ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal
Regular Issue, Vol. 14 (2025), e31857
eISSN: 2255-2863
DOI: https://doi.org/10.14201/adcaij.31857

Comparison of Leading AI Models an Analytical Study of ChatGPT Google Bard and Microsoft Bing

Pascal Adomako, Talha Ali Khan, Raja Hashim Ali, and Rand Koutaly

Department of Business, University of Europe for Applied Sciences. Think Campus, Konrad-Zuse-Ring 11, 14469 Potsdam, Germany
✉ pascal.adomako@ue-germany.de, talhaali.khan@ue-germany.de, hashim.ali@ue-germany.de, rand.kouatly@ue-germany.de

KEYWORDS

prompt engineering; LLMs; ChatGPT; Google Bard; Microsoft Bing Chat; encoder-decoder structure; attention mechanism; conversational skills; creative text; generation; translation; question answering; code generation; NLP

ABSTRACT

This comparative analysis delves into the capabilities of three prominent Conversational AI models: ChatGPT, Google Bard, and Microsoft Bing Chat. The study encompasses a meticulous exploration of their conversational skills, natural language processing abilities, and creative text generation. Methodologically, this study crafts a comprehensive evaluation framework, including complexity levels and tasks for each dimension. Through user-generated responses, key metrics of fluency, coherence, relevance, accuracy, completeness, informativeness, creativity, and relevance were assessed. The results reveal distinctive strengths of the AI models. The theoretical implications lead to recommendations for dynamic learning, ethical considerations, and cross-cultural adaptability. Practically, avenues for future research were proposed, including real-time user feedback integration, multimodal capabilities exploration, and collaborative human-AI interaction studies. The analysis sets the stage for benchmarking and environmental impact assessments, underlining the need for standardized metrics.

1. Introduction

The era of artificial intelligence (AI) has seen a dramatic shift from traditional rule-based systems to sophisticated machine learning (ML) techniques (Velásquez-Henao et al., 2023). Among these advancements, natural language processing (NLP) has emerged as a pivotal domain, enabling seamless human-computer interactions. Large language models (LLMs) represent a transformative milestone in AI, particularly in the field of conversational AI, which includes models such as ChatGPT, Google Bard, and Microsoft Bing Chat (Kim, 2010; Qin et al., 2022). These models leverage deep learning architectures, such as the transformer (Vaswani et al., 2017), to process and generate human-like text.

ChatGPT, developed by OpenAI, has gained widespread attention for its ability to generate contextually coherent and semantically rich responses (Wu et al., 2023). Google Bard, with its foundation in Google's advanced PaLM 2 architecture, excels in multilingual capabilities and creative text generation (Anil et al., 2023). Microsoft Bing Chat, powered by GPT-4 and Microsoft's proprietary Prometheus model, integrates conversational AI directly into search, offering highly contextual and relevant responses (Chowdhery et al., 2022). Despite their individual strengths, the comparative evaluation of these models across critical metrics remains an underexplored area, which this study aims to address.

To provide a nuanced understanding of these conversational AI models, this paper adopts a structured evaluation framework. By leveraging user-generated responses, we assess key metrics such as fluency, coherence, relevance, accuracy, creativity, and completeness. This study contributes to an evolving field by identifying the distinctive strengths of each model, offering insights into their practical applications, and proposing future research directions.

Despite discussions on the topic, there is a research gap in state-of-the-art literature regarding systematic and comparative studies. With the aim of filling this gap, the present paper proposes a comprehensive comparison using prompt engineering, an AI technique refining LLMs' performance through specific prompts. Trial questions assess Google Bard AI, ChatGPT, and Microsoft Bing AI, aiming to provide insights for informed decisions and contribute to the evolving landscape of AI technology.

2. Introduction to Key Technologies and Concepts

2.1. Natural Language Processing

Natural language processing (NLP), situated at the intersection of linguistics, computer science, and artificial intelligence, focuses on the interaction between human language and computational systems. It strives to enable machines to understand, interpret, and respond meaningfully to human language, addressing challenges such as natural language production, semantic analysis, grammar, pragmatics, and understanding. Drawing from various disciplines, NLP plays a crucial role in empowering computers to engage with human language across diverse applications.

2.2. Evolution of Natural Language Processing

The evolution of NLP has transitioned from rule-based systems to ML techniques and, more prominently, to deep learning approaches. Initially, NLP utilized rule-based approaches, with linguists and experts manually setting rules to understand language, which had inherent limitations due to the intricate nature of language (Kim, 2010). This phase was succeeded by ML-based techniques, particularly

deep learning models, which enable computers to automatically learn language patterns directly from data (Samant et al., 2022).

One pivotal advancement in this journey is the development of word embeddings, which are vector representations of words that capture their semantic and syntactic relationships. Notable models such as Word2Vec and GloVe have played crucial roles in generating these embeddings from vast textual data, thus allowing for more profound language comprehension (Biswas and De, 2022).

Subsequently, techniques such as recurrent neural networks (RNNs), especially long short-term memory (LSTM) networks, have been instrumental in modelling sequential language data, such as in machine translation. Meanwhile, transformers have emerged as powerful tools for addressing long-range dependencies in text, proving beneficial for tasks such as text classification and sentiment analysis (Qin et al., 2022) .

Driving this progress in NLP are a series of factors, including the surge in available text data from sources such as the internet and social media, as well as the advances in computational hardware, such as graphics processing units (GPUs). These have facilitated the development, training, and deployment of complex models at an exponential rate.

The surge in computational methods and their applications in NLP, coupled with vast data availability, has strongly propelled the academic community to focus on this research area. In a recent study conducted by Chen et al. (2022), the bibliographic data of NLP scientific papers were retrieved from the Web of Science (WoS) database. The search query used in the study was based on prior reviews of NLP (Kreimeyer et al., 2017; Pons et al., 2016). The search generated a total of 31,485 NLP papers. The study analyzed the number of NLP papers by year and found that there has been an overall growing tendency in the number of NLP-related scientific papers from 1999 to 2021 (Chen et al., 2022). The study also identified three stages of development in NLP research. From 1999 to 2005, there was a slow growth tendency. From 2006 to 2016, a steady growth tendency was witnessed. From 2017 to 2021, there was a fast growth in the number of NLP papers (Chen et al., 2022). Furthermore, the study examined the funding information in the Acknowledgements section of each NLP paper and identified the top seven agencies that provided the most grants for supporting NLP research. These agencies included the National Natural Science Foundation of China, the United States Department of Health and Human Service, the National Institutes of Health, the European Commission, Fundamental Research Funds for the Central Universities, the National Key Research and Development Program of China, and National Science Foundation (Chen et al., 2022).

2.2.1. NLP Market Size Worldwide

Considering the evolution of NLP and the growing academic attention it has received; it is imperative to compare this progression with real-world market dynamics. The research landscape and technological breakthroughs often serve as precursors to market growth and application viability. This is seen in current market insight forecasts from August 2023 by Statista (2023).

The market size of the NLP market, as illustrated in Figure 1 from Statista Market Insights Forecast (Statista, 2023), was estimated at €22.24bn in 2023. The market size is expected to also show a compound annual growth rate (CAGR) of 15.47 %, resulting in a market volume of €60.89bn by 2030. This projection is driven by several factors, including the increasing demand for NLP solutions in the enterprise sector, the growing popularity of chatbots and virtual assistants, and the growing need to extract insights from unstructured data. This projected market trajectory of NLP not only validates academic endeavors but also offers a pragmatic lens through which we can assess the tangible impact of these innovations as NLP continuously evolves.

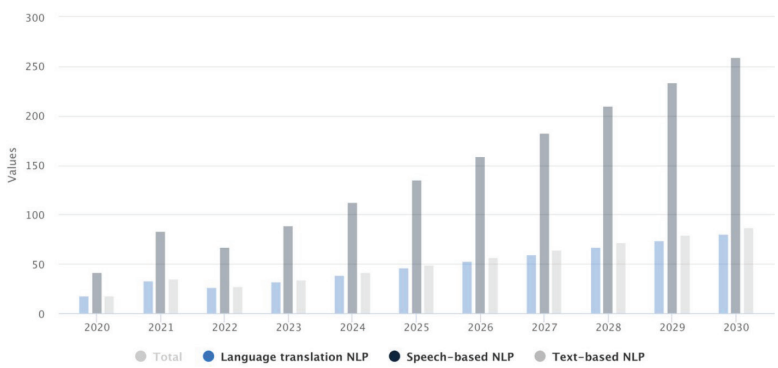

Figure 1. Natural language processing worldwide market size (Statista, 2023)

2.3. Large Language Models

2.3.1. Emergence of Large Language Models

In recent years, notable paradigm shifts in AI, particularly in NLP, have been marked by the evolution of LLMs. LLMs have significantly impacted diverse domains, from language translation to code generation and drug discovery, owing to their sophisticated architectures, extensive parameters, and rigorous training methodologies. The historical transition from rule-based or statistical methods to data-driven approaches, which leverage neural networks and deep learning techniques, reflects the growth of computational capacities.

The emergence of language models is discussed in the "Linguistic Data Model for Natural Languages and Artificial Intelligence. Part 4. Language" journal (Kim, 2010) attributing it to the development of a world model that necessitates language in an open form. Dynamic knowledge graphs, as highlighted by Kingma and Ba (2014), contribute to enhancing language models' comprehension of textual data, creating a dynamic knowledge base that can differentiate between facts and stories. Koubaa et al. (2023) emphasize that LLMs exhibit not only excellent few-shot learning with task-specific exemplars but also decent zero-shot reasoning, showcasing their capacity for complex reasoning even in novel situations.

The study of Kovan and Márta (2023) delves into human interactions with LLMs, revealing participants' strategies, such as explicit reasoning requests and pursuing multiple approaches for each question. These studies collectively underscore the diverse capabilities of LLMs in language understanding, reasoning, scalable oversight, and human-machine interactions, advancing the field of natural language processing.

The true emergence of LLMs began with the realization that increasing model size and utilizing vast training data could enhance language understanding and generation tasks significantly. Symbolic models such as generative pre-trained transformer (GPT) by OpenAI, bidirectional encoder representations from transformers (BERT) from Google, and text-to-text transfer transformer (T5) mark this transformative era in AI. The field has rapidly advanced with the introduction of powerful models such as Google's language model for dialogue applications (LaMDA) and parallel language model (PaLM) 2, capable of generating more human-like text, accurate language translation, and comprehensive question answering. Table 1 shows the summary of major Large Language Models architectures.

Table 1. Key architectures and innovations of selected large language models

Model	Architecture	Innovations	Release Date
BERT	Transformer	Masked language modelling, next sentence prediction	October 2018
GPT 3 **GPT 3.5** **GPT 4**	Transformer	Self-attention, positional encoding, mixture of experts	June 2020 March 2022 March 2023
T5	Transformer	Text-to-text transfer learning, unified objective	October 2019
LaMDA	Transformer	Pathways, decoder-only training	May 2021
PaLM **PaLM 2**	Transformer	Pathways, decoder-only training, pathways with feedback	April 2022 May 2023

2.3.2. Key Architectures in LLMs

The cornerstone of most LLMs is their underlying architecture. The architecture defines how data is processed, relationships are modelled, and outputs are generated.

1. Transformer architecture: Introduced by Vaswani et al. (2017), the authors' transformer architecture has quickly become the gold standard for LLMs. Its self-attention mechanism allows models to dynamically weigh input tokens, making it particularly adept at capturing long-range dependencies in data. The authors demonstrated that this architecture significantly outperformed previous models in tasks such as machine translation (Vaswani et al., 2017).
2. RNNs: Before the rise of transformers, RNNs, especially their variants such as LSTM, were the preferred choice. LSTM was introduced as a solution to the vanishing gradient problem observed in traditional RNNs, making the training of deeper networks feasible.

Recent advances in LLM architectures include:

1. Sparse LLMs: Sparse LLMs have fewer parameters but can achieve comparable performance to dense LLMs. This is achieved by using techniques such as parameter pruning and knowledge distillation.
2. Transformer-based architectures for specific tasks: Transformer-based architectures have been extended to support a variety of tasks, such as text summarization and question answering. This has led to significant improvements in performance in these tasks.

2.3.3. LLMs Training and Parameters

LLM training is a complex interplay of data, optimization strategies, and regularization techniques.

- Data and curriculum learning: While most LLMs rely on vast web data, the sequence in which data is presented also impacts performance. Bengio et al. (2009) proposed a curriculum learning

strategy, where models are first exposed to simpler tasks before transitioning to complex ones, analogous to human learning.

- Optimization: Techniques such as the Adam optimizer have become standard in LLM training. Kingma and Ba (2014) introduced Adam, demonstrating its superiority over traditional methods such as stochastic gradient descent in terms of convergence speed and stability.
- Regularization and robustness: Regularization techniques prevent overfitting in large models. Srivastava et al. (2014) introduced dropout as a simple yet effective regularization method, randomly deactivating a subset of neurons during training to improve generalization. Moreover, the robustness of models, especially concerning adversarial attacks, has gained attention. In a pioneering study, Goodfellow et al. (2014) explored the susceptibility of neural networks to adversarial examples, sparking a plethora of research on enhancing model robustness.

Recent advances in LLM training dynamics:

- Self-supervised learning: Self-supervised learning techniques have enabled LLMs to be trained on large amounts of unlabeled data. This has led to significant improvements in performance in a variety of tasks.
- Adversarial training: Adversarial training is a technique that can be used to improve the robustness of LLMs to adversarial attacks. This is achieved by training the LLM to distinguish between real and adversarial examples.

Parameters are the heart and soul of neural networks, storing the learned knowledge and shaping the model's performance. Recent advances in LLM parameter scaling:

- Sparse LLMs: Sparse LLMs have fewer parameters but can achieve comparable performance to dense LLMs. This is achieved by using techniques such as parameter pruning and knowledge distillation.

Efficient training algorithms: New training algorithms have been developed that can train LLMs with trillions of parameters in a reasonable amount of time.

2.4. Introduction to Chatbots

Chatbots are commonly internet-based computer programs, specifically designed to simulate conversations with human users. These conversational agents leverage advanced technologies such as NLP and AI to comprehend and respond to user inputs in a manner that mimics human interaction. The fundamental objective of chatbots is to foster meaningful conversations by furnishing information, addressing queries, or aiding in specific tasks. The key characteristics of chatbots include:

- Simulation of human conversation in understanding and responding to natural language queries, thus maintaining conversations with users over multiple turns.
- They are powered by AI and NLP to understand the intent of user queries and generate appropriate responses. This allows them to be more flexible and responsive than traditional chatbots that are based on rule-based systems.

In recent years, chatbots have emerged as valuable tools in the realm of artificial intelligence, revolutionizing the way humans interact with technology. Chatbots, also known as conversational agents or virtual assistants, are software applications that utilize NLP and machine learning algorithms to simulate human-like conversations, enabling users to interact with systems, services, or information through textual or auditory methods (Gupta et al., n.d.). They serve a diverse range of purposes, from performing specific tasks, such as booking tickets or handling customer queries, to providing entertainment, helping in healthcare, and more (Kovan and Márta, 2023).

The utilization of chatbots has garnered attention across various domains, including education, healthcare, business, and customer service, owing to their potential to enhance user experiences, improve operational efficiency, and provide personalized support (Kovan and Márta, 2023). This surge in chatbot applications can be attributed to advancements in natural language processing and the increasing demand for seamless, interactive digital experiences. Figure 2 shows the use of the chatbots and their performance across various sectors from 2014 to 2025.

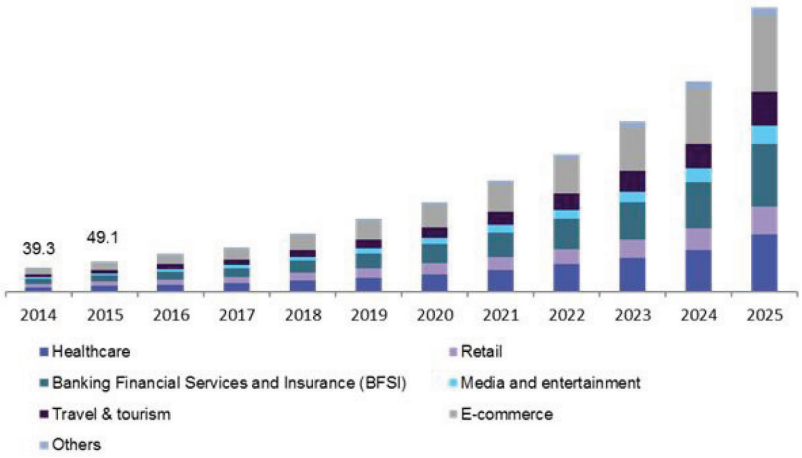

Figure 2. Evolution of chatbots and their performance (DivyaSingh456, 2019)

Despite their growing prominence, the effectiveness of chatbots is a subject of ongoing research and discussion. Studies have examined THE social and emotional implications of interacting with chatbots, exploring whether users can derive genuine feelings of connection and happiness from engagement with these virtual agents (Psyarxiv Manuscript, n.d.).

The potential for chatbots to offer social connection and companionship has been of particular interest in the context of human-computer interaction, raising questions about the extent to which users can establish meaningful relationships with these artificial entities (Psyarxiv Manuscript, n.d.). The development of chatbots involves the use of various platforms and APIs, such as Chatfuel and Facebook Messenger, which provide modern development, integration, and communication methods for creating interactive conversational interfaces (Kovan and Márta, 2023). Additionally, generative AI chatbots such as ChatGPT, Bard and Bing chart are dominating the market and are widely accessible.

2.4.1. Types of Chatbots

Chatbots can be broadly classified into five main types as seen in Table 2.

Table 2. Types of chatbots

Chatbot	Characteristics
Rule-based	The simplest type of chatbot is based on a set of rules
Keyword based	More advanced than rule-based, can recognize key words in user input
AI-powered	Uses AI to understand user intent and generate responses
Hybrid	Combines the strengths of rule-based, keyword-based and AI-powered chatbots
Conversational AI	The most advanced type of chatbot is used for tasks that require a high level of NLP.

2.4.2. Applications and Use-Cases of Chatbots

Chatbots have emerged as versatile tools with a wide-ranging impact across various industries and domains. Their inherent ability to simulate human conversation and process natural language has propelled their adoption in diverse applications, redefining the way we interact with technology including customer service, e-commerce, healthcare, finance, education, human resources, travel and hospitality.

3. Literature Review of Three Conversational AI Models

3.1. ChatGPT: A Comprehensive Overview

3.1.1. Large Language Models

Open AI developed ChatGPT, a model that leads the market in the areas of NLP and conversational AI. OpenAI has released this model as a branch of the GPT model family. ChatGPT was specifically developed to ensure the output is a human-like conversational response to a prompt rather than just filled in text. It is a stunning new step in the evolution of language models, revealing the immense power of planning for language nuances, context, and coherence. ChatGPT's core philosophy is based on the advancements in transformer architectures. We know that transformers capture dependencies and contextual information within sequential data of great length quite well. Therefore, it becomes possible for ChatGPT to understand and formulate responses in a conversation which makes it applicable in a large number of areas ranging from chatbots to content creation and others. ChatGPT has been profoundly developed by the evolution of GPT models. And as we explore its development it is impossible to ignore the fact that language models have undergone rigorous refining processes over the years. With each new version of GPT, OpenAI has demonstrated its dedication to expanding the frontier of language understanding, and each newer model attempts to build upon the successes and address the limitations of its predecessors. This iterative approach has played a pivotal role in shaping the capabilities of ChatGPT.

In terms of sheer technical architecture, ChatGPT inherits the fundamental architecture of GPT models, which relies on a transformer-based neural network. The transformer architecture, introduced by Vaswani et al. (2017) has become a cornerstone in NLP, facilitating parallel processing of input

sequences and enabling the model to attend to different parts of the input text simultaneously. This architecture, which is explored further in subsequent sections, has proven to be highly effective in capturing intricate linguistic patterns and dependencies.

3.1.2. Generative Pre-Trained Transformer Models

Natural language processing (NLP) has undergone a paradigm shift with the introduction of generative pre-trained transformer (GPT) models, which are distinguished by their capacity to learn and produce coherent, contextually relevant text. The foundation of GPT models, such as ChatGPT, is the transformer architecture, which was first presented by Vaswani et al. (2017). An important breakthrough in natural language processing (NLP) is the transformer design, which allows the model to use self-attention processes to capture contextual information and long-range dependencies.

Generative pre-trained transformer (GPT) models are a series of language models developed by OpenAI, specifically designed for natural language processing (NLP) tasks (Ahmed et al., n.d.; Ray, 2023; Koubaa et al., 2023). The importance of these models in AI increased primarily due to their impact on NLP in the recent past. Additionally, these models have had a ripple effect on the deep learning research community. The GPT models are based on the transformer architecture, which is a disruptive technology in sequence to sequence modeling, as it outperforms previous models based on RNNs or CNNs. The transformer architecture has a self-attention mechanism in which the model learns to attend to the specific parts of the variable length input sequence, which enables the model to capture long range dependencies. It helps the models to manage and analyze the more complex structures of a language and to perform a number of other related tasks such as understanding and generation of language. Additionally, the transformer architecture makes models more efficient in NLP applications.

First, these models are pre-trained for specific tasks, such as goal oriented conversational agents, where the model is required to understand the language and generalize its rules: syntax, grammar, and semantics. After pre-training, the model is prepared for a specific task and then modified with task specific training in order to change the weights and biases of the model to suit the new specific case and application.

This idea about pre-training on large and diverse datasets is key to the procedure followed by GPT. This stage is called pre-training and consists of teaching the model its language, syntax, and how to make semantic connections by providing it with a vast corpus of texts collected from the Internet and other sources. There is also a specific fine-tuning stage later when the model is further trained for specific tasks such as language translation, sentiment detection, or even conversation generation, similarly to ChatGPT. Another unique aspect of GPT models, including ChatGPT, is their autoregressive feature during text generation. Autoregressive models are designed to generate outputs sequentially in the form of tokens. Each token is generated based on the previous one. The transformer's attention mechanism allows the model to generate text in a coherent manner. During this process, context is preserved, and information remains relevant to the topic. One of the advantageous features that the GPT model has is its parameters. Models with millions and billions of parameters are classified as large-scale architectures which in turn define the GPT model. The enormous size of the model allows it to effortlessly identify details and complex patterns from various types of language data, making it extremely practical.

ChatGPT was made possible by all the developments and emerging trends present in the previous versions of the GPT series. The predecessor of ChatGPT, GPT-2, also significantly evolved by enhancing the structure and increasing the scale of the model. These changes helped in improving language

generation and interpretation. Consequently, ChatGPT expands on this basis by improving the model's conversational skills and resolving issues encountered in previous versions.

When combined with the autoregressive generation process, the pre-training and fine-tuning method presents GPT models as adaptable instruments that can perform well in a variety of NLP applications. This adaptability also applies to ChatGPT, where the model stands out among AI-driven chatbots due to its ability to produce contextually relevant responses in conversational scenarios.

3.1.3. Technical Architecture and Training of ChatGPT

The architecture and training of ChatGPT are rooted in advanced technologies such as deep learning, unsupervised learning, multi-task learning, in-context learning, reinforcement learning from human feedback, and the integration of large language models. As detailed by Wu et al. (2023), ChatGPT is an amalgamation of these technologies, with iterative improvements leading to the development of GPT-1 to GPT-4.

The initial GPT-1, developed in 2018, laid the groundwork for training a generative language model based on the transformer framework through unsupervised learning. Similarly to InstructGPT, ChatGPT was trained via reinforcement learning from human feedback (RLHF). Supervised fine-tuning was used to train the first model. AI trainers who were human, acted as both the user and the AI assistant in chats. To assist in crafting answers, trainers were provided with access to sample written suggestions. The InstructGPT dataset, which we converted into a conversation format, was then combined with a brand-new dialogue dataset.

Subsequently, the GPT-2, introduced in 2019, expanded its capabilities with more network parameters and data for training, enabling the model to generalize to most supervised subtasks without further fine-tuning. The evolutionary leap to GPT-3 in 2020 marked the incorporation of meta-learning and in-context learning, greatly enhancing the model's generalization ability across various downstream tasks. GPT-3 also surpassed existing methods on various benchmarks and became the first language model to surpass a parameter scale of 100 billion. GPT-4, an advanced multimodal model capable of processing image and text inputs and emitting text outputs, exhibits human-level performance on professional and academic benchmarks (Wu et al., 2023). The enlargement of model capacity and the volume of data for pre-training were instrumental in enhancing the model's comprehension of textual content.

Comparative data, which rated the quality of two or more model replies, needed to be gathered to build a reward model for reinforcement learning. During training, recordings of chatbot chats conducted by AI trainers were saved to gather data. Figure 3 shows the training steps of the Open AI chatGPT model.

The parameters and data used in the GPT series models have significantly increased from GPT-1 to GPT-3, culminating in GPT-3 with 175 billion parameters and 45 TB of data. The increase in model capacity allows the model to better understand the meaning and intent behind a given text, as discussed in Wu et al. (2023). The training of ChatGPT also entails the integration of reinforcement learning from human feedback to incrementally train the model, improving its ability to align with users' intent.

3.2. Google Bard: A Comprehensive Overview

3.2.1. Introduction to Google Bard

Created by the esteemed Google AI team in 2023, Google Bard is a noteworthy advancement in the field of large language models. This innovative artificial intelligence (AI) system is evidence of the

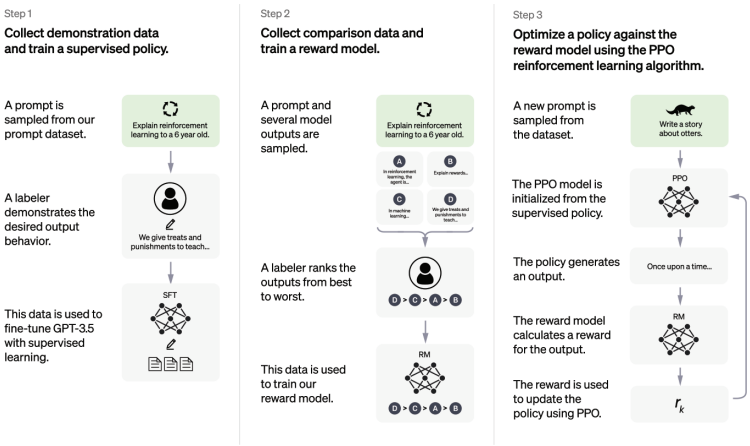

Figure 3. Steps in training ChatGPT (OpenAI, 2022)

progress made in both artificial intelligence (AI) and natural language processing (NLP). In contrast to conventional chatbots that depend on pattern matching and preset responses, Google Bard's capabilities go above and beyond what is often expected. It can have in-depth and educational dialogues and produce smoothly flowing, coherent prose.

LaMDA demonstrated remarkable progress in its ability to understand and respond to natural language prompts. However, the quest for even more sophisticated language processing capabilities led to the development of PaLM 2, a successor that surpassed LaMDA's achievements. PaLM 2's prowess in understanding and generating text, coupled with its ability to perform complex tasks such as coding and scientific research, paved the way for the creation of Google Bard.

Google Bard's roots can be traced back to LaMDA, the first generation of Google's LLM, which was unveiled in 2022. This was highlighted by Ahmed et al. in the study «ChatGPT vs. Bard: A Comparative Study» (Ahmed et al., n.d.), as well as in Borji et al.'s «The AI Race is On! Google's Bard and OpenAI's ChatGPT Head to Head: An Opinion Article» (Borji and Mohammadian, 2023), which discusses the emergence of Google Bard in the context of the rapidly evolving field of NLP and conversational AI. The opinion article also draws attention to Google's unexpected announcement of Bard's release, which came just after OpenAI's ChatGPT launch. Google presented Bard as their first attempt at releasing chat-based generative language models for public use. The paper highlights Bard's potential as a strong competitor in the area of AI and highlights its access to large swaths of the internet, laying the groundwork for a competitive dynamic with current models, most notably ChatGPT. Moreover, the introduction highlights the strategic reasons for Google's release of Bard, speculating that the popularity and success of ChatGPT may have had a role. This creates the conditions for a competitive environment driven by technological developments and innovation, adding to the dynamic character of AI chatbots and generative language models.

3.2.2. PaLM 2: The Next Generation of Google's Large Language Model

PaLM 2 is a revolutionary LLM developed by Google AI. Unveiled in October 2023, PaLM 2 represents a significant advancement in the field of AI, pushing the boundaries of what is possible for

LLMs. It surpasses its predecessors, PaLM and LaMDA, in every aspect, demonstrating remarkable capabilities in a wide range of tasks.

The technical report on PaLM 2 by Anil et al. (2023), presented a brand-new, cutting-edge language model that outperforms PaLM in terms of computation efficiency while displaying increased multilingual and reasoning skills. PaLM 2 exhibits its better performance across a range of downstream activities and capacities thanks to its transformer design and combination of training objectives. Additionally, the report discussed the models' availability in small, medium, and large varieties. The transformer architecture's stacked layers are used, with the parameters changing according to the size of the model. The study highlights the model's capacity to accept text as input and produce text as output, even if the specifics of the model's design and size are kept from external disclosure.

Designed for accelerating research on language models and serving as a building block in various Google products, PaLM 2 offers potential for use in experimental applications such as Bard and Magi. Through MakerSuite and APIs, PaLM 2 is also made available to other developers, along with further procedural and technological safety measures pertaining to policy. Reviewing guidelines and resources for responsible development is advised by the report, especially in light of potential risks and biases unique to downstream uses. Significant gains over PaLM on a range of reasoning tests show that PaLM 2 has strong reasoning capabilities. Additionally, the model allows for inference-time management of toxicity without causing extra expense or compromising other capabilities, and it displays consistent performance across a range of responsible AI evaluations. According to the technical assessment, PaLM 2 performs at the cutting edge across a wide range of activities and functionalities. The model's strong language competency across tested languages, which showcases its language creation, translation, and reasoning capabilities, demonstrates its multilingual competence. Numerous statistics in the technical study demonstrate the capabilities of PaLM 2, including multilingual competency, improved translation capabilities, multilingual creative production, and cross-programming language coding abilities. Figure 4 shows the performance comparison of PaLM 2 and PALM in a language proficiency test

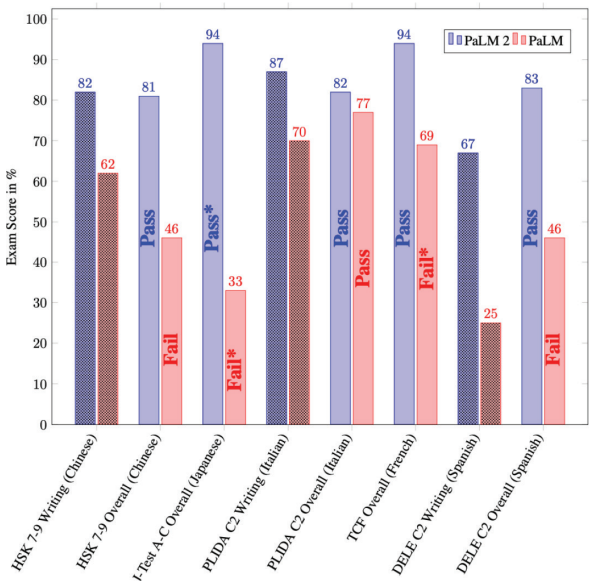

Figure 4. Performance of PaLM 2 and PALM in a language proficiency test (Anil et al., 2023)

3.2.3. Dataset Training and Evaluation

The technical report by Anil et al. (2023) also discusses the dataset training and evaluation of processes. The pre-training corpus for PaLM 2 was compiled from a variety of sources, including books, web pages, code, mathematical data, and conversational data. Compared to the corpus used to train PaLM, the pre-training corpus was noticeably larger (Chowdhery et al., 2022). Because PaLM 2 has been exposed to a greater range of languages and cultures, it is more suited for multilingual tasks such as translation and multilingual question answering. This is because PaLM 2 was trained on a dataset that has a higher percentage of non-English data than prior large language models. This lets the model notice the subtleties of every language.

PaLM 2 was trained on parallel data comprising hundreds of languages, which consisted of source and target text pairings with one side in English, in addition to non-English monolingual data. The model's comprehension and production of multilingual text are further enhanced by the addition of parallel multilingual data. Additionally, it imbues the model with an innate capacity for translation, which is advantageous for a variety of applications.

Numerous techniques for data cleaning and quality screening were used, such as de-duplication, sensitive-PII removal, and filtering. Tokens denoting text toxicity were incorporated into a tiny portion of the pre-training data. As an inference time control technique, evaluation is accomplished by conditioning the effectiveness of control tokens. Crucially, the assessments show that control tokens have no detrimental effect on performance in unrelated tasks.

The model's context length was trained to be substantially longer than PaLM 2's. This enhancement is essential for enabling activities that call for the model to consider numerous contexts, such as lengthy dialogues, long-range reasoning and comprehension, summarization, and other similar tasks.

3.3. Microsoft Bing Chat: A Comprehensive Overview

3.3.1. Introduction to Microsoft Bing Chat

Microsoft Bing Chat, initially launched as Bing Chat in February 2023, stands as a groundbreaking AI tool embedded within the Microsoft Bing search engine and the Microsoft Edge web browser. This sophisticated technology, currently powered by GPT-4 LLM, surpasses conventional chatbots by enabling natural and informative interactions with users.

It is integrated into Microsoft's search engine, Bing and designed to engage users in natural language conversations and provide information and assistance through text or voice interactions. As a part of the larger Bing search ecosystem, Bing Chat not only functions as a chatbot but also serves as a search engine, incorporating AI capabilities for conversation and information retrieval. When the chatbot is prompted with a question or statement, it performs the following actions to provide a response:

1. Interprets the prompt
2. Searches the knowledge base
3. Crafts the answer
4. Delivers the response

The capabilities of Microsoft Bing Chat go beyond just having natural language chats. It has several characteristics that make it a useful tool for a variety of tasks. It also offers thorough and educational answers to a broad range of queries, indicating its in-depth knowledge of the world. It translates across languages with amazing accuracy and fluidity. In addition to being proficient in creating a wide range of artistic text formats, Microsoft Bing Chat can also generate code in several computer languages, including

Python, Java, and C++, as well as poems, scripts, musical compositions, emails, and letters. Additionally, Microsoft Bing Chat can produce various creative text formats of text content, such as poems, code, scripts, musical pieces, emails, letters, etc. It can also help users with research by summarizing factual topics.

3.3.2. Introduction to Microsoft Bing Chat

Bing Chat is powered by the three models below:

1. **ChatGPT-4 (OpenAI):** This version of Open AI's LLM is fast, optimized for search, and capable of producing imaginative writing in a broad range of formats.
2. **Prometheus model (Microsoft):** This is a proprietary model developed by Microsoft to interface with the OpenAI LLM. It is specifically made to improve the safety, relevance, and timeliness of AI-powered search results.
3. **DALL-E 3:** This is the text-to-image generative neural network from OpenAI. With DALL-E 3, AI-generated content transcends language boundaries and enters the visual domain.

4. Shared Technical Architectures

4.1. Transformer Architecture

Figure 5 highlights the input and output stages of the basic transformer architecture. The transformer neural network architecture, which transformed the field of NLP, is the foundation of these LLMs. This transformer architecture uses self-attention techniques to capture long-range relationships between words in a sequence, doing away with RNNs, which were the popular architecture at the time. Because of this change in architecture, LLMs can now comprehend and produce text more accurately and fluently.

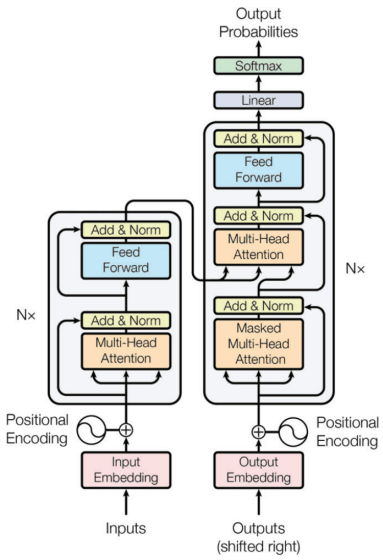

Figure 5. Transformer Architecture (Vaswani et al., 2017)

In summary, an input sequence is mapped by the encoder, located on the left half of the transformer architecture, to a sequence of continuous representations, which is subsequently supplied into a decoder. To produce an output sequence, the decoder, located on the right half of the architecture, gets both the encoder's and the decoder's output from the previous time step.

4.2. Encoder-Decoder Structure

Figures 6 and 7 show the basic encoder and decoder blocks for the transformer. ChatGPT, Google Bard, and Microsoft Bing Chat all utilize the encoder-decoder structure, a common architecture in NLP tasks that consists of two key components:

- **Encoder:** After processing the input text, the encoder creates a representation that accurately conveys the text's meaning and context.
- **Decoder**: This component creates the output text by using the encoder's representation to create a word sequence that represents the intended answer.

4.3. Attention Mechanism

The attention mechanism is a fundamental component of the transformer architecture, which has dramatically changed the landscape of natural language processing (NLP). Unlike the older recurrent neural networks (RNNs), transformers use self-attention to capture long-range relationships between words in a sequence. By surpassing the capability of RNNs, large language models (LLMs) are made more effective in generating and comprehending text due to this technology.

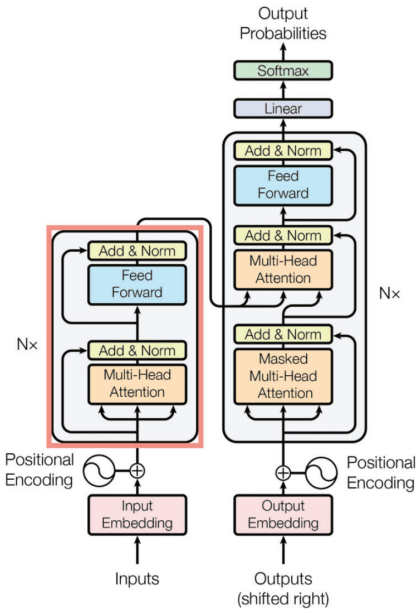

Figure 6. Encoder block of transformer architecture (Vaswani et al., 2017)

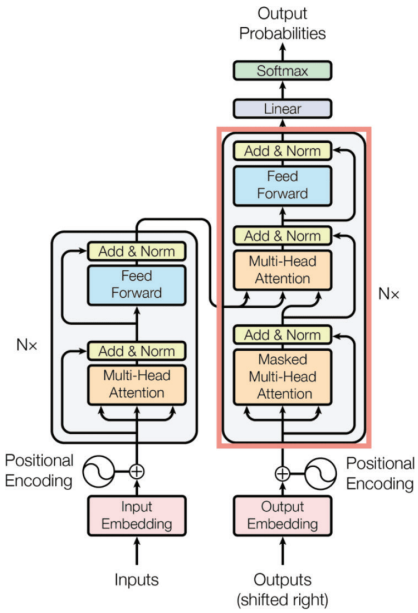

Figure 7. Decoder block of transformer architecture (Vaswani et al., 2017)

LLMs are capable of zeroing in on sections of the input text when generating output thanks to the attention mechanism. To illustrate, the model assigns each word a score based on how relevant it is in relation to the current task. Outputs with the highest scores are preserved while the rest are not as influential. This is one of many examples of how the model addresses the challenge of semantics and how words that follow each other are related. When an LLM translates an English sentence into French, it takes into account all the words in the English sentence so that it can accurately and natively convert it. By training the LLM using the attention mechanism, the model is able to concentrate on the important aspects of the English sentence while translating it into French.

4.4. Prompt Engineering

4.4.1. Theoretical Foundations of Prompt Engineering

Prompt engineering is deeply rooted in the theoretical foundations of machine learning, natural language processing (NLP), and reinforcement learning. These theoretical frameworks are essential for developing efficacious prompts and enhancing the efficiency of large language models (LLMs). The foundations listed below offer a framework for comprehending the ideas and methods that successful quick design is built upon.

1) **Knowledge representation** is a field of artificial intelligence, dedicated to creating computer-interpretable representations of abstract ideas and complex facts. When it comes to LLMs, knowledge representation is essential to the model's ability to comprehend and produce language. The prompts we give the model are essentially a way for us to represent our knowledge in a way that the model can understand.

2) **Language modelling** is a subfield of NLP that involves predicting the likelihood of a sequence of words. Large volumes of text data are used to train LLMs, such as ChatGPT, Google Bard, and Microsoft Bing Chat, so they can understand the statistical structure of language. This lets them produce language that seems human depending on the commands they get. Knowing the fundamentals of language modelling can assist us in creating prompts that correspond with the model's acquired patterns, leading to more precise and well-organized answers.
3) **Reinforcement learning** is a type of machine learning where an agent learns to make decisions by taking actions in an environment to maximize some notion of cumulative reward. Some LLMs adjust their models using reinforcement learning based on user feedback. This entails directing the model's learning process using prompts and matching desired replies. We can create efficient incentive systems and training schedules for rapid engineering by having a solid understanding of the fundamentals of reinforcement learning.

Prompt engineering is directly related to LLM capabilities and cannot be done in isolation. LLMs can process more complicated instructions and produce more nuanced responses as they advance in sophistication. On the other hand, skillfully written prompts can stretch LLMs' abilities, allowing them to take on tasks that are progressively more difficult and provide outputs that are more imaginative and educational. The interaction between LLM capabilities and prompt engineering is what drives the ongoing development of language processing technology. As a result, the ideas of language modelling, reinforcement learning, and knowledge representation are all woven into the theoretical underpinnings of rapid engineering. These ideas help us predict how LLMs will react to various prompts and modify our prompt engineering techniques accordingly. This theoretical understanding, combined with practical experience, forms the bedrock of effective prompt engineering.

4.4.2. Types of Prompts

Prompts can be broadly categorized into three main types:

- Open-ended prompts: These prompts offer less direction, giving LLMs greater creative latitude in producing the output. They are frequently employed in activities such as creative text generation, where the objective is to create unique and creative content.
- Closed-ended prompts: These prompts give the LLM more precise instructions on what to generate by outlining the intended result in greater detail. They are frequently employed for jobs such as answering questions, where the objective is to give a succinct and accurate response.

Specific instruction prompts: The format, structure, and style requirements are all included in these prompts, along with comprehensive instructions on how to generate the output. They are frequently employed in jobs such as code creation, where the objective is to generate code that complies with predetermined guidelines.

4.4.3. Factors Influencing Prompt Engineering Effectiveness

The effectiveness of a prompt in guiding an LLM to the desired output depends on several factors. These factors can be further grouped into several major factors which are all crucial in measuring the effectiveness of the prompt.

1) **Clarity and Specificity:** When a task or a query is given to the LLM in a precise manner, it eliminates a lot of uncertainty and ensures that the model will reach the end goal. Never use ambiguous or hard to define terms in attempt to prevent a range of interpretations. A case in point is "What is the weather today in Berlin?" can be improved to "When does winter typically start in Germany?"
2) **Contextual Relevance:** A context accurate and to the point answer provides the LLM with the background information necessary for creating a good answer. Consider the topic and the context of the dialogue and add useful information to the prompt. For example, when a specific question concerning a particular place in history is asked, provide the relevant time and space to help the LLM narrow the scope of its knowledge search.
3) **Language Style:** Alter the style of speech in the prompt to achieve the desired outcome. For example, engage in more imaginative and intensely descriptive speech which incorporates different figures of speech and stirs emotions if the aim is to generate more text types. In contrast, utilize more succinct and objective prompts for more formal tasks such as answering questions or creating code.
4) **Length and Complexity:** Strive to make the prompt sufficiently elaborate and brief by taking into account its length and complexity. Think through the task's nature and the capabilities of the LLM. For more complicated tasks, a longer prompt may be necessary in order to provide the LLM with adequate context. At the same time, having too much information may be a challenge for the model and lead to poorer quality outputs.

Taking these aspects into account and adjusting prompts to fit the particular task or question allows users to better these systems and maximize their capability to produce useful, creative, and informative outputs. As techniques in LLMs begin to improve, new prompt engineering methods will be developed in order to more efficiently guide large models towards the expected output.

4.4.4. Case Studies and Applications of Prompt Engineering

In the field of healthcare, prompt engineering has gained significance in the development of NLP models for medical applications. It has been instrumental in task guidance for question-answering systems, text summarization, and machine translation, and offers the potential to utilize large language models more effectively and efficiently in healthcare-related tasks (Wang et al., 2023). In the context of AI art, prompt engineering has become an essential skill for practitioners of text-to-image generation, offering an intuitive language-based interface to AI. It allows for the iterative and interactive creation of inputs (prompts) for generative models, enabling dialogue between humans and AI in an act of co-creation. It has been particularly useful for improving the quality of digital visual art, as seen in the development of image generation systems such as DALL-E, Midjourney, and Imagen, which have been trained on large collections of text and images. Prompt modifiers, such as specific keywords and phrases, contribute to the skillful application of prompt engineering to distinguish expert practitioners from novices in text-to-image generation (Oppenlaender et al., 2023).

"A Systematic Survey of Prompt Engineering on Vision-Language Foundation Models" by Gu et al. (2023) highlights the application of prompt engineering on three types of vision-language models: multimodal-to-text generation models, image-text matching models, and text-to-image generation models. The survey outlines various methodologies of prompt engineering on these models and discusses their applications, challenges, future directions, and research opportunities in this domain. Overall, prompt engineering offers versatile applications and has the potential to revolutionize various

domains, including healthcare, artistic endeavors, and computer vision, by providing targeted prompt information to guide model generation and task resolution.

5. Methodology

5.1. Scope of Analysis

In the pursuit of a comprehensive understanding of the capabilities and applications of the prominent conversational AI models, this study seeks to address the following scope:

1) Conversational Capabilities:
    a. How do Chat GPT, Google Bard, and Microsoft Bing Chat differ in their conversational capabilities?
    b. What metrics can be effectively employed to measure the quality and effectiveness of their respective conversational interactions?
2) Natural Language Processing Abilities:
    a. To what extent do Chat GPT, Google Bard, and Microsoft Bing Chat exhibit advanced natural language processing abilities?
    b. How do these models handle intricate sentence structures, contextual nuances, and diverse user intents?
3) Generating Creative Text Formats:
    a. In what ways do Chat GPT, Google Bard, and Microsoft Bing Chat demonstrate creativity in generating text formats?
    b. Can the outputs of these models be objectively evaluated for originality, coherence, and relevance across varied contexts?
4) Suitability for Specific Applications:
    a. How well-suited are Chat GPT, Google Bard, and Microsoft Bing Chat for specific applications, such as customer support, content creation, or educational purposes?
    b. Can discernible strengths and weaknesses be identified in each model, aiding in targeted application deployment?

This analysis scope has been meticulously crafted to guide the investigation into key aspects of the three chatbots under consideration. By focusing on conversational capabilities, natural language processing abilities, creative text generation, and suitability for specific applications, this study aims to provide a nuanced and insightful comparative analysis. The results of this study are anticipated to contribute valuable knowledge to the field of AI and inform stakeholders, researchers, and practitioners about the diverse functionalities of these ever-developing AI models in the field of technology. This methodological framework ensures a structured and systematic approach to the study, aligning with the academic standards of rigor and depth.

5.2. Test Questions Generation

In the context of this comparative analysis, prompt engineering will be used to ensure that all three models are evaluated on an equal footing. Specifically, the following steps will be taken:

1) Define a set of benchmark tasks: A set of benchmark tasks will be defined that span a range of LLM capabilities, including creative text generation, and question answering.

2) Design prompts for each task: Clear, specific, and contextually relevant prompts will be designed for each benchmark task. These prompts will be tailored to the capabilities of each model, ensuring that all three models are given a fair chance to succeed.
3) Evaluate the generated outputs: The generated outputs will be evaluated against a set of pre-defined criteria, such as accuracy, fluency, and creativity.

The use of prompt engineering in this comparative analysis will ensure that the results are fair and accurate, providing a valuable assessment of the relative strengths and weaknesses of ChatGPT, Google Bard, and Microsoft Bing Chat.

To provide a new dimension with respect to the scope of the analysis, questions for each of the categories for analysis were curated to facilitate in generation of test questions for the AI chatbots. The questions are well curated to ascertain the nuanced aspects of each model's capabilities within the specified dimensions such as Conversational Capabilities, Natural Language Processing Abilities and Generating Creative Text.

5.3. Model Versions and Technical Brief

For tests, the current base version of each chatbot was used. The model version and technical details have a significant impact on its performance. For example, chatbots with larger model sizes are typically able to generate more accurate and comprehensive responses. Similarly, chatbots with more advanced neural network architectures can better understand complex and nuanced languages. Consequentially, the training data size is also an important factor to consider. Larger and more diverse datasets will typically generate more informative and relevant responses.

It is important to note that the performance of these chatbots can also be affected by other factors, such as the quality of the code that implements the chatbot and the quality of the user interface. Additionally, the performance of a chatbot can vary depending on the specific task that it is being asked to perform.

The following table shows the model version, technical details, and training data size for each chatbot. This serves as the base model for operations for analyses. Table 3 represents the comparison of different models.

Table 3. Overview of model versions used for testing

Chatbot	Version-Release	Model Size	Training Data Size
ChatGPT	GPT -3.5	175 billion parameters	Not publicly disclosed
Google Bard	Current release (January 2023)	1.56 trillion parameters	1.3 trillion words
Microsoft Bing Chat	Not publicly disclosed	175 billion parameters	Not publicly disclosed

5.4. Evaluation Framework

A well-constructed evaluation framework plays a pivotal role in the assessment and comparison of advanced conversational AI models, exemplified by ChatGPT, Google Bard, and Microsoft Bing Chat. This framework serves as an indispensable tool for researchers and developers,

providing valuable insights into the strengths and weaknesses of these models and facilitating informed decisions for future enhancements. The comprehensive evaluation framework employed is designed to address three fundamental tasks: conversational capabilities, natural language processing (NLP) abilities, and creative text generation. Each of these tasks undergoes meticulous evaluation utilizing a predefined set of metrics, wherein scores ranging from excellent (10) to poor (1) are assigned. This structured approach enables quantitative assessment and seamless cross-chatbot comparisons.

5.4.1. Complexity Levels

The evaluation framework incorporates three levels of question complexity for each metric within the designated tasks: simple, medium, and complex. This ensures a nuanced evaluation that challenges the AI models across various cognitive loads, providing a comprehensive understanding of their capabilities. The levels are carefully curated to encompass a spectrum of difficulty, enabling a robust assessment of the models' performance.

5.4.2. Evaluation Tasks

5.4.2.1. Conversations Skills Task

This task evaluates the chatbot's proficiency in engaging users in natural and meaningful conversations. The following metrics were utilized. Table 4 shows the evaluation criteria used for testing the models.

Table 4. Evaluation metrics for the conversational skills task

Metric	Description
Fluency	This metric assesses the grammatical correctness and ease of reading of the chatbot's responses
Coherence	This metric assesses the logical connection between the chatbot's responses
Relevance	This metric assesses the chatbot's ability to stay on topic and avoid veering off on unrelated tangents

5.4.2.2. NLP Abilities Task

This task gauges the chatbot's capacity to comprehend user queries and provide accurate and comprehensive responses. Table 5 shows the metrics used for this evaluation include

Table 5. Evaluation metrics for NLP abilities task

Metric	Description
Accuracy	This metric assesses the chatbot's ability to correctly understand user queries and generate accurate responses
Completeness	This metric assesses the chatbot's ability to provide comprehensive and well-rounded responses to user queries
Informativeness	This metric assesses the chatbot's ability to provide factual information that is relevant to user queries

5.4.2.3. Creative Text Generation Task

This task focuses on the chatbot's capability to generate creative text formats, such as poems, code, scripts, musical pieces, emails, letters, etc. Table 6 shows the evaluation metrics for this task include

Table 6. Evaluation metrics for the creative text generation task

Metric	Description
Fluency	This metric assesses the grammatical correctness and ease of reading of the chatbot's generated text
Creativity	This metric assesses the originality and inventiveness of the chatbot's generated text
Relevance	This metric assesses the chatbot's ability to generate text that is relevant to the user's input and the overall topic of the conversation

The overall performance of the chatbot is determined by calculating the average score across all three tasks. A high overall score signifies excellent performance in conversational capabilities, NLP abilities, and creative text generation. This systematic evaluation approach provides valuable insights into the chatbot's strengths and weaknesses, empowering researchers and developers to make informed decisions for further improvements.

By employing this robust evaluation framework, stakeholders can gain a comprehensive understanding of their chatbot's capabilities and iterate on its design and functionality, thereby advancing the state-of-the-art in conversational AI.

5.5. Evaluation Questions

This section introduces the set of carefully crafted evaluation questions designed to assess the performance of advanced conversational AI models, exemplified by ChatGPT, Google Bard, and Microsoft Bing Chat. These questions are tailored to evaluate the models across three fundamental tasks: conversational capabilities, natural language processing (NLP) abilities, and creative text generation as already discussed, with varying levels of complexity, ensuring a comprehensive evaluation that challenges the AI models across different cognitive loads. Table 7 shows the evaluations metrics and the level of the task solved for the conversational tasks. Similarly, Table 8 shows for the NLP tasks and Table 9 for the text generation tasks.

Table 7. Evaluation questions for the conversational task

Metrics	Simple	Medium	Complex
Fluency	Briefly explain the concept of cloud computing in a way that is easy for a non-technical person to understand	What is the impact of artificial intelligence on job markets? Be as brief as possible	Briefly explain the ethical considerations of using AI in healthcare
Coherence	Describe the process of photosynthesis	Tell a very short story that involves characters from different cultural backgrounds	Briefly explain the implications of climate change
Relevance	What are the benefits of regular exercise?	Briefly discuss the challenges faced by the education system today	What is the future of space exploration?

Table 8. Evaluation questions for NLP abilities task

Metrics	Simple	Medium	Complex
Accuracy	Define the term 'artificial intelligence' in a way that accurately captures its meaning	With the current trends in technology, are we on the verge of inventing a toaster that not only toasts bread but also delivers inspirational quotes on the meaning of life?	Briefly explain the concept of quantum computing
Completeness	What are the basic principles of Newtonian physics?	What are the latest developments in renewable energy?	What are the ethical considerations in advanced genetic engineering?
Informativeness	What is generative AI?	What is the current state of affairs in Israel and Palestine?	What is your view on using electric tanks in wars?

Table 9. Evaluation questions for the creative text generation task

Metrics	Simple	Medium	Complex
Fluency	Generate a grammatically correct poem about the beauty of nature	Craft a brief narrative script for a short film that maintains a consistent and engaging tone throughout	Incorporate metaphors and similes to enhance the quality of a short description of exploring the concept of consciousness in machines
Creativity	Create an original and creative code snippet for list sorting in Kotlin	Write a humorous paragraph	Generate truly innovative and groundbreaking creative text that envisions the future of human-computer symbiosis
Relevance	Generate an email response relevant to an inquiry about a new product launch	Write a short story about futuristic technologies	What is the future of autonomous vehicles by 2025?

After obtaining responses from the AI models, the next step involved ranking these responses. A selection of 20 users evaluated the responses on a scale of 1 to 10, according to the established evaluation framework.

6. Results

6.1. Conversational Skills Results

The evaluation of conversational capabilities involved assessing three key metrics: Fluency, Coherence, and Relevance. The results, based on user responses (average scores from 20 users who ranked chatbot responses on a scale of 1-10 as per the evaluation framework), are presented in Table 10:

Table 10. Conversational skills results

Chatbot	Fluency (Avg)	Coherence (Avg)	Relevance (Avg)	Overall (Avg)
ChatGPT	8.8	8.0	7.7	8.2
Google Bard	8.7	8.2	9.0	8.6
Microsoft Bing Chat	8.5	8.0	8.0	8.2

These scores represent the average ratings across simple, medium, and complex conversational skills tasks. Google Bard emerged as a top performer, excelling in relevance across various complexities, while ChatGPT showcased outstanding fluency. The overall average score indicates that users found Google Bard to be the most well-rounded in conversational capabilities, while both ChatGPT and Microsoft Bing Chat were equally rounded behind Google Bard.

6.2. NLP Abilities Results

The evaluation of NLP abilities included assessing three key metrics: Accuracy, Completeness, and Informativeness. The results, based on the user responses, are presented in Table 11:

Table 11. NLP abilities results

Chatbot	Accuracy (Avg)	Completeness (Avg)	Informativeness (Avg)	Overall (Avg)
ChatGPT	9.7	5.0	8.2	7.6
Google Bard	9.8	6.2	9.3	8.3
Microsoft Bing Chat	9.2	7.3	8.5	8.3

Google Bard demonstrated superior performance in accuracy and informativeness, and lagged behind Microsoft Bing Chat only in completeness. ChatGPT also performed well in accuracy but very poorly in completeness. Microsoft Bing Chat showcased impressive performance, particularly in accuracy, achieving an overall average score equal to Google Bard. These scores provide valuable insights into the nuanced strengths of each chatbot in natural language processing tasks.

6.3. Creative Text Generation Results

The evaluation of creative text generation involved assessing three key metrics: Fluency, Creativity, and Relevance. The results, based on user responses, are presented in Table 12:

Google Bard consistently outperformed in creative text generation, displaying high scores in creativity and relevance. ChatGPT showcased excellent fluency, earning an overall average score of 9.0. Microsoft Bing Chat demonstrated commendable performance, achieving an overall average score of 8.5. These scores provide valuable insights into the creative capabilities of each chatbot, with Google Bard standing out as a top performer.

Table 12. Creative text generations results

Chatbot	Fluency (Avg)	Creativity (Avg)	Relevance (Avg)	Overall (Avg)
ChatGPT	9.7	8.5	9.0	9.0
Google Bard	9.3	9.2	9.2	9.2
Microsoft Bing Chat	8.8	8.5	8.2	8.5

6.4. Overall Performance Results

To provide a comprehensive overview of the overall performance, the average scores across all tasks were calculated. The results, based on user responses, are presented in Table 13:

Table 13. Overall performance results

Chatbot	Overall Performance (Avg)
ChatGPT	8.3
Google Bard	8.7
Microsoft Bing Chat	8.3

Google Bard emerges as the top-performing chatbot with an overall average score of 8.7, showcasing its prowess in conversational capabilities, NLP abilities, and creative text generation. ChatGPT and Microsoft Bing Chat demonstrated commendable performance but fell slightly behind Google Bard. These overall performance scores provide a consolidated view of each chatbot's capabilities, guiding insights into their strengths and areas for further enhancement.

The evaluation results were derived from the responses of 20 users, who rated the outputs of the AI models on a scale of 1 to 10 based on the established evaluation framework. While this provides initial insights into the models' comparative performance, we acknowledge that a larger sample size would yield more statistically robust conclusions. Future studies will expand the participant base to enhance the generalizability and reliability of the findings.

7. Discussion

7.1. In-Depth Results

7.1.1. Conversation Skills

The evaluation of conversational capabilities unveiled intricate insights into three pivotal dimensions: Fluency, Coherence, and Relevance. These metrics, critical for assessing the quality of chatbot responses, were scrutinized based on average scores derived from user responses employing a 1-10 ranking system.

Fluency, encompassing grammatical correctness and ease of comprehension, serves as a cornerstone for effective communication.

Google Bard emerged as the top performer in this evaluation, earning an impressive average score of 8.8 across simple, medium, and complex conversational tasks. The users' strong approval of responses that are simple and grammatically correct is strong. ChatGPT came second with an average score of 8.7, which shows that it is capable of producing fluent and cohesive responses. Although Microsoft Bing Chat is good too, with an average score of 8.3, it was ranked lower than the other two in this metric.

In regards to coherence, which is the metric concerned with how connected and smooth the conversation is, Google Bard yet again came out on top with an average score of 8.2. It was, nonetheless, able to demonstrate strong logical coherence and flow even with complex and multi-faceted discussions. ChatGPT and Microsoft Bing Chat were both good too, with an average score of 8, reflecting their versatility and skill in handling different types of conversations.

Relevance measures the capacity of the chatbot to remain focused on topic, and as such, it points out some important differences between them. In contrast with the other two AI models that showed capability in remaining engaged with user prompts, Google Bard earned the highest ranking for its precision. Google Bard earned an average score of 9.0, and this is where it stood out most for remaining relevant. Users praised it for directly answering queries and remaining focused on the topic.

ChatGPT attained, on average, a relevance score of 7.7 which indicates one of the highest marks of relevance that the system could reach irrespective of slight divergence that was caused due to extra details. The relevance of Microsoft Bing Chat was good as well with an average score of 8.0 which conformed to user expectations. Bard was the best for conversational skills tasks due to its fluency, coherence, and relevance. Users liked the way the system self-generated and responded, which made interactions more enjoyable. ChatGPT also performed relatively well, with users reporting that the system's fluency and coherence was exceptionally good but irrelevant details were sometimes included. Microsoft Bing Chat, while performing admirably, showed slight room for improvement in coherence and maintaining relevance.

These results unveil nuanced disparities in how each chatbot engages in conversations, aligning with user preferences and expectations. The feedback from 20 users provides invaluable insights, guiding the trajectory of further enhancements in the development of conversational AI models.

7.1.2. NLP Abilities

This study evaluated the proficiency of three Chatbots—Google Bard, ChatGPT, and Microsoft Bing Chat—through three core metrics: accuracy, completeness, and informativeness. User responses shed light on the strengths of each bot as well as areas in need of improvement. A constantly prevailing component when evaluating the provided information was its accuracy, which tells a user whether the information is relevant and trustworthy in nature. For all three bots, this was quite strong, although Bard was ahead with an outstanding score of 9.8. In this domain, ChatGPT performed acceptably well with 9.7, and even Microsoft Bing Chat was able to score an average of 9.2—competent score, which further proved this chatbot's ability to provide reliable information.

When measuring the degree, strength, and detail of the answers given through the metric of completeness, there was a now slight variation in performance. It was, however, Microsoft Bing Chat that achieved the highest performance as it, with a notable difference, scored an average of 7.3. This means the answers given by the chatbot were relatively well developed and were able to cover most of what was asked for by the users. Bard's performance in scoring an average of 6.2 regarding the comprehension of answers signifies that it still has some gaps to fill. On the other hand, ChatGPT scored 5.0, demonstrating much weaker capabilities in this dimension. Interestingly, all chatbots tended to

get lower scores because they were too careful with delicate issues that, in turn, suggest they were programmed to avoid.

Moreover, informativity, which is a graded aspect concerning the quantity and quality of information being received, is an area where Google Bard stood out the most with an average score of 9.3. Microsoft Bing Chat came second, and ChatGPT's performance was just slightly lower with an average score of 8.2. All of these results illustrate that apart from the outstanding performance of Microsoft Bing Chat, Google Bard also managed to provide sufficient detail and information to the users.

All of these findings bring to light the differences in the capabilities of these chatbots when performing NLP tasks. In this regard, ChatGPT scored well in accuracy but struggled in completeness, revealing that it needs to improve in how thoroughly the prompts are used.

Microsoft Bing Chat stood out as one of the competitors in the same league as Google Bard, as it performed well in terms of completeness and informativeness. Microsoft Bing Chat compared and reported their superiorirty in overall performance in natural language processing, crafting a unique competitive standing alongside its contemporaries. This assessment focused on the strengths and weaknesses of each individual chatbot, as for these results, there is a set of areas to work on and further enhance the findings that will help inform further development. These advanced building blocks may in turn help advance chatbot technology and specialize in one particular insight.

7.1.3. Creative Text Generation

Creativity is defined in three categories: Fluency, Relevance, and Creativity. These categories serve as the basis for the user's qualitative feedback and clearly outline the parameters of viability for each of the chatbots. For the creative text, Google Bard was best at grammar and legibility with a score of over 9.3. Microsoft Bing Chat's scores for fluency were over average and were rated at 8.8, while ChatGPT was the best at over 9.7.

According to a surveyed creativity score, Google Bard Elite scored the highest with 9.2. ChatGPT captured an astonishing 8.5 on creativity, while Microsoft Bing Chat garnered an average score of 8.5 as well. Google Bard scored the highest in average relevance with 9.2 and ChatGPT showcased a high relevance with an average score of 9.0. Microsoft Bing Chat also demonstrated good relevance with an average score of 8.2. Summarizing the main points, Google Bard was superior in all aspects of creative text generation compared to the other models as indexed by the high scores in creativity and relevance. ChatGPT showcased excellent fluency and creativity, earning an overall average score of 9.0. Microsoft Bing Chat demonstrated commendable performance, achieving an overall average score of 8.5. These results highlight the creative capabilities of each chatbot, with Google Bard standing out as a top performer. These nuanced findings provide a comprehensive understanding of each chatbot's creative text generation capabilities, offering valuable insights into their strengths and areas for improvement.

8. Future Research and Limitations

This section highlights the possible future research and the limitations of the current study. We make the following suggestions for research in the future that can also address the existing limitations of this study.

As for future lines of research, we could look into and study in depth the more recent deep learning approaches that form the backbone of the recent conversational models for AI. One of the avenues

for exploration are the methods that facilitate continuous learning from user interactions. It should be examined how existing models can be enhanced in terms of their adaptability to the changing trends in language and user preferences. By looking at key points, especially the user related principles as well as the styles and preferences of each individual, we hope that studies in the future can positively impact the model development in line with user expectations.

Furthermore, in the future, the present analysis could be extended by evaluating performance in linguistic or cultural contexts, or both, across a diverse set of languages and cultures. This would aid in the creation of AI-based systems that are inclusive and globally applicable across a variety of regions. Another interesting direction is the exploration of ethical considerations in the design of AI-based models as conversational agents. In the future, we can focus on developing special guidelines and frameworks that ensure ethical practices. These can aid in addressing concerns for bias and privacy. Similarly, developing an understanding of how techniques such as fine tuning and transfer learning are applied to conversational AI models is of significant importance in achieving greater efficiency in some domain-specific applications. Other dimensions to be explored include multimodal capabilities that seamlessly integrate text and images.

Similarly, we noted several limitations of the present study. One limitation of this study is the relatively small sample size of 20 users for evaluating the chatbot responses. While their feedback offers valuable insights, the limited participant pool may not capture the full spectrum of user experiences and perspectives. Expanding the user base in future research will enable more diverse feedback and provide a stronger foundation for assessing the models' performance. Additionally, employing stratified sampling techniques to include users with varying backgrounds, expertise levels, and linguistic proficiencies can further enrich the analysis.

Incorporating these will contribute significantly to advancing the field of conversational AI. These future research avenues are poised to address current challenges, enhance user experiences, and ensure the responsible and ethical development of AI systems.

Despite their impressive capabilities, LLMs such as ChatGPT, Google Bard, and Microsoft Bing Chat face limitations, including:

- **Prompt Engineering Challenges:** Effectively crafting prompts that guide the LLM towards generating the desired output remains a challenge.
- **Bias Mitigation:** LLMs may inadvertently encode biases from their training data, which can lead to unfair or discriminatory outputs.
- **Transparency and Explainability:** Understanding the decision-making processes of LLMs remains a challenge, limiting their interpretability and trustworthiness.

9. Conclusion

We have comprehensively explored prompt engineering for large language models (LLMs) in this study, where a series of LLMs, namely ChatGPT, Google Bard, and Microsoft Bing Chat, were discussed in depth. In this study, we discussed the theoretical foundations of prompt engineering with a special focus on factors that influence the effectiveness of prompts. Additionally, we looked at various techniques and tools for prompt engineering. This study's main contribution is the detailed comparison between the three popular LLMs on multiple benchmarking tasks. We have therefore explored the strengths and weaknesses of these LLMs with respect to available prompts. In

agreement with older studies, we found that prompt engineering plays a central role in the effectiveness of utilization of LLMs. By writing clear, concise, contextually relevant, and specifically tailored prompts for the LLMs, a user can enhance the performance of a LLM significantly for various tasks. The comparison performed in this study revealed that all three LLMs performed well on the benchmark datasets and displayed their unique stamps in results, e.g., ChatGPT excelled in creative text generation, Google Bard excelled in question answers while Microsoft Bing Chat was superior for its conversational abilities. However, all three models showed that LLMs are susceptible to bias in the data, which becomes more apparent with an incorrectly designed or incomplete prompt. To conclude, we underscore the importance of prompt engineering for delivering better LLM performance in a variety of tasks. We also reiterate our belief that with the continuous evolution and improvements in the performance of LLMs, ultimately prompt engineering will gain central importance in unlocking the full potential of LLMs.

References

Ahmed, I., Kajol, M., Hasan, U., & Datta, P. P. (n.d.). ChatGPT *vs. Bard: A comparative study*.

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., Chu, E., Clark, J. H., Shafey, L. El, Huang, Y., Meier-Hellstern, K., Mishra, G., Moreira, E., Omernick, M., Robinson, K., … Wu, Y. (2023). *PaLM 2 technical report*. http://arxiv.org/abs/2305.10403

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. *Journal of the American Podiatry Association, 60*(6). https://doi.org/10.1145/1553374.1553380

Biswas, R., & De, S. (2022). A comparative study on improving word embeddings beyond Word2Vec and GloVe. *PDGC 2022 - 2022 7th International Conference on Parallel, Distributed and Grid Computing*. https://doi.org/10.1109/PDGC56933.2022.10053200

Borji, A., & Mohammadian, M. (2023). Battle of the wordsmiths: Comparing ChatGPT, GPT-4, Claude, and Bard. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4476855

Chen, X., Xie, H., & Tao, X. (2022). Vision, status, and research topics of Natural Language Processing. *Natural Language Processing Journal, 1*, 100001. https://doi.org/10.1016/j.nlp.2022.100001

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., … Fiedel, N. (2022). *PaLM: Scaling language modeling with pathways*. http://arxiv.org/abs/2204.02311

DivyaSingh456. (2019, May 2). Evolution of chatbots & their performance. *DataScienceCentral.Com*. https://www.datasciencecentral.com/evolution-of-chatbots-amp-their-performance/

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). *Explaining and harnessing adversarial examples*. http://arxiv.org/abs/1412.6572

Gu, J., Han, Z., Chen, S., Beirami, A., He, B., Zhang, G.,... & Torr, P. (2023). A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*.

Gupta, A., Hathwar, D., & Vijayakumar, A. (n.d.). *Introduction to AI chatbots*. www.ijert.org

Kim, T. H. (2010). Emerging approach of natural language processing in opinion mining: A review. *Communications in Computer and Information Science, 75* CCIS. https://doi.org/10.1007/978-3-642-13467-8_12

Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization*. http://arxiv.org/abs/1412.6980

Koubaa, A., Boulila, W., Ghouti, L., Alzahem, A., & Latif, S. (2023). Exploring ChatGPT capabilities and limitations: A survey. *IEEE Access, 11*, 118698–118721. https://doi.org/10.1109/ACCESS.2023.3326474

Kovan, M., & Márta, T.-S. (2023). Chatbot development using APIs and integration into the MOOC. *Journal of New Technologies in Research, 5*(1).

Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S. F., Forshee, R., Walderhaug, M., & Botsis, T. (2017). Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *Journal of Biomedical Informatics, 73*. https://doi.org/10.1016/j.jbi.2017.07.012

OpenAI. (2022, November 20). *Introducing ChatGPT*. OpenAI. https://openai.com/blog/chatgpt

Oppenlaender, J., Linder, R., & Silvennoinen, J. (2023). *Prompting AI art: An investigation into the creative skill of prompt engineering*. http://arxiv.org/abs/2303.13534

Pons, E., Braun, L. M. M., Hunink, M. G. M., & Kors, J. A. (2016). Natural language processing in radiology: A systematic review. *Radiology, 279*(2). https://doi.org/10.1148/radiol.16142770

Psyarxiv Manuscript. (n.d.).

Qin, R., Huang, M., Liu, J., & Miao, Q. (2022). Hybrid attention-based transformer for long-range document classification. *Proceedings of the International Joint Conference on Neural Networks, 2022-July*. https://doi.org/10.1109/IJCNN55064.2022.9891918

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems, 3*, 121–154. KeAi Communications Co. https://doi.org/10.1016/j.iotcps.2023.04.003

Samant, R. M., Bachute, M. R., Gite, S., & Kotecha, K. (2022). Framework for deep learning-based language models using multi-task learning in natural language understanding: A systematic literature review and future directions. *IEEE Access, 10*. https://doi.org/10.1109/ACCESS.2022.3149798

Srivastava, N., Hinton, G., Krizhevsky, A., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research, 15*.

Statista (2023, August). *Natural language processing - Global | Market forecast*. *Statista*. https://www.statista.com/outlook/tmo/artificial-intelligence/natural-language-processing/worldwide#market-size

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 2017-December*.

Velásquez-Henao, J. D., Franco-Cardona, C. J., & Cadavid-Higuita, L. (2023). Prompt engineering: A methodology for optimizing interactions with AI-language models in the field of engineering. *DYNA, 90*(230), 9-17. https://doi.org/10.15446/dyna.v90n230.111700

Wang, J., Shi, E., Yu, S., Wu, Z., Ma, C., Dai, H., Yang, Q., Kang, Y., Wu, J., Hu, H., Yue, C., Zhang, H., Liu, Y., Li, X., Ge, B., Zhu, D., Yuan, Y., Shen, D., Liu, T., & Zhang, S. (2023). *Prompt engineering for healthcare: Methodologies and applications*. http://arxiv.org/abs/2304.14670

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q. L., & Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo and potential future development. IEEE/CAA Journal of Automatica Sinica, 10(5), 1122–1136. https://doi.org/10.1109/JAS.2023.123618