Journal of Information Systems Engineering and Management
2025, 10(34s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

How to Choose the Best AI LLM: A Guide to Navigating the Diversity of Models

Bery Leouro Mbaiossoum[1*]; Atteib Doutoum Mahamat[1] ; Narkoy Batouma[1], Lang Dionlar[1] , Batoure Bamana Apollinaire[2], Idriss Oumar Adam[1]
[1]*University of Ndjamena, Faculty of Exact and Applied Sciences, Chad*
[2]*University of Ngaoundere, Cameroon*
*Corresponding author: email: bery.mbaiossoum@gmail.com Ph: 00235-664008779

ARTICLE INFO

Received: 22 Dec 2024
Revised: 07 Feb 2025
Accepted: 18 Feb 2025

ABSTRACT

This article presents an in-depth analysis of Large Language Models (LLMs), a rapidly expanding technology. The main objective is to provide a comprehensive state-of-the-art overview, exploring recent advancements, challenges, and future perspectives. The methodology employed is based on a systematic review of scientific literature, combined with a comparative analysis of existing models. Some of the most popular LLMs, such as GPT-4, LaMDA, Claude, etc., are presented with their architectures and applications. This comparative analysis highlights the strengths and weaknesses of each model. The article also offers selection guidelines to help users choose the most suitable LLM for their needs. These guidelines are based on a multi-criteria approach, considering factors such as model size, performance, required resources, and ethical considerations. The impact of this article is significant, as it provides a valuable resource for researchers, developers, and LLM users. By offering a clear and structured overview of this complex field, it facilitates the understanding and adoption of these technologies. By following the selection guidelines presented in this article and staying informed about the latest trends, one can choose the LLM that will achieve its objectives effectively and efficiently.

Keywords: Artificial Intelligence, Large Language Models, ChatGPT, LaMDA, Choosing LLM

INTRODUCTION

AI Large Language Models (LLMs) have revolutionized many fields, from machine translation to content generation. More and more users are venturing into the use of these models in various domains such as human resources management [1], medicine [2, 3, 4], physics [5], education [6, 7], e-commerce [8], and applications [9, 10]. As [11] points out, "LLMs are becoming the core technology of AI, powering a new wave of applications and services that are transforming how we work, communicate, and interact with the world." It should be noted that this revolution comes with a major challenge: the diversity of available models. Indeed, the plethora of available models does not make the task easy for users and can make choosing an AI model difficult. It can be difficult even to find one's way around. A Gartner report highlights this complexity: "The LLM market is booming, with new players and new models constantly emerging. This rapid growth can make choosing the right model complex for businesses" [12]. We see no standardized norms or criteria for evaluating and comparing different LLM models. This can make it difficult for businesses to determine which models are the most efficient and best suited to their needs.

It is in this context that this article takes on its full meaning. It aims to provide a state-of-the-art overview of LLMs, present some popular models, and offer selection guidelines to help users choose the most suitable LLM for their needs.

Its primary objective is to explore the foundations of these models, their capabilities, and their limitations. A focus on their architectures will be made. Its second objective is to provide key guidelines for choosing the LLM best suited to specific needs, considering factors such as the task to be accomplished, budget, ease of use, confidentiality, and security.

The state of the art has covered many articles interesting to LLMs, but these articles do not focus on the issue of LLM selection. [13] introduced the concept of few-shot learning for LLMs, a technique that has revolutionized their ability to perform various tasks with little specific training. [14] explored the impact of model size and training datasets on LLM performance, paving the way for the creation of increasingly large and efficient models. [15] highlights the potential risks associated with LLMs, such as the propagation of biases, false information, and hate speech, emphasizing the need for a responsible and ethical approach to their development and use. [16] explores the intersection of Large Language Models (LLMs) and cognitive science, examining similarities and differences between LLMs and human cognitive processes. [17] presents a comprehensive review of evaluation methods for LLMs, focusing on three key dimensions: what to evaluate, where to evaluate, and how to evaluate. [6, 7] also explore various application scenarios of LLMs in classroom teaching, such as teacher-student collaboration, personalized learning, and assessment automation. [18] provides an overview of the history of LLMs, their evolution over time, delves into the working principles of language models, and analyzes different architectures of the GPT. [19] provide a survey with multiple perspectives on the utilization of LLMs in the multilingual scenario. It aims to help the research community address multilingual problems and provide a comprehensive understanding of the core concepts, key techniques, and latest developments in multilingual natural language processing based on LLMs. [5, 20] provide an overview of works relating to LLMs, the history of LLMs, their evolution over time, the architecture of transformers in LLMs, the different resources of LLMs, and the different training methods that have been used to train them.

This article will serve as a guide for some users by giving them confidence when they navigate the LLM landscape and allow them to make an informed choice that will propel their projects to success.

The rest of the article is organized as follows: the presentation of the research methodology is described in section II, section III presents the results of our bibliometric study; contributions are discussed in Section IV and Section V focuses on the conclusion and future work.

METHODOLOGY

Our approach to developing this article consisted of an in-depth review of the scientific and technical literature devoted to Large Language Models (LLMs). This exploration was conducted through the use of academic search engines such as Google Scholar, JSTOR, and ACM Digital Library, as well as specialized databases in the field of artificial intelligence, such as arXiv [https://arxiv.org/list/cs.AI/recent] and Semantic Scholar [https://www.semanticscholar.org/]. We favored peer-reviewed research articles, conference proceedings, and reference publications to ensure the rigor and validity of the information collected.

In parallel with this academic review, we also consulted blog articles, industry publications, and technical reports from recognized organizations in the field of AI, such as OpenAI [https://openai.com/], Google AI [https://ai.google/latest-news/], DeepMind [https://deepmind.google], Meta AI [https://research.facebook.com], and Anthropic [https://www.anthropic.com/]. This approach allowed us to supplement our analysis with up-to-date information on the latest LLM models, their practical applications, and the challenges they face.

By combining these different sources of information, we were able to provide a comprehensive and up-to-date state-of-the-art overview of LLMs, identify the most popular models and their applications, and develop relevant selection guidelines to help users choose the LLM best suited to their needs.

RESULTS

State of the Art of LLMs

Large Language Models (LLMs) have become key players in the field of artificial intelligence. Trained on immense corpora of textual data, they have the ability to understand and generate text with remarkable sophistication. This feat opens the door to a multitude of applications, ranging from machine translation to creative content creation. Indeed, LLMs excel in understanding and generating natural language, which allows them to perform various tasks:

- Machine Translation: They can translate text from one language to another with surprising accuracy.
- Text Summarization: They can condense long documents into a few key sentences, preserving the essence of the message.

- Creative Text Generation: They can write poems, stories, articles, etc., with a style sometimes indistinguishable from that of a human.
- Question Answering: They can provide accurate and relevant answers to questions asked in natural language, relying on their encyclopedic knowledge.
- Sentiment Analysis: They can determine whether a text expresses a positive, negative, or neutral opinion, which is useful for analyzing customer reviews or social media conversations.

However, despite their spectacular advances, LLMs still face challenges, including:

- Bias: LLMs can reflect the biases present in the data on which they were trained, which can lead to discrimination or stereotypes.
- Lack of Transparency: The internal workings of LLMs can be opaque, making it difficult to understand their decisions.
- Cost: Training and using LLMs can be expensive, limiting their access to certain organizations or individuals.

However, research continues to address these challenges and improve LLMs. The prospects are promising, with constant advances in areas such as bias reduction, transparency improvement, and cost reduction.

Popular LLMs

In the literature, we encounter several LLMs, the most popular currently available, each with its strengths and specialties, are presented in this work.

GPT-4 (OpenAI)

GPT-4 is OpenAI's latest model, succeeding GPT-3. It is recognized for its creative text generation capabilities, its understanding of natural language, and its ability to handle complex tasks. According to OpenAI, GPT-4 is "more creative and collaborative than ever" and can "handle much more complex instructions" [21]. It can generate ideas, songs, creative texts, etc. It can also learn a user's writing style and adapt to it" [21].

GPT-4 is a multimodal language model based on the Transformer architecture (Fig. 1), but with major innovations in design and capabilities [20, 22]. It marks a significant evolution compared to its predecessors, notably thanks to its hybrid architecture and multimodal approach.

GPT-4 utilizes a Mixture of Experts (MoE) architecture, composed of 16 specialized sub-networks (or "experts"), each containing approximately 100 billion parameters [22]. For each query, only 2 experts are activated, which enhances efficiency and reduces computational costs. This approach allows for specialization in various tasks while sharing 50 billion common parameters for attention mechanisms. GPT-4 can process up to 128,000 tokens (approximately 96,000 words), compared to 4,096 tokens for GPT-3.5. This enables it to analyze long documents (up to 50 pages) or complex conversations without losing track of the exchange [22]. Unlike GPT-3.5, GPT-4 integrates a native multimodal capability. For example, it can analyze a provided medical diagram and generate a textual diagnosis [23]. Table 1 presents a comparison of some characteristics of GPT-4 and GPT-3.5.

[Table 1]

A significant innovation was introduced with the GPT-4o ("omni") version, which unifies modalities (text, audio, image) into a single model, unlike previous versions that used disjoint systems. This allows for:

- Real-time interaction with human-latency voice responses.
- Better understanding of emotions and nuances in audio inputs [27].
- Superior performance in multilingual speech recognition and translation [26].

GPT-4 has been refined through Reinforcement Learning from Human Feedback (RLHF), incorporating human feedback to reduce biases and dangerous responses. Emphasis is placed on security to test and limit abusive uses such as malicious code generation [26]. Three points are noted as limitations of GPT-4: transparency, data recency, and cost. Regarding transparency, details about training and architecture are not presented [28]. Concerning data recency, GPT-4's knowledge is primarily based on data prior to 2023 [26]. As for cost, the required infrastructure (computation, storage) remains prohibitive for replication outside of OpenAI [29].

LaMDA (Google AI)

LaMDA (Language Model for Dialogue Applications) is a language model developed by Google, designed to be conversational and informative. It has been presented as a model capable of engaging in natural and engaging conversations on a variety of topics. [30] emphasizes that LaMDA is "capable of understanding and responding to complex questions" and that it can "generate original and creative text."

LaMDA is built on the Transformer architecture, which was introduced by Google Research in 2017 [31]. Transformer is a neural network architecture consisting of an encoder and a decoder. The encoder takes a sequence of words as input and transforms it into a vector representation. This vector representation contains information about each word in the sequence and their relationship to each other. The decoder uses the vector representation generated by the encoder to generate an output. The encoder and decoder can be parallelized. Figure 1 presents this architecture.

[Figure 1]

This architecture is fundamental to modern natural language processing and allows the model to:

- Analyze relationships between words: The Transformer uses attention mechanisms to understand how words in a sentence or paragraph interact with each other, which is essential for predicting the next words in a dialogue.
- Manage long sequences: Thanks to its ability to process longer text sequences, LaMDA can maintain context across multiple exchanges in a conversation.

Unlike other language models that are often trained on various corpora, LaMDA has been specifically trained on dialogues [32]. This allows it to learn the nuances and subtleties that characterize human conversations, such as:

- Contextual sensitivity: The ability to generate responses that make sense in the context of a given conversation.
- Fluency and relevance: LaMDA is designed to produce responses that seem natural and engaging, which is crucial for realistic human interactions.

These characteristics make LaMDA particularly suitable for applications such as:

- Advanced chatbots: Used in customer service and other areas where simulated human interaction is desired [33].
- Virtual assistants: Capable of handling complex and varied queries.

When LaMDA answers a question or engages in a conversation, it generates several potential responses. These responses are then evaluated and ranked based on their relevance and quality, allowing the model to select the most appropriate response [33, 34]. Regarding security and ethics, measures are in place to prevent the generation of harmful or biased content [35]. The model has been tested to ensure it meets high standards of safety and ethics.

Claude (Anthropic)

Claude is an LLM developed by Anthropic, a company co-founded by former OpenAI employees. It distinguishes itself by its approach focused on safety and ethics [37]. It was designed to be a helpful, honest, and harmless assistant, with a conversational tone. Claude is capable of understanding and answering questions, generating creative text, translating languages, writing different types of creative content, and answering questions informatively. Its design was made to avoid inappropriate or dangerous responses, using techniques such as "Constitutional AI" [36].

Although the precise technical details of Claude's architecture are not publicly available, it is known that it is a large language model based on transformers, similar to other LLMs like GPT-3 [20, 31]. It was trained on a large amount of textual data to learn the relationships between words and concepts. Anthropic emphasizes safety and ethics in the development of its models, using techniques to reduce the risks of bias, false information, and undesirable behaviors. Claude stands out in several aspects:

- Emphasis on safety and ethics: Anthropic has implemented rigorous safety measures to minimize the risks associated with LLMs, such as the generation of inappropriate or biased content [37].
- Design for conversation: Claude is optimized for conversational interactions, with the ability to understand context and provide relevant and natural responses [38].
- Versatility: Claude can be used for various tasks, ranging from creative writing to answering complex questions [39].

Anthropic offers different access options to Claude, including an API for developers and a Pro version for individual users. Prices vary depending on usage and needs. Like all LLMs, Claude has certain limitations [40]:

- Bias: Although efforts are made to reduce biases, they can still be present in the training data and affect Claude's responses.
- Incorrect information: Claude can sometimes generate incorrect or misleading information, as it does not possess a real-world understanding and relies on textual data.
- Lack of common sense: Claude may struggle with certain tasks that require common sense or a deep understanding of context.

Despite its limitations, Claude offers many advantages [37, 40]:

- Writing assistance: Claude can help with writing creative texts, emails, reports, etc.
- Quick responses: Claude can provide quick answers to complex questions, which can be useful for information retrieval.
- Language translation: Claude can translate text from one language to another.
- Learning: Interacting with Claude can be an interesting learning experience, as it can provide new perspectives and information.

Llama 2 (Meta)

Llama 2 is an open-source model developed by Meta. It offers great flexibility to developers, who can adapt it to their specific needs. [41] indicates that Llama 2 is "a powerful and versatile model" that can be used for "a variety of tasks, from text generation to machine translation." Llama 2 was released in July 2023 and is available in different sizes, with a number of parameters ranging from 7 billion to 70 billion [41].

Llama 2 is based on the Transformer architecture [20, 26], which is a standard architecture for LLMs. It uses an attention mechanism to process text, which allows it to understand context and generate coherent text. The tuned versions use supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align to human preferences for helpfulness and safety [41]. Llama 2 was pretrained on 2 trillion tokens of data from publicly available sources.

Llama 2 stands out for its open-source nature. It is freely accessible for research and commercial use, which differentiates it from many other proprietary LLMs. This open approach fosters innovation and allows a wide range of developers and researchers to work with the model. Access to Llama 2 is free, however, it is important to note that running Llama 2 may require significant computing resources, which can lead to indirect costs.

Llama 2 has the following limitations [41]:

- Resources: Running Llama 2 may require significant computing resources, which can be an obstacle for some users.
- Use: Like any LLM, Llama 2 can be used for malicious purposes, such as generating false information or hate speech.

On the other hand, Llama 2 has some advantages, including [41]:

- Open source: Access to Llama 2 is free. This allows a wide range of people to use and improve it.
- Performance: Llama 2 has demonstrated competitive performance compared to other LLMs in its class.
- Flexibility: Llama 2 is available in different sizes, allowing it to be adapted to different needs and resources.

Mistral

Mistral is a large language model (LLM) developed by the French startup of the same name. It stands out for its innovative architecture and promising performance. It offers several models, with numerous variations, including multimodal and open-source versions. Thanks to its ability to understand natural language in French, Mistral can be used for a wide variety of tasks, such as text generation, machine translation, text classification, information extraction, and speech synthesis [42]. Mistral Large has been trained on a diverse corpus of texts, including news articles, books, websites, and online conversations, which allows it to understand a wide range of writing styles and language registers.

Mistral uses a transformer architecture [26], like most recent LLMs. However, it distinguishes itself by using a "grouped-query attention" (GQA) mechanism [43, 44]. This technique reduces the complexity and computational cost during inference, making Mistral faster and less resource-intensive than some of its competitors.

Mistral has the following particularities:

- Efficiency: Mistral's architecture is designed to be performant while requiring fewer resources, making it potentially more accessible for large-scale use.
- Focus on French: Although Mistral can handle multiple languages, it has been trained with particular attention to French, which gives it a good command of this language.
- Open source: Mistral has been released as open source, which means its code is accessible to everyone. This fosters collaboration and innovation around this model.

Since Mistral is open source, its direct usage cost is zero. However, the costs associated with the infrastructure needed to run it (servers, computing power, etc.) must be taken into account. These costs may vary depending on the usage and scale of the project.

Like any LLM, Mistral has its limitations [43]:

- Limited knowledge: Although it is trained on a large amount of data, Mistral has gaps in certain knowledge domains.
- Bias: LLMs are likely to reflect biases present in the data they were trained on. Mistral is no exception and may sometimes produce biased or discriminatory results.
- Content generation: Although it can generate text fluently and coherently, it can sometimes produce incorrect or misleading information.

Despite its limitations, Mistral has several advantages:

- Performance: It offers good performance in various tasks, including understanding and generating text in French.
- Efficiency: Its optimized architecture makes it faster and less expensive to use than some other LLMs.

- Open source: Its open-source nature fosters innovation and collaboration.

LLM Gemini

Gemini is a large language model (LLM) developed by Google DeepMind [45]. It was designed to be multimodal, meaning it can understand and process different types of information, such as text, code, images, and audio [46].

Gemini's architecture is based on the Transformer model [47], which is a neural network architecture commonly used for LLMs. However, Gemini stands out for its ability to handle multiple modalities, which allows it to better understand the world around it and provide more complete and relevant responses [48, 49].

Gemini's particularities include:

- Multimodality: Gemini can understand and integrate information from different sources, which allows it to provide richer and more contextual responses [46].
- Advanced capabilities: Gemini has been trained on a large amount of data, which allows it to perform complex tasks such as translation, creative text generation, and answering difficult questions [48].
- Integration with other Google products: Gemini is integrated with other Google products, such as the search engine and Bard, which allows it to improve their functionalities [49].

The cost of using Gemini will depend on the specific use you make of it. Google offers different options, including free and paid versions. Gemini has the following limitations:

- Bias: Like all LLMs, Gemini can exhibit biases because it was trained on data that may contain prejudices.
- Hallucinations: Gemini can sometimes generate incorrect or invented information, which is a common problem for LLMs.
- Resource requirements: Using Gemini may require significant resources in terms of computing power and memory.

Gemini has the following advantages [50]:

- Versatility: Gemini can be used for a wide range of tasks, making it a powerful tool for many fields.
- Performance: Gemini has demonstrated high performance in various benchmarks, making it one of the most powerful LLMs.
- Continuous improvement: Google continues to develop and improve Gemini, which means its capabilities are likely to continue to improve in the future.

There are many other LLM we do not present here such as BERT, Deep seek, T5, Bart, etc. This state-of-the-art allows us to know that there are several LLMs and that each LLM has its limitations and advantages. We also note that LLMs use different architectures, but the dominant architecture is the Transformer model [20, 37, 50]. Some LLMs are specific and oriented to precise domains, others are intended to be generalists and attempt to respond to all requests from any domain.

Factors in Choosing an LLM: An In-Depth Analysis

Choosing a large language model (LLM) is a strategic decision that must be guided by a clear understanding of specific needs, budget constraints, and security requirements. Below, we present an exploration of the key factors to consider when selecting an LLM.

Task to be Accomplished

The task to be accomplished can guide the choice of LLM based on its nature and complexity. The nature of the task can be:

- Answering questions: For applications such as chatbots or search engines, LLMs focused on accuracy and information retrieval are suitable. Appropriate examples are GPT-4, Gemini, BERT (Devlin et al., 2018).

- Text generation: If the goal is to generate original content (poems, stories, articles), LLMs like GPT-4 [26] or LaMDA [30] are excellent choices due to their advanced creative capabilities [11].
- Sentiment analysis: If analyzing opinions or emotions is required, specialized models like XLNet, GPT-4 can offer better performance.

As for the complexity of the task, it is noted that there are:

- Simple tasks like basic text classification, where smaller and faster models may suffice.
- Complex tasks requiring larger and more sophisticated LLMs, capable of handling linguistic and contextual nuances [13].

Budget

The budget can also play a crucial role in choosing an LLM. Indeed, depending on the budget, one can decide to use a free or paid LLM. This involves the cost of use. One can choose open-source models: in fact, open-source LLMs like Llama 2 (Meta) offer a free alternative. However, these require computing resources for training and deployment. One can choose a paid model (API): Commercial LLMs like GPT-4 are accessible via paid APIs, with costs varying based on usage.

Alongside the cost of use, it is noted that there may be a need for computing resources. This can be:

- A need for computing power: Larger LLMs require considerable computing power for training and inference, which can lead to high costs [51].
- A need for cost optimization: Techniques exist to optimize costs, such as model compression, knowledge distillation, and the use of specialized hardware [52].

Ease of Use

Another quite important factor in the choice of LLMs is ease of use. Some LLMs have an intuitive interface or produce easily integrable APIs. For example, most commercial LLMs are accessible via APIs, which facilitates their integration into existing applications. It is also noted that libraries like Transformers [53] simplify the use of open-source LLMs. Good documentation and usage examples are essential to facilitate the adoption of LLMs. The existence of responsive technical support that can help solve problems and optimize the use of an LLM is a major asset in its selection.

Confidentiality and Security

Data protection and model security can influence the choice of an LLM. Indeed, some data can be very sensitive (for example, personal, banking, or medical information), the chosen LLM must comply with current regulations (GDPR, HIPAA, etc.), in order to protect this data [54]. We can consider encryption and anonymization as security measures to protect data during its processing by the LLM. It should be noted that LLMs can be vulnerable to attacks, such as data poisoning or adversarial attacks. It is necessary to ensure whether security measures to protect the model against these attacks and guarantee its integrity are proposed by the LLM or are possible locally.

Other Factors

Other means that can guide the choice of an LLM that can be mentioned include:

- Performance and accuracy: consists of evaluating the performance of LLMs on specific tasks, and comparing the results on benchmark datasets and choosing the best. But, this has a cost that can be exorbitant [55].
- Scalability: when it is recommended to use the LLM for large-scale applications, and manage large volumes of data and frequent queries without compromising performance, this dimension must be taken into account [56].
- Personalization and adaptation: some LLMs can be customized or adapted to specific use cases. This flexibility must be considered in the choice of the LLM.

- Community: It is also necessary to ensure that the LLM has complete documentation and an active community of users. These elements are important in case of difficulties during the use of the LLM. Their existence could encourage the choice of the LLM.

Choosing the ideal LLM is an iterative process that requires careful evaluation of needs, resources, and constraints. By considering these key factors and relying on the provided bibliographic references, you will be able to make an informed decision and select the LLM that will allow you to achieve your goals successfully.

DISCUSSION

Choosing the best LLM is an important decision that can have a significant impact on the success of a project. This choice is an iterative process that requires careful evaluation of needs, resources, and constraints. By considering the factors mentioned above and having information on the different models available, you can choose the LLM best suited to your project.

The nature of the work to be done is a determining factor. It will lead us to consider LLMs that excel in our type of work. We have identified the main types of tasks to be performed and have proposed some LLMs. For "question and answer" type work, chatbots or search engines with AI, LLMs focused on accuracy and information retrieval are best suited. This type of work can be the subject of use of weak or specialized AI in the fields of work. If the goal is to generate content such as poems, stories, or articles, GPT-4 or LaMDA are good candidates, due to their advanced creative abilities [11]. However, we can be wary of biases and data imaginations. For the choice, it would be interesting to take into account the types of data on which the work is based.

For a simple task, most LLMs can do it with more or less satisfactory results that experience or trials can help in the choice of the tool. But for complex tasks such as sentiment analysis or automatic language translation, we must look towards larger and more sophisticated LLMs, capable of managing linguistic and contextual nuances [13]. [11] finds that GPT-4 and LaMDA are good candidates, due to their advanced capabilities. DeepSeek, which aims to be a competitor to Chat-GPT-4, can also be interesting [57].

The proprietary or open-source nature of the LLM can also influence its choice, depending on the importance of the budget and the nature of the project. For learning or experimentation work, open-source LLMs are good candidates.

It is important to note that the field of LLMs is in full swing, with new models and new features appearing regularly. It is therefore essential to stay informed of the latest advances to make the best possible choice. Useful resources for following LLM news include specialized artificial intelligence journals that regularly publish articles on LLMs. Among the most prestigious, we can mention the "Journal of Machine Learning Research" and "Artificial Intelligence". AI conferences and symposia are an opportunity to discover the latest research and developments in LLMs. The NeurIPS (Neural Information Processing Systems) and ICML (International Conference on Machine Learning) conferences are particularly renowned in this field. Blogs and websites specializing in AI also offer articles and analyses on LLMs. Among the most popular, we can mention "Towards Data Science" and "Synced Review".

CONCLUSION

The current landscape of LLMs offers a multitude of options to meet diverse needs. Choosing the ideal model involves carefully evaluating your objectives, allocated budget, and specific project requirements. LLMs distinguish themselves by their performance, ability to handle different tasks, cost, and ease of integration. Some excel in creative text generation, while others specialize in translation or answering questions. It is crucial to stay informed about the latest advancements in this constantly evolving field. Research publications, conferences, and online communities are excellent sources of information. Don't hesitate to experiment with different models to determine which one best suits your needs. The key is to find a balance between performance, cost, and ease of use.

By following these tips and adapting to the constant progress of technology, you will be able to choose the LLM that will allow you to achieve your goals effectively and efficiently.

REFRENCES

[1] Ajzen, M., Patesson, L., & Inglebert-Frydman, A. (2024). Intelligence Artificielle et Gestion des Ressources Humaines. À quelles conditions le recours à l'IA peut-il être un atout pour la GRH?.

[2] Tiffany H. Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. 2023. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PloS Digital Health 2, 2 (2023), e0000198.

[3] Nov, O., Singh, N., & Mann, D. (2023). Putting ChatGPT's medical advice to the (Turing) test: survey study. JMIR Medical Education, 9, e46939.

[4] Jeblick, K., Schachtner, B., Dexl, J., Mittermeier, A., Stüber, A. T., Topalis, J., ... & Ingrisch, M. (2024). ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. European radiology, 34(5), 2817-2825.

[5] Holmes, J., Liu, Z., Zhang, L., Ding, Y., Sio, T. T., McGee, L. A., Lisa A.; Ashman, Jonathan B., Li, X., Liu, T., Shen, J. & Liu, W. (2023). Evaluating large language models on a highly-specialized topic, radiation oncology physics. Frontiers in Oncology, 13, 1219326.

[6] Tan, K., Pang, T., Fan, C., & Yu, S. (2023). Towards applying powerful large ai models in classroom teaching: Opportunities, challenges and prospects. arXiv preprint arXiv:2305.03433.

[7] Kamalov, F., Santandreu Calonge, D., & Gurrib, I. (2023). New era of artificial intelligence in education: Towards a sustainable multifaceted revolution. Sustainability, 15(16), 12451.

[8] Palen-Michel, C., Wang, R., Zhang, Y., Yu, D., Xu, C., & Wu, Z. (2024). Investigating LLM Applications in E-Commerce. arXiv preprint arXiv:2408.12779.

[9] Topsakal, O., & Akinci, T. C. (2023, July). Creating large language model applications utilizing langchain: A primer on developing llm apps fast. In International Conference on Applied Engineering and Natural Sciences (Vol. 1, No. 1, pp. 1050-1056).

[10] Chen, X., Gao, C., Chen, C., Zhang, G., & Liu, Y. (2025). An Empirical Study on Challenges for LLM Application Developers. ACM Transactions on Software Engineering and Methodology

[11] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv:2303.12712.

[12] Linden, A., & Fenn, J. (2003). Understanding Gartner's hype cycles. Strategic Analysis Report Nº R-20-1971. Gartner, Inc, 88, 1423.

[13] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

[14] Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

[15] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big? 🦜. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 610-623).

[16] Niu, Q., Liu, J., Bi, Z., Feng, P., Peng, B., Chen, K., ... & Fei, C. (2024). Large language models and cognitive science: A comprehensive review of similarities, differences, and challenges. arXiv preprint arXiv:2409.02387.

[17] Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., ... & Xie, X. (2024). A survey on evaluation of large language models. ACM transactions on intelligent systems and technology, 15(3), 1-45.

[18] Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., ... & Mirjalili, S. (2023). A survey on large language models: Applications, challenges, limitations, and practical usage. Authorea Preprints, 3.

[19] Huang, K., Mo, F., Zhang, X., Li, H., Li, Y., Zhang, Y., ... & Liu, Y. (2024). A survey on large language models with multilingualism: Recent advances and new frontiers. arXiv preprint arXiv:2405.10936.

[20] Raiaan, M. A. K., Mukta, M. S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M. J., ... & Azam, S. (2024). A review on large language models: Architectures, applications, taxonomies, open issues and challenges. IEEE access, 12, 26839-26874.

[21] OpenAI, (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

[22] Schreiner, M. (2023). GPT-4 architecture, datasets, costs and more leaked. The decoder, 11.

[23] [23] Reddy, S., Schwartzman, G., & Flowers, R. H. (2023). ChatGPT in dermatology clinical practice: potential uses and pitfalls. Cutis, 112(2), E15-E17.

[24] Karttunen, P. (2023). LARGE LANGUAGE MODELS IN HEALTHCARE DECISION SUPPORT. Tampere University.

[25] Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 30(4), 681–694. https://doi.org/10.1007/s11023-020-09548-1

[26] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.
[27] Hern, A., & Bhuiyan, J. (2023). OpenAI says new model GPT-4 is more creative and less likely to invent facts. The Guardian, 14.
[28] Sanderson, K. (2023). GPT-4 is here: what scientists think. Nature, 615(7954), 773.
[29] Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., ... & Kivlichan, I. (2024). Gpt-4o system card. arXiv preprint arXiv:2410.21276.
[30] Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H. T., ... & Le, Q. (2022). Lamda: Language models for dialog applications. arXiv preprint arXiv:2201.08239.
[31] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
[32] Buchanan, M. (2023). Define the IQ of a chatbot. Nature Physics, 19(4), 465-465.
[33] Cohen, A. D., Roberts, A., Molina, A., Butryna, A., Jin, A., Kulshreshtha, A., ... & Chen, Z. (2022). Lamda : Language models for dialog applications. arXiv preprint arXiv :2201.08239.
[34] O'Leary, D. E. (2023). An analysis of three chatbots: BlenderBot, ChatGPT and LaMDA. Intelligent Systems in Accounting, Finance and Management, 30(1), 41-54.
[35] Yuan, Y., Jiao, W., Wang, W., Huang, J. T., He, P., Shi, S., & Tu, Z. (2023). Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher. arXiv preprint arXiv:2308.06463.
[36] Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., ... & Kaplan, J. (2022). Constitutional ai: Harmlessness from ai feedback. arXiv preprint arXiv:2212.08073.
[37] Menon, S. (2024). Being "LaMDA" and the Person of the Self in AI. In AI, Consciousness and The New Humanism: Fundamental Reflections on Minds and Machines (pp. 331-349). Singapore: Springer Nature Singapore.
[38] Dhar, P., Gupta, K., & Ranjan, R. (2024, September). Lamda: label agnostic mixup for domain adaptation in iris recognition. In 2024 IEEE international joint conference on biometrics (IJCB) (pp. 1-10). IEEE.
[39] Matarazzo, A., & Torlone, R. (2025). A Survey on Large Language Models with some Insights on their Capabilities and Limitations. arXiv preprint arXiv:2501.04040.
[40] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.
[41] Kostina, A., Dikaiakos, M. D., Stefanidis, D., & Pallis, G. (2025). Large Language Models For Text Classification: Case Study And Comprehensive Review. arXiv preprint arXiv:2501.08457.
[42] Li, M., Yang, H., Liu, Z., Alam, M. M., Sack, H., & Gesese, G. A. (2024). KGMistral: Towards boosting the performance of large language models for question answering with knowledge graph integration. In Workshop on Deep Learning and Large Language Models for Knowledge Graphs.
[43] Fernández, F. C., Garcia, R. L. S., & Caarls, W. (2024, November). Comparison of LLM Models and Strategies for Structured Query Construction from Natural Language Queries. In 2024 IEEE Latin American Conference on Computational Intelligence (LA-CCI) (pp. 1-6). IEEE.
[44] Islam, R., & Ahmed, I. (2024, May). Gemini-the most powerful LLM: Myth or Truth. In 2024 5th Information Communication Technologies Conference (ICTC) (pp. 303-308). IEEE.
[45] Team, G., Anil, R., Borgeaud, S., Alayrac, J. B., Yu, J., Soricut, R., ... & Blanco, L. (2023). Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805.
[46] Li, Y., Zhang, Y., Wang, C., Zhong, Z., Chen, Y., Chu, R., ... & Jia, J. (2024). Mini-gemini: Mining the potential of multi-modality vision language models. arXiv preprint arXiv:2403.18814.
[47] Imran, M., & Almusharraf, N. (2024). Google Gemini as a next generation AI educational tool: a review of emerging educational technology. Smart Learning Environments, 11(1), 22.
[48] Pande, A., Patil, R., Mukkemwar, R., Panchal, R., & Bhoite, S. (2024). Comprehensive Study of Google Gemini and Text Generating Models: Understanding Capabilities and Performance. Grenze International Journal of Engineering & Technology (GIJET), 10.
[49] Rane, N., Choudhary, S., & Rane, J. (2024). Gemini versus ChatGPT: applications, performance, architecture, capabilities, and implementation. Journal of Applied Artificial Intelligence, 5(1), 69-93.
[50] Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. T., Abid, A., Fisch, A., ... & Liang, P. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv preprint arXiv:2206.04615
[51] Wu, S., Fei, H., Qu, L., Ji, W., & Chua, T. S. (2024, July). Next-gpt: Any-to-any multimodal llm. In Forty-first International Conference on Machine Learning.
[52] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020, October). Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations (pp. 38-45).

[53] Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024). A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. High-Confidence Computing, 100211.

[54] Li, S., Ning, X., Hong, K., Liu, T., Wang, L., Li, X., ... & Wang, Y. (2023). Llm-mq: Mixed-precision quantization for efficient llm deployment. In NeurIPS 2023 Efficient Natural Language and Speech Processing Workshop (pp. 1-5).

[55] Ganesh, A. (2024, November). Program Scalability Analysis for LLM Endpoints: Ahmdal's Law Analysis of Parallelizability Benefits of LLM Completions Endpoints. In 2024 4th International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME) (pp. 1-6). IEEE.

[56] Peng, Y., Malin, B. A., Rousseau, J. F., Wang, Y., Xu, Z., Xu, X., ... & Bian, J. (2025). From GPT to DeepSeek: Significant gaps remains in realizing AI in healthcare. Journal of Biomedical Informatics, 104791.