Gemini-the most powerful LLM: Myth or Truth

Raisa Islam*, Imtiaz Ahmed†
Dept of Computer Science & Engineering
New Mexico Institute of Mining and Technology
Socorro, NM 87801 USA
Email: *raisa.islam@student.nmt.edu, †imtiaz.ahmed@student.nmt.edu

***Abstract*—Gemini models excel in various tasks including image generation and interpretation, video understanding, and solving mathematical problems, among others. The Vertex AI Gemini API and Google AI Gemini API both enable developers to integrate Gemini model functionalities into their applications. This paper offers a concise summary of the Gemini Framework, focusing on its distinctive modalities that distinguish it from current systems. In our research, we explored the details of its architecture, pointing out the innovative strategies employed to improve generative AI capabilities. Furthermore, we conduct a comparative study, assessing Gemini's performance against other top generative AI models.**

***Index Terms*—Gemini Framework, Generative AI, OpenAI GPT, Multimodal AI Models, Comparative AI Analysis, Google AI Innovations, Conversational AI, AI in Human-Machine Interaction, Digital Assistants Evolution, AI Multimodality**

I. Introduction

Technological shift brings the opportunity to advance scientific discovery, accelerate human progress, and improve lives [1]. The advancement in the field of artificial intelligence (AI) holds the potential to create new horizon for knowledge, learning, creativity and productivity. This requires pursuing ambitious research objectives and aiming for capabilities that can provide substantial benefits to people and society. In this competitive field, Google is strategically positioned to enter into a contest with Microsoft and its collaborator, OpenAI. Google has tactically rebranded its Bard chatbot and Duet AI, consolidating them into the newly launched Google Gemini framework. This initiative marks a significant shift in the approach to AI. Text-based interaction models played a significant role in the evolution of AI, however, human beings navigate a dynamically evolving environment which requires processing of more complex information. Moreover, human communication transcends mere textual exchanges, encompassing sophisticated modalities such as speech and visual imagery. Google Gemini represents an endeavor to bridge this gap, aiming to achieve a more comprehensive and nuanced understanding of the world, akin to human cognition.

The evolution of digital interaction tools introduced voice-activated digital assistants such as Siri, Alexa, and Google Assistant, followed by the development of online chatbots, ChatGPT and Google Bard. Now, a more sophisticated technology has been introduced with Google's 2024 launch of Gemini [2]. Gemini is the next generation of Google's large language model (LLM). Gemini integrates the features of both talking digital assistants and conversational chatbots. It is adept at handling voice and text inputs, enabling it to perform a wide array of tasks. Gemini's design aims to cater to diverse needs, functioning as a personal tutor, aiding programmers with coding endeavors, and preparing job seekers for interviews, showcasing its versatility and the broad scope of its capabilities.

A multimodal model is a model that is capable of processing and relating information from multiple modalities [3]. Gemini is a family of GenAI models developed by Google DeepMind that is designed for multimodal use cases. Gemini models are capable of understanding and generating images, understanding videos, solve mathematical problems to name a few. The Vertex AI Gemini API and Google AI Gemini API both allow the users to incorporate the capabilities of Gemini models into applications.

In this paper, we present a brief overview of the Gemini Framework, emphasizing the unique modalities it incorporates, which set it apart from existing systems. We delve into the specifics of its architecture, highlighting the innovative approaches it adopts to enhance generative AI capabilities. Additionally, we undertake a thorough comparative analysis, meticulously evaluating Gemini's performance in relation to a variety of other leading generative AI models across multiple dimensions.

II. Background

A. Generative AI (GenAI)

GenAI refers to a subset of AI technologies capable of generating new content that is similar to human-generated content. This is achieved through learning from a vast dataset of existing content in a specific domain. The primary goal of GenAI includes to understand and interpret data to create new, original content which is coherent, contextually relevant, and often indistinguishable from content created by humans [4]. Based on the type of input, GenAI may operate as either unimodal or multimodal; unimodal systems accept a single type of input, while multimodal systems are capable of processing various types of input.

B. Multimodal AI

Multimodal AI marks a significant evolution, synthesizing diverse data modalities i.e.,text, images, audio, and video, to improve comprehension, analysis, and decision-making [3]. Additionally, the use of multimodal fusion methods, including self-attention mechanisms and sparse fusion techniques, has demonstrated significant performance improvements in various

AI applications. In areas like affective computing and sentiment analysis, multimodal AI integrates visual, acoustic, and language modalities to improve the accuracy of human emotions and sentiments analysis.

Early AI models, such as Word2Vec, VGG, ResNet, and DeepSpeech, were designed to operate in unimodal fashion, focusing on specific types of data such as text, images, or speech. Despite their proficiency within specific domains, these models lacked the capability to integrate information across various modalities for a comprehensive understanding of data. The evolution towards multimodal AI necessitated the development of various fusion techniques to synthesize data from different modalities. However, overcoming feature alignment and data type heterogeneity posed significant challenges [5].

With the advent of LLMs, evaluation criteria shifted towards assessing natural language nuances and the generation of coherent text. Noteworthy LLMs like BERT [6] and GPT garnered attention for their contextual understanding and generation abilities [7], subsequently evolving into conversational AI models such as LaMDA [8]. The recent introduction of advanced multimodal AI models like OpenAI GPT-4V [9], Meta ImageBind, and Google Gemini marks a significant shift in multimodal AI research, unlocking the potential to understand and generate content integrating text and images [10]. However, assessing multimodal AI presents challenges, including knowledge transfer between modalities and discerning causality within multimodal datasets. Given the novelty of multimodal AI, its full spectrum of applications and benefits remains under exploration, necessitating adaptable evaluation methodologies to fully comprehend its complexity.

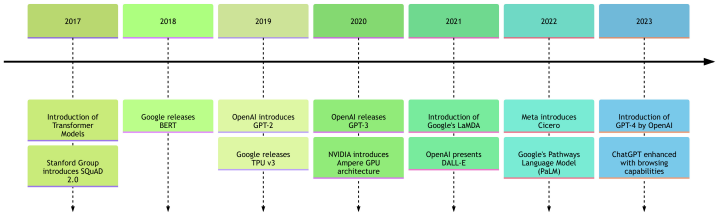

Fig. 1. Journey of LLMs

C. Journey of Google

Google's trajectory from ML tools to multimodal AI exemplifies a series of pivotal advancements that underscore the company's dedication to AI innovation. In the early stages, Google integrated ML applications, such as spell correction, translate e.t.c., into its products to enhance user experiences. A significant milestone in Google's AI journey is the introduction of TensorFlow [11], as this open-source ML framework revolutionized the accessibility of AI technologies and accelerated global AI research endeavors. In 2016, Google's DeepMind achieved acclaim with *AlphaGo*, a deep learning system that defeated a world champion Go player, showcasing the potential of deep learning in tackling complex challenges. Moreover, the introduction of the Transformer architecture revolutionized language understanding and paved the way for future AI models [12].

The emergence of LLMs marked another noteworthy phase in Google's AI evolution. In 2021, Google Research launched LaMDA [8], a conversational LLM that represented a significant advancement in natural language processing (NLP). Followed by the introduction of Bard [13] in 2023, which integrated GenAI capabilities into Gmail, Workspace, and Google Search, further enhancing user interactions with AI-driven features. Google's journey towards multimodal AI culminated in the development of Gemini, a groundbreaking multimodal model capable of processing diverse data types including text, code, audio, image, and video. Gemini's native multimodal design, coupled with its pre-training on various modalities and fine-tuning capabilities, underscores Google's commitment to advancing AI technologies.

Google is persistently advancing its AI technologies, especially through the Gemini initiative. Future directions include advancements in planning, memory, and increasing the context window size to process even larger volumes of information [14]. This relentless pursuit of innovation underscores Google's dedication to crafting more advanced, versatile AI solutions poised to revolutionize AI interactions into adept assistants.

D. Prompting Strategies

- ***Zero-Shot (0-shot) prompting*** refers to the ability of a language model to perform a task without any explicit training data or examples for that task. Instead, the model can use its knowledge of language and the relationships between words to perform the task based on a textual description or prompt.
- ***Few-Shot (n-shot) prompting*** involves providing the model with a small set of inputs and outputs before the final input, allowing them to leverage previous examples to better understand and respond to new questions [15].
- ***Chain of Thought (CoT@n) prompting*** utilizes intermediate steps of reasoning before presenting the sample answer. This approach involves decomposing complex problems into smaller, more manageable steps, which is believed to aid a foundation model in producing a more precise answer.
- ***maj1@k***, a variant of CoT, denotes evaluations where k samples were generated for each problem and only the majority vote (most common answer) was selected [16]. In cases where CoT prompting is ineffective, using Maj1@k oftentimes improves results.

III. A Brief Overview of Gemini Framework

Initially, multimodal models were developed by training separate components for each modality and then integrating them, which worked well for tasks like image description but struggled with complex reasoning. Gemini was designed to be inherently multimodal, pre-training on diverse modalities. Simultaneous training on text, programming code, images, audio and video, has enabled Gemini to more efficiently cope

with multimedia input compared to other GenAI models [17]. Subsequently, it was fine-tuned with additional multimodal data to enhance efficiency enabling Gemini to effortlessly comprehend and reason about a wide array of inputs.

In this section, we will discuss all the Gemini models and technologies used to enhance their performance.

A. Prototypes

Gemini comes in three versions tailored for different levels of computing power:

a) ***Gemini Pro*** *(GPro):* is an advanced LLM designed for understanding and generating human language. It enables user interaction through both single-turn interactions and multi-turn conversations, including capabilities for understanding and generating code. As a foundational model, GPro excels across a wide range of NLP tasks. More about GPro will be discussed in section IV.

b) ***Gemini Nano*** *(GNano):* is claimed to be *most efficient model*, specifically engineered to run on smartphones and is available in two distinct versions to accommodate different memory capacities. This versatility ensures that users can benefit from its advanced capabilities regardless of device's memory size. It introduces innovative features aimed at enhancing user experience and productivity. A prominent feature is the ability to summarize dialogues within its Recorder application, providing users with concise summaries of their recorded conversations. Another key feature, in collaboration with Google's Gboard, is offering intelligent response suggestions for WhatsApp. This functionality streamlines communication by suggesting contextually appropriate replies, thereby saving time and improving the efficiency of text-based interactions.

c) ***Gemini Ultra*** *(GUltra):* a paid premium version of Gemini that uses Google's biggest and most advanced LLM. It can perform across a wide range of highly complex tasks, including reasoning and multimodal tasks. It is efficiently serveable at scale on Tensor processing unit (TPU) accelerators due to the Gemini architecture. It was reported that GPro requires only a small portion of the computational resources needed for GUltra, suggesting that GUltra is a significantly larger model.

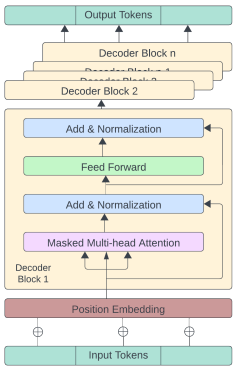

Fig. 2. Decode-only transformer diagram

B. Technologies enhancing Gemini performance

a) Decoder-only Transformer model: Similar to many generative AI models, Gemini models build on top of decoder-only transformers (base model [18]). However, the standard decoder-only architecture was modified to enhance efficiency and stabilize training at scale and optimized inference on Google's TPUs [2]. They employed multi-query attention, a method that augments multi-head attention's efficiency by allowing attention heads to share key and value vectors. Additionally, Gemini leverages some of the optimization and architectural tricks, i.e., Lion optimizer[1], Low Precision Layer Normalization, Flash Attention[1], and, Flash Decoding (build on top of Flash Attention) [19].

b) TPU Accelerators: Gemini models were trained using TPUv4 and TPUv5e, based on their respective sizes and configurations. These specially designed AI accelerators are core to Google's AI-driven products, empowering cost-effective training of AI models. TPU v4 includes with SparseCores, specialized dataflow processors that enhance the performance of models dependent on embeddings by 5 to 7 times, while consuming merely 5% of the die area and power [20]. The training of GUltra, which utilized numerous TPUv4 accelerators spread across various data centers, resulted in a proportional decrease in the mean time between hardware failures throughout the system.

TPUv5e is the newest generation of AI accelerators, a successor of TPUv4 lite. It features a compact 256-chip configuration per Pod wherein TPUv4 have 4096 chips per Pod. These Pods are tailored for training, fine-tuning, and deploying transformer-based, text-to-image, and CNN-based models. TPUv5e enables Google to inference models that are larger than OpenAI at the same cost as OpenAI's smaller model.

c) Retrieval-Augmented Generation (RAG): Fundamentally, RAG is an AI framework designed for information optimization minimize the amount of irrelevant information to the model by feeding more relevant, external information.

Due to limited context window, GPro integrates RAG for information retrieval with text generation, resulting in factually grounded outputs. RAG access useful passages from the book; indexes them using TF-IDF; and stores the results in an external database. By utilizing cosine similarity, the passages are re-ranked, and the most relevant passages are retrieved (up to 4k tokens). The retrieved passages are then put into context following a temporal ordering.

C. Safety Policies

Privacy issues in multimodal AI arise due to the ability to correlate various data sources, potentially leading to invasive surveillance and profiling which raises concerns about individual consent and rights. In accordance with the AI Principles[2], protective measures are incorporated at every development phase to address potential risks, including bias and toxicity.

[1]research supported by Google.

[2]https://ai.google/responsibility/principles

	GUltra	GPro	GPT-4	GPT-3.5	PaLM 2-L	Claude 2	Inflection-2	Grok 1	LLAMA-2
MMLU MCQ in 57 subjects (professional & academic)	90.04% CoT@32*	79.13% CoT@8*	87.29% CoT@32 (via API**)						
	83.7% 5 shot	71.8% 5 shot	86.4% 5 shot (reported)	70.0% 5 shot	78.4% 5 shot	78.5% 5 shot	79.6% 5 shot	73.0% 5 shot	68.0%***
GSM8K Grad school Math	94.4% Maj1@32	86.5% Maj1@32	92.0% SFT & 5 shot CoT	57.1% 5 shot	80.0% 5 shot	88.0% 0 shot	81.4% 8 shot	62.9% 8 shot	56.8% 5 shot
Math 5 difficulty level & 7 sub-disciplines	53.2% 4 shot	32.6% 4 shot	52.9% 4 shot (via API**)	34.1% 4 shot (via API**)	34.4% 4 shot		34.8%	23.9% 4 shot	13.5% 4 shot
Big-Bench-Hard subset of hard task written as CoT problems	83.6% 3 shot	75.0% 3 shot	83.1% 3 shot (via API**)	66.6% 3 shot (via API**)	77.7% 3 shot				51.2% 3 shot
HumanEval Python coding tasks	74.4% 0 shot (PT****)	67.7% 0 shot (PT****)	67.0% 0 shot (reported)	48.1% 0 shot		70.0% 0-shot	44.5% 0 shot	63.2% 0 shot	29.9% 0 shot
Natural2Code Python code generation (New held-out set with no leakage on web)	74.9% 0 shot	69.6% 0 shot	73.9% 0 shot (via API**)	62.3% 0 shot (via API**)					
DROP reading comprehension & arithmetic. (metric: F1 score)	82.4 variable shot	74.1 variable shot	80.9 3 shot (reported)	64.1 3 shot	82.0 variable shot				
HellaSwag (validation set) common-sense MCQ	87.8% 10 shot	84.7% 10 shot	95.3% 10 shot (reported)	85.5% 10 shot	86.8% 10 shot		89.0% 10 shot		80.0%***
WMT23 Machine translation (metric: BLEURT)	74.4 1 shot (PT****)	71.7 1 shot	73.8 1 shot (via API**)		72.7 1 shot				

TABLE I
PERFORMANCE COMPARISON BY GOOGLE

Novel research into potential risk areas, such as cyber-offense, persuasion, and autonomy, and adversarial testing techniques to uncover critical safety issues makes the system more robust [21]. Moreover, a diverse group of external experts and partners are engaged in rigorously testing the internal evaluation methods to uncover any blind spots. The *Real Toxicity Prompts* benchmark is utilized to assure content safety throughout the training process and to ensure the output complies with established policies.

In an effort to mitigate harm, safety classifiers are designed and implemented to detect, label, and segregate content related to violence or negative stereotypes. This multi-tiered strategy, enhanced by robust filters, aims to render safety and inclusivity for all users. Additionally, the team is actively working on ongoing challenges associated with model performance, such as factuality, grounding, attribution, and corroboration. To establish best practices and safety and security benchmarks, Gemini has formed partnerships with MLCommons [22], the Frontier Model Forum [23], and the Secure AI Framework (SAIF) [24].

IV. GEMINI PRO (GPRO)

GPro is an advanced large language model designed for understanding and generating human language. It enables user interaction through both single-turn interactions and multi-turn conversations, including capabilities for understanding and generating code. The key features of GPro are the followings [25, 26]:

- *Text Summarization*: Create summary from any document retaining essential information, such as summarizing a textbook chapter.
- *Object Recognition*: Identify objects within images and videos with detailed precision.
- *Content Understanding*: Extract and generate descriptions from various digital content forms, such as charts, figures, and many more, and provide answers based on given content.
- *Content Generation*: Create content based on specific requirements or context, for instance, composing an email in a particular tone for a given scenario, or crafting HTML/JSON responses based on provided prompts.
- *Extrapolation*: Predict unseen elements in an image or events that occur before or after a captured video sequence.
- *Classification*: Categorize text by assigning labels that describe its characteristics, such as detecting sentiment conveyed in a text, identifying whether the text expresses positive or negative emotions.

V. PERFORMANCE ANALYSIS

Benchmarking is a critical part of evaluating progress in large language models. The scores provide a snapshot of progress, with new state-of-the-art (SOTA) results heralded as breakthroughs. LLMs are usually evaluated in a zero-shot setting, without explicit training on the test set, to gauge their general abilities. However, both Gemini and GPT reported respective performance evaluation after using different prompting strategies (described in Section II-D).

In this section, we will discuss Gemini vs other generative AI models performance reported by Goggle and other researchers.

A. Analyzing Performance Reported by Google

Google Gemini team published a report comparing Gemini performance with other available MML model [2]. GUltra is claimed to be the first model to surpass human-expert performance on *MMLU* with a score above 90%. Interestingly, to achieve the stated accuracy, GUltra uses *uncertainty-routed CoT@32*, which makes a direct comparison of these results somewhat misleading. With only CoT@32, GUltra accuracy is 84.99% (see [2] Appendix section). Nevertheless, GUltra, being a newer model, should win on 5-shot itself against GPT-4. For 5-shot, GPT-4 outperforms Gemini with a good margin (+2.7%). Similarly, to achieve new SOTA on *GSM8K* and *DROP*, GUltra uses Maj1@32 which is a variant of CoT prompting, and variable shot prompting respectively. However,

GUltra outperforms GPT-4 on coding benchmark *HumanEval* with an impressive margin (+7.4%). In most of the other cases where same type of prompting strategies were used (Math, Big-Bench-Hard, Natural2Code and WMT23), performance difference between Gemini and GPT-4 is $\leq 1\%$. There are reservations about the claim that GUltra performs better than GPT-4 as this version has not been released yet.

Gemini models shine the most on multimodal tasks. For all the benchmarks related to image understanding, GUltra outperforms prior SOTA; however GPro fails to surpass them except for *InfographicVQA* test ([2] Table 7). In audio understanding tasks, i.e., automatic speech recognition and automatic speech translation, GPro surpasses well-known models such as Google USM [27] and OpenAI Whisper [28] across all benchmarks. Gemini models represent a pioneering effort in the integration of video content with LLMs. Considering that videos constitute a vast, yet unexplored, data reservoir for AI, such an approach holds significant potential.

B. Analysis by Other Researchers

Wang et al. [29] identified a limitation in evaluating Gemini's capability for commonsense reasoning, since the assessment was based only on HellaSWAG dataset. To bridge the gap and provide a more thorough assessment on commonsense reasoning tasks, the researchers performed experiments involving 12 different commonsense reasoning datasets spanning a wide range of domains, including general, physical, social, and temporal reasoning. The researchers selected zero-shot and CoT@n prompting methods for this experiment. The study shows that the efficiency of GPro is comparable with GPT-3.5 in language-only commonsense reasoning tasks, demonstrating logical and contextual reasoning processes. Nonetheless, it lags behind of GPT-4 in terms of accuracy, and encounters challenges in temporal and social reasoning, and emotion recognition within images.

Lee et al. [30] conducted a comparison between GPro and GPT-4V within educational frameworks, utilizing the Novel Educational Rubric Interpretation Framework (NERIF) and few-shot prompting. The objective was to evaluate the ability of both models to understand text-based educational rubrics and independently assess student-generated diagrams in science education, utilizing visual question answering (VQA) methodologies. The findings indicated that GPT-4V gives better accuracy in image classification as well as processing detailed text in images. Despite adjustments to NERIF, GPro was unable to reach the performance level of GPT-4V, making GPT-4V a more suitable for educational applications. In another study, Liu et al. [31] evaluated GPT-4V and GPro for VQA task using *VQAonline* dataset. VQAonline dataset consists authentic information needs of everyday online users,and each ground truth answer is verified by the user who posed the question. In a zero-shot prompting setting, the average accuracy of GPT-4V and Gemini is 0.53 and 0.42, respectively.

In a study [15][3], researchers evaluated GPT-4, GPro and MedPaLM 2 in the context of medical reasoning, hallucination detection, and medical visual question answering (VQA) tasks. The findings demonstrated that Gemini underperformed in comparison to both MedPaLM 2 and GPT-4, with GPro achieving an accuracy rate of 61.45%, while GPT-4V attained an 88% accuracy rate. Additionally, the study exposed Gemini's vulnerability to hallucinations, overconfidence, and knowledge deficiencies, highlighting potential risks associated with its deployment without careful consideration. Another medical case study [33], reported that GPT-4 is able to achieve approx. 90.2% accuracy on MMLU dataset using a modified version of Medprompt (Medprompt+@31) which is better than GUltra.

A language specific, impartial and reproducible evaluation of the capabilities of GPT, Gemini, and MixTral model classes was conducted to present fully transparent outcomes in [32]. The study aims to thoroughly examine the results to pinpoint the areas where each model class performs exceptionally. Findings show that GPro is defeated by GPT-4; and in times GPT-3.5 and Mixtral performs better.

Findings of mentioned studies in this section are presented in Table II.

C. Other

GPT-4 was tested on academic and professional exams, originally designed for humans, and final score was graded according to exam specific rubrics. GPT team claimed that a minority of the test set questions in the exams might be present in the training set; but they didn't train specifically for those tests [34]. However, Gemini did not provide any such reports. Using a specialized version of Gemini, AlphaCode [35] and AlphaCode 2 [36] was created, the first AI code generation system to reach a competitive level of performance in programming competitions. While GPT was tested on Leetcode platform, AlphaCode was tested on Codeforces platform and ranked within the top 15% of entrants.

VI. Conclusion

Gemini models excel in image generation, video understanding, and solving mathematical problems, among other capabilities. Our study offers an overview of the Gemini Framework, showcasing its distinctive modalities and architectural innovations that enhance Gemini. We also provided a comparative analysis, evaluating the performance of Gemini against other leading GenAI models. Despite claims that GUltra outperforms other models in the report provided by Gemini becomes questionable due to the use of different prompting methods. Since, GUltra is not available to public, other researcher conducted performance comparison between GPro and other models where GPro underperforms in most of the cases.

The Gemini team has revealed intentions to release GUltra to the public in the near future. This includes undergoing rigorous scrutiny through red-teaming by reputable external groups, as well as enhancing the model with fine-tuning and reinforcement learning from human feedback (RLHF) to ensure its readiness for widespread use.

[3] 5-shot GPT-4 base model, best available model for other MMLs

Research	Dataset	Prompting Method	GPro	GPT-4	GPT-3.5	Other	
						Llama-2-70b	
[29]	CommonsenseQA	0 shot	76.5	78.0	73.0	72.0	
		CoT@k	79.0	80.0	76.0	76.5	
	Cosmos QA	0 shot	81.5	86.5	75.0	77.0	
		CoT@k	84.5	88.0	78.5	81.0	
	αNLI	0 shot	79.5	87.0	75.5	77.5	
		CoT@k	81.5	88.0	78.0	80.5	
	HellaSWAG	0 shot	76.0	94.0	78.0	73.0	
		CoT@k	78.5	95.0	80.0	77.0	
	TRAM	0 shot	73.5	79.5	68.5	66.0	
		CoT@k	76.0	82.0	72.0	70.0	
	NumerSense	0 shot	80.0	85.0	81.5	74.0	
		CoT@k	82.0	86.0	82.5	75.5	
	PIQA	0 shot	89.0	94.5	87.0	74.0	
		CoT@k	90.5	95.5	89.5	78.5	
	QASC	0 shot	80.0	91.5	83.0	78.0	
		CoT@k	82.5	92.5	85.0	82.0	
	RiddleSense	0 shot	75.0	94.0	71.5	62.5	
		CoT@k	82.5	95.0	75.0	66.0	
	Social IQa	0 shot	73.0	82.0	73.0	71.0	
		CoT@k	78.5	84.5	78.0	77.5	
	ETHICS	0 shot	87.0	97.0	94.0	88.0	
		CoT@k	87.5	98.0	95.0	89.5	
[30]	MMMU	n shot	47.9%	56.8%			
	TextVQA	n shot	74.6%	78.0%			
	DocVQA	n shot	88.1%	88.4%			
	ChartQA	n shot	74.1%	78.5%			
	InfographicVQA	n shot	75.2%	75.3%			
	MathVista	n shot	45.2%	49.9%			
	AIZ2D	n shot	73.9%	78.2%			
	VQA v2	n shot	71.2%	77.2%			
[15]						**Flan-PaLM**	**Med-PaLM 2**
	MedQA USMLE		67.0	86.1		67.6	86.5
	PubMedQA		70.7	80.4		79.0	81.8
	MedMCQA		62.2	73.7		57.6	72.3
	MMLU clinical knowledge		78.6	88.7		80.4	88.7
	MMLU medical genetics		81.8	97.0		75.0	92.0
	MMLU anatomy		76.9	85.2		63.7	84.4
	MMLU pro. medicine		83.3	93.8		83.8	92.3
	MMLU college biology		89.5	97.2		88.9	95.8
	MMLU college medicine		79.3	80.9		76.3	83.2
[31]	VQAonline	0 shot	0.42	0.53			
[32]						**Mixtral**	
	MMLU	5 shot	65.22	80.48	67.75	68.81	
		CoT@k	62.09	78.95	70.07	59.57	
	Big-Bench-Hard		67.53	83.90	71.02	60.76	
	GSM8K		76.42	92.72	78.01	71.65	
	SVAMP		81.10	92.60	82.30	81.60	
	ASDIV		85.31	92.75	89.07	83.16	
	MAWPS		96.50	98.67	98.00	96.00	
	HumanEval		59.76	76.83	74.39	45.12	
	ODEX		39.86	45.79	52.62	40.55	
	FLORES unblocked	5 shot	53.31	54.00	52.43	40.97	
	FLORES all	5 shot	21.68	48.24	40.00	30.27	
	WebArena		7.12	14.90	8.87	1.39	

TABLE II
PERFORMANCE COMPARISON BY RESEARCH STUDIES

REFERENCES

[1] D. H. Sundar Pichai. Introducing gemini: our largest and most capable ai model. *Google Blog*, Dec. 6, 2023. URL: https://blog.google/technology/ai/google-gemini-ai. Last accessed 2 March, 2024.

[2] Gemini Team, Google. Gemini: A Family of Highly Capable Multimodal Models, 2024. DOI: 10.48550/arXiv.2312.11805.

[3] T. Baltrušaitis et al. Multimodal Machine Learning: A Survey and Taxonomy, 2019. DOI: 10.1109/TPAMI.2018.2798607.

[4] Z. Lv. Generative artificial intelligence in the metaverse era, 2023. DOI: 10.1016/j.cogr.2023.06.001.

[5] T. McIntosh et al. From Google Gemini to OpenAI Q* (Q-Star): A Survey of Reshaping the Generative Artificial Intelligence (AI) Research Landscape, Dec. 2023. DOI: 10.48550/arxiv.2312.10868.

[6] J. Devlin et al. Bert: pre-training of deep bidirectional transformers for language understanding, 2018. DOI: 10.48550/arXiv.1810.04805.

[7] OpenAI. Chatgpt: optimizing language models for dialogue. https://openai.com/blog/chatgpt/, Nov. 2022. Last accessed 29 February 2024.

[8] R. Thoppilan et al. LaMDA: Language Models for Dialog Applications, 2022. DOI: 10.48550/arXiv.2201.08239.

[9] L. Teixeira. The New OpenAI GPT-4 Vision on ChatGPT: Bridging the Gap Between Text and Image Understanding. URL: https://medium.com/@lawrenceteixeira/the-new-open-ai-gpt-4-vision-on-chatgpt-bridging-the-gap-between-text-and-image-understanding-9337ed4c1a61. Last accessed 29 February 2024.

[10] R. Girdhar et al. Imagebind one embedding space to bind them all. DOI: 10.1109/CVPR52729.2023.01457.

[11] M. Abadi et al. Tensorflow: a system for large-scale machine learning, 2016. DOI: 10.5555/3026877.3026899.

[12] Google. A timeline of google's biggest ai and ml moments. https://blog.google/technology/ai/google-ai-ml-timeline/, Sept. 2023. Last accessed 2 March 2024.

[13] J. Manyika and S. Hsiao. An overview of Bard: an early experiment with generative AI. https://ai.google/static/documents/google-about-bard.pdf. Last accessed 3 March 2024.

[14] Artificial Intelligence News. AI Google News: Latest AI Developments at Google, 2024. URL: https://www.artificialintelligence-news.com/categories/ai-companies/google. Last accessed 2 March 2024.

[15] A. Pal and M. Sankarasubbu. Gemini Goes to Med School: Exploring the Capabilities of Multimodal Large Language Models on Medical Challenge Problems & Hallucinations. DOI: 10.48550/arXiv.2402.07023.

[16] A. Lewkowycz et al. Solving Quantitative Reasoning Problems with Language Models, 2022. DOI: 10.48550/arXiv.2206.14858.

[17] S. Shankland. Google Gemini AI Tries Outsmarting ChatGPT Using Photos and Videos. *CNET*, Dec. 13, 2023. URL: https://www.cnet.com/tech/computing/google-gemini-ai-tries-outsmarting-chatgpt-using-photos-and-videos. Last accessed 2 March, 2024.

[18] A. Vaswani et al. Attention is all you need, 2023. DOI: 10.48550/arXiv.1706.03762.

[19] C. R. Wolfe. Google Gemini: Fact or fiction?, Dec. 23, 2023. URL: https://cameronrwolfe.substack.com/p/google-gemini-fact-or-fiction. Last accessed 2 March, 2024.

[20] N. Jouppi et al. TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings, 2023. DOI: 10.1145/3579371.3589350.

[21] Kathy Meier-Hellstern. Responsible AI at Google Research: Adversarial testing for generative AI safety, Nov. 16, 2023. URL: https://blog.research.google/2023/11/responsible-ai-at-google-research_16.html. Last accessed 27 February 2024.

[22] MLCommons Research Working Group. Mlcommons, 2023. URL: https://mlcommons.org/working-groups/. Last accessed 2 March, 2024.

[23] P. Policy. Frontier model forum: a new partnership to promote responsible ai. *Google Blog*, June 28, 2022. URL: https://blog.google/outreach-initiatives/public-policy/google-microsoft-openai-anthropic-frontier-model-forum. Last accessed 27 February 2024.

[24] P. V. Royal Hansen. Introducing google's secure ai framework. *Google Blog*, June 8, 2022. URL: https://blog.google/technology/safety-security/introducing-googles-secure-ai-framework. Last accessed 27 February 2024.

[25] Gemini Pro. URL: https://console.cloud.google.com/vertex-ai/publishers/google/model-garden/gemini-pro. Last accessed 2 March, 2024.

[26] Overview of multimodal models. Google Cloud. URL: https://cloud.google.com/vertex-ai/generative-ai/docs/multimodal/overview#visual_understanding. Last accessed 5 March 2024.

[27] Y. Zhang et al. Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages, Mar. 2023. DOI: 10.48550/arXiv.2303.01037.

[28] A. Radford et al. Robust Speech Recognition via Large-Scale Weak Supervision, 2023. DOI: 10.5555/3618408.3619590.

[29] Y. Wang and Y. Zhao. Gemini in Reasoning: Unveiling Commonsense in Multimodal Large Language Models, 2023. DOI: 10.48550/arXiv.2312.17661.

[30] G.-G. Lee et al. Gemini Pro Defeated by GPT-4V: Evidence from Education, 2023. DOI: 10.48550/arXiv.2401.08660.

[31] M. Liu et al. An Evaluation of GPT-4V and Gemini in Online VQA, 2024. DOI: 10.48550/arXiv.2312.10637.

[32] S. N. Akter et al. An In-depth Look at Gemini's Language Abilities, 2023. DOI: 10.48550/arXiv.2312.11444.

[33] H. Nori et al. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. Dec. 2023. DOI: 10.48550/arXiv.2311.16452.

[34] OpenAI Team. GPT-4 Technical Report, 2023. DOI: 10.48550/arXiv.2303.08774.

[35] Y. Li et al. Competition-level code generation with alphacode, Dec. 2022. DOI: 10.1126/science.abq1158.

[36] AlphaCode Team. Alphacode 2 technical report. URL: https://api.semanticscholar.org/CorpusID:266058988.