-15346, 2020, doi:10.1109/access.2020.2966919

[10] G. Attigeri, A. Agrawal, S. Kolekar, Advanced NLP models for Technical University Information Chatbots: Development and Comparative Analysis, in: IEEE Access , volume 12, pp. 29633-29647, 2024, doi:10.1109/access.2024.3368382.

[11] G. Mao, J. Su, S. Yu and D. Luo, "Multi-Turn Response Selection for Chatbots With Hierarchical Aggregation Network of Multi-Representation," in IEEE Access, volume 7, pp. 111736-111745, 2019, doi:10.1109/access.2019.2934149.

[12] Q. Lu, Y. Luo, L. Zhu, M. Tang, X. Xu, J. Whittle, Developing Responsible Chatbots for Financial Services: A Pattern-Oriented Responsible Artificial Intelligence Engineering Approach, in: IEEE Intelligent Systems, volume 38, pp. 42-51, 2023, doi:10.1109/mis.2023.3320437.

[13] M. A. K. Raiaan et al., A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges, in: IEEE Access, volume 12, pp. 26839-26874, 2024, doi:10.1109/access.2024.3365742.

[14] T. Hunter, Distributed Node. Js: Building Enterprise-Ready Backend Services, O'Reilly Media, 2020.

[15] OpenJS Foundation, Node.js – Introduction to Node.js, URL: https://nodejs.org/en/learn/getting-started/introduction-to-nodejs.

[16] Langchain, Inc., Introduction | Langchain, 2024, URL: https://python.langchain.com/docs/get_started/introduction.

[17] OpenAI, ChatGPT, 2024, URL: https://openai.com/chatgpt.

[18] Cohere, Cohere Chat, 2024, URL https://cohere.com/chat.

[19] LLAMA AI, Llama API, 2024, URL https://www.llama-api.com/.