

Article

Putting GPT-4o to the Sword: A Comprehensive Evaluation of Language, Vision, Speech, and Multimodal Proficiency

Sakib Shahriar ¹, Brady D. Lund ^{2,*}, Nishith Reddy Mannuru ², Muhammad Arbab Arshad ³, Kadhim Hayawi ⁴, Ravi Varma Kumar Bevara ², Aashrith Mannuru ⁵ and Laiba Batool ⁶

¹ School of Computer Science, University of Guelph, Guelph, ON N1G 2W1, Canada; shahrias@uoguelph.ca

² Department of Information Science, University of North Texas, Denton, TX 76201, USA;
nishithreddymannuru@my.unt.edu (N.R.M.); ravivarmakumarbevara@my.unt.edu (R.V.K.B.)

³ Department of Computer Science, Iowa State University, Ames, IA 50011, USA; arbab@iastate.edu

⁴ College of Interdisciplinary Studies, Zayed University, Dubai, United Arab Emirates; abdul.hayawi@zu.ac.ae

⁵ Naveen Jindal School of Management, University of Texas at Dallas, Richardson, TX 75080, USA;
arm210018@utdallas.edu

⁶ National University of Computer & Emerging Sciences, Islamabad 44000, Pakistan;
laiba.batool.1426@gmail.com

* Correspondence: brady.lund@unt.edu

Abstract: As large language models (LLMs) continue to advance, evaluating their comprehensive capabilities becomes significant for their application in various fields. This research study comprehensively evaluates the language, vision, speech, and multimodal capabilities of GPT-4o. The study employs standardized exam questions, reasoning tasks, and translation assessments to assess the model's language capability. Additionally, GPT-4o's vision and speech capabilities are tested through image classification and object-recognition tasks, as well as accent classification. The multimodal evaluation assesses the model's performance in integrating visual and linguistic data. Our findings reveal that GPT-4o demonstrates high accuracy and efficiency across multiple domains in language and reasoning capabilities, excelling in tasks that require few-shot learning. GPT-4o also provides notable improvements in multimodal tasks compared to its predecessors. However, the model shows variability and faces limitations in handling complex and ambiguous inputs, particularly in audio and vision capabilities. This paper highlights the need for more comprehensive benchmarks and robust evaluation frameworks, encompassing qualitative assessments involving human judgment, as well as error analysis. Future work should focus on expanding datasets, investigating prompt-based assessment, and enhancing few-shot learning techniques to test the model's practical applicability and performance in real-world scenarios.

Keywords: GPT; large language models; exam; translation; classification; vision–language; review; omni; reasoning



Citation: Shahriar, S.; Lund, B.D.; Mannuru, N.R.; Arshad, M.A.; Hayawi, K.; Bevara, R.V.K.; Mannuru, A.; Batool, L. Putting GPT-4o to the Sword: A Comprehensive Evaluation of Language, Vision, Speech, and Multimodal Proficiency. *Appl. Sci.* **2024**, *14*, 7782. <https://doi.org/10.3390/app14177782>

Academic Editor: Andrea Prati

Received: 23 July 2024

Revised: 27 August 2024

Accepted: 28 August 2024

Published: 3 September 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In the past few years, the emergence of large language models has led to paradigm shifts across various disciplines and professions. The pursuit of building and implementing the most powerful and accurate models has captured both researchers and industry. In late 2023 and early 2024, competitors to OpenAI, including Google and Anthropic, introduced advanced large language models: Google's Gemini and Anthropic's Claude 3 [1,2]. These models surpassed the capabilities of the original GPT-3, GPT-3.5, and GPT-4 models that powered ChatGPT. To stay competitive, OpenAI needed to develop an upgraded model with more parameters, enhanced capabilities, and improved speed. This led to the launch of GPT-4 Omni (GPT-4o) in May 2024.

GPT-4o introduces several major innovations that improve upon previous large language models. The model includes a massive number of parameters—estimated to be

well over one trillion—which dwarfs GPT-3, at 175 billion parameters, and GPT-1, at an estimated 117 million parameters [3]. The model is able to process and generate text, image, and audio content and does so at a speed that is much faster than competitor models. Importantly, the model also integrates the improved handling of ambiguous and complex queries, where a misunderstanding could emerge between the user and the model, and enhances its ethical and safety protocols to mitigate the prevalence of harmful or incorrect outputs, as has been an issue with competitor models in recent months [4,5]. Though all these innovations appear to be a tremendous boon for the model, there are many areas where the efficacy of the model has not yet been formally evaluated.

1.1. Research Purpose

The purpose of this study is to comprehensively evaluate the capabilities of GPT-4 Omni (GPT-4o) across various domains, including language, vision, speech, and multi-modal tasks. By systematically assessing GPT-4o's performance on a wide range of benchmarks and real-world tasks, we aim to understand its capabilities, strengths, and limitations. This evaluation will provide insights into the advancements made by GPT-4o compared to previous models, such as GPT-3 and GPT-4, and other contemporary models, like Google's Gemini and Anthropic's Claude 3. These findings will contribute to ongoing investigations of the practical applications and future development of large language models.

1.2. Related Work

GPT-4o is the latest development in a string of innovations to generative pre-trained transformers in recent years. In order to situate the development of GPT-4o within the context of the greater developments occurring in artificial intelligence (AI), it may be helpful to view these technologies as a series of nested boxes, as in Figure 1. AI as a concept encompasses a wide range of developments, of which machine learning and deep learning are but one area [6]. Within deep learning, there are further divisions, with generative AI being only one (albeit major) area. The same is true for large language models, as one application of generative AI. We already know of other types of generative AI that are not language-based, such as image generators. The generative pre-trained transformer is but one large language model (LLM), developed by OpenAI. GPT-4o is the latest version of this model. As such, while GPT-4o is a very important innovation, it is but one element within the broad AI landscape that exists today.

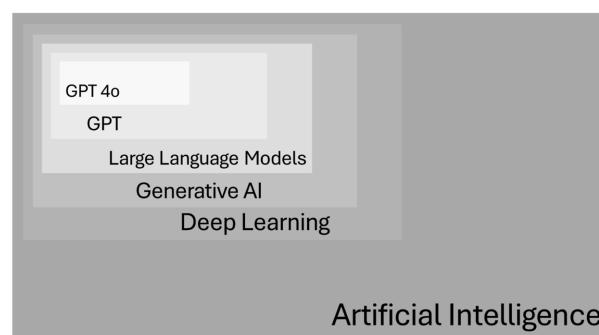


Figure 1. Visualization of the relationship between general AI and GPT-4o.

As illustrated in Figure 1, GPT-4o belongs to the class of technologies known as large language models (LLMs). These models are notable for their ability to mimic human language usage so closely that it can be difficult for a human observer to distinguish between text generated by a human and that generated by a machine [7,8]. This innovation marks a significant advancement towards passing the Turing test and underscores the practicality of AI in writing and research [9,10]. However, it also introduces significant risks, including potential invasions of privacy and the generation of inaccurate, misleading, biased, or harmful information [11]. Therefore, it is crucial to carefully evaluate these

LLMs and scrutinize their outputs. Failure to do so could lead to the proliferation of misinformation and malicious content on the Internet [12].

Given the serious issues associated with some LLMs, it is essential to critically examine each new model for its limitations. Recent versions of GPT have shown significant improvements over their predecessors in various areas. For example, Koubaa (2023) found substantial improvements in GPT-4 compared to GPT-3.5 on tests, such as the Graduate Record Examination (GRE), SAT, and Bar exam, with GPT-4's performance placing it in the top tenth percentile on most of these exams [13]. Similarly, Coyne et al. (2023) reported improvements in grammatical error correction for GPT-4 compared to GPT-3.5 [14]. However, having more parameters in a model does not inherently guarantee better performance on all tasks. Overfitting can occur when a model is extensively trained on a large dataset but fails to generalize well to real-world data [15].

The evaluation of the GPT-4o model is currently very limited. Research has explored various aspects of the model, including the potential threats [16,17], diagnostic ability [18,19], and multilingual capabilities [20]. One study by Sonoda et al. (2024) found that GPT-4o underperforms compared to Claude 3 Opus in radiology diagnosis tasks [21]. Other studies investigated the sentiment of the general public regarding GPT-4o [22]. Many studies focus on ChatGPT, the chatbot powered by GPT models, rather than the models themselves. These studies provide some additional insights into the quality of the models, such as their performance in English language teaching tasks [23]. However, a comprehensive evaluation of GPT-4o itself remains a gap in the literature.

GPT-4o lends itself to new forms of evaluation beyond the language and reasoning evaluation of past model versions due to its new capabilities in vision, speech, and cross-modal activities. GPT-4 with Vision (GPT-4V) was previously evaluated based on vision tasks; however, it is clear from these studies that the model was not ready for the visual challenges to which it was exposed [24,25]. Meanwhile, the speech capacity of GPT-4o is a new innovation, one that has already been met with some criticism due to the choices for the voice of the model [26]. While cross-modal activities have been theorized in LLMs for some time, GPT-4V stands out as among the first models to actualize this potential, paving the way for its evaluation [27].

2. Language Capacity of GPT-4o

Language capacity is foundational to developing intelligent models capable of understanding, generating, and interacting with human language. This capacity encompasses a range of skills that enable models to process and produce coherent and contextually appropriate responses in natural language. The objective of this section is to comprehensively assess the language performance of GPT-4o (omni) by testing it on exams, reasoning tasks, and translation activities. Each of these tasks is significant for evaluating different aspects of the model's language capabilities.

2.1. Performance on Exams

In this subsection, we evaluate GPT-4o's performance on various standardized and board exam questions. This helps us gauge the model's ability to comprehend complex problems and generate coherent, relevant, and accurate responses. Standardized exams are designed to measure a range of cognitive abilities and knowledge across different subjects. This task measures the model's proficiency in handling structured questions across various subjects. Our methods involve presenting GPT-4o with questions from a variety of standardized and board exams. The responses generated by GPT-4o are then analyzed based on the correctness of the answers provided.

2.1.1. Performance on USMLE

The United States Medical Licensing Examination (USMLE) Step 1 is a rigorous and comprehensive assessment designed to evaluate a candidate's understanding and ability to apply key concepts in medical science necessary for the practice of medicine [28]. Jointly

developed by the Federation of State Medical Boards and the National Board of Medical Examiners, this examination serves as a milestone for medical students and professionals aiming to obtain their medical licensure in the United States. The USMLE Step 1 primarily focuses on testing the examinee's grasp of foundational medical knowledge and their ability to apply this knowledge to clinical scenarios. The sample test questions provided in the USMLE Step 1 Sample Items booklet encompass various disciplines, including anatomy, biochemistry, microbiology, pathology, pharmacology, physiology, and interdisciplinary areas, such as genetics, immunology, and molecular biology. The dataset used for evaluating GPT-4o's performance includes 119 sample test questions from the USMLE Step 1 booklet, updated as of January 2024 (https://www.usmle.org/sites/default/files/2021-10/Step_1_Sample_Items.pdf, accessed on 1 February 2024). We utilized all questions in this sample booklet, which we believe to be an exhaustive evaluation of the model's proficiency in this area. This approach of USMLE Step 1 is consistent with previous works [29,30].

Out of the total 118 questions, GPT-4o correctly answered 98 questions. This corresponds to an accuracy of 83.1%. Table 1 provides a comparison of GPT-4o with its predecessor models, as reported by Gilson et al. [30] and Brin et al. [31]. Compared to its predecessor, GPT-3.5, which achieved an accuracy of 51.67%, GPT-4o shows significant improvement. GPT-4o, despite being designed for faster and more efficient tasks, offers a notable enhancement in language comprehension and problem-solving capabilities. However, GPT-4o's performance is slightly lower than that of GPT-4, which achieved an accuracy of 90.00%. This decline can be attributed to the design focus of GPT-4o on efficiency and speed, while GPT-4 remains the model for more complex and demanding tasks.

Table 1. Performance comparison of GPT models on USMLE.

Model	Total Questions	Correct Answers	Accuracy
GPT-3.5	389	201	51.67%
GPT-4	80	72	90.00%
GPT-4o	118	98	83.05%

The results indicate that GPT-4o can serve as a valuable tool in medical education, offering fast, interactive learning experiences that are crucial for students needing immediate feedback and guidance [32]. While GPT-4 excels in handling more intricate questions, its slower response time may limit its practicality for real-time learning scenarios. Meanwhile, GPT-4o's accuracy and efficiency make it suitable for dynamic educational environments.

2.1.2. Performance on CFA

The Chartered Financial Analyst (CFA) Level 1 exam is a globally recognized certification offered by the CFA Institute, aimed at financial and investment professionals (CFA Institute, n.d.). The exam covers a broad range of topics, including ethical and professional standards, quantitative methods, economics, corporate finance, equity investments, fixed income, derivatives, and portfolio management. The CFA Level 1 exam is known for its rigorous and comprehensive assessment of a candidate's foundational knowledge and skills in finance and investment. It tests both theoretical understanding and the practical application of financial concepts and principles.

However, benchmarking a model on the CFA exam presents challenges, as the CFA Institute does not publicly release past exams taken by registered candidates, making it impossible to directly obtain official questions and answers [33]. Additionally, a significant portion of the exam requires written responses that require costly grading by human experts. Hence, for this evaluation, we utilized the dataset from the 300Hours CFA Level 1 Mock Exam, which includes questions developed to mirror the style and difficulty of the actual exam (<https://300hours.com/free-cfa-level-1-mock-exam/>, accessed on 15 June 2024). GPT-4o correctly answered 76 out of the 89 questions, yielding an overall accuracy of 85.39%. Table 2 summarizes the performance in comparison to GPT-3.5 and GPT-4, as reported by Callanan et al. [33]. We compare the results obtained using zero-shot prompting

since we did not provide the models with any hints or specific instructions during our prompting. The results indicate that GPT-4o noticeably outperforms both its predecessors. The increased accuracy of GPT-4o (despite being designed for faster and more efficient tasks) indicates that it can provide reliable and timely assistance for financial exam preparation.

Table 2. Performance comparison of GPT models on CFA Level 1 Exam.

Model	Accuracy
GPT-3.5	58.80%
GPT-4	73.20%
GPT-4o	85.39%

2.1.3. Performance on SAT

The Scholastic Assessment Test (SAT) is a standardized test widely used for college admissions in the United States [34]. Developed and administered by the College Board, the SAT assesses a student's readiness for college and provides colleges with a common data point for comparing all applicants. The SAT covers areas, including reading, writing and language, and mathematics, with an optional essay section. This test is designed to measure a range of skills necessary for academic success in college, including critical thinking, problem-solving, and analytical abilities.

The dataset used for evaluating GPT-4o's performance consists of questions from the SAT Practice Test #1, which includes a variety of reading, writing, and math questions that reflect the format and content of the actual SAT exam (<https://satsuite.collegeboard.org/sat/practice-preparation?excmpid=mtg796-st-1-bk>, accessed on 15 June 2024). The practice test consisted of two modules, each containing a reading and writing exam, as well as a math exam. The performance on each module are outlined in Table 3.

Table 3. GPT-4o performance based on SAT.

Test	Total Questions	Correct Answers	Accuracy
Reading & Writing M1	33	31	93.94%
Reading & Writing M2	33	29	87.88%
Math M1	27	25	92.59%
Math M2	27	22	81.48%

For a comparison with previous GPT models, we refer to the comprehensive report by the Open AI team [35]. In this context, we average the results of M1 and M2 for GPT-4o, as summarized in Table 4.

Table 4. Performance comparison of GPT models on SAT.

Model	Reading & Writing	Math
GPT-3.5	83.75%	73.75%
GPT-4	88.75%	87.50%
GPT-4 (no vision)	88.75%	86.25%
GPT-4o	90.91%	87.04%

GPT-4o demonstrates the highest accuracy in the Reading & Writing section with 90.91%, surpassing all the older models. In the Math section, GPT-4o achieves a strong performance with 87.04%, slightly lower than GPT-4 but higher than the rest. Figures 2–5 provide examples of GPT-4o correct and incorrect responses on each of the SAT categories.

Lucía Michel of the University of Chile observed that alkaline soils contain an insoluble form of iron that blueberry plants cannot absorb, thus inhibiting blueberry growth. If these plants were grown in alkaline soil alongside grasses that aid in iron solubilization, _____ Michel was determined to find out.

B - The blueberries could thrive.

- Explanation: This phrasing correctly completes the sentence as a conditional statement.

Which choice completes the text so that it conforms to the conventions of Standard English?

- could the blueberries thrive.
- the blueberries could thrive.
- the blueberries could thrive?
- could the blueberries thrive?

Choice D is the best answer. The convention being tested is end-of-sentence punctuation. This choice correctly uses a question mark to punctuate the interrogative clause "could the blueberries thrive," which asks a direct question at the end of the sentence.

Figure 2. Example of GPT-4o incorrect answer on SAT Reading & Writing: Question (Top), GPT-4o response (Bottom Left), and correct answer (Bottom Right).

While researching a topic, a student has taken the following notes:

- In the midst of the US Civil War, Susie Taylor escaped slavery and fled to Union-army-occupied St. Simons Island off the Georgia coast.
- She began working for an all-Black army regiment as a nurse and teacher.
- In 1902, she published a book about the time she spent with the regiment.
- Her book was the only Civil War memoir to be published by a Black woman.
- It is still available to readers in print and online.

D - Taylor was the only Black woman to publish a Civil War memoir.

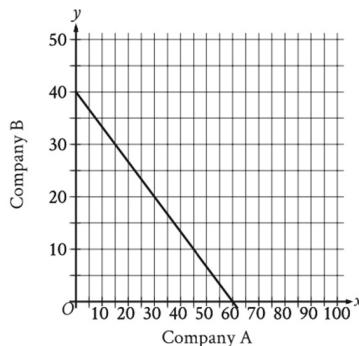
- Explanation: This option emphasizes the uniqueness of Taylor's accomplishment.

The student wants to emphasize the uniqueness of Taylor's accomplishment. Which choice most effectively uses relevant information from the notes to accomplish this goal?

- Taylor fled to St. Simons Island, which was then occupied by the Union army, for whom she began working.
- After escaping slavery, Taylor began working for an all-Black army regiment as a nurse and teacher.
- The book Taylor wrote about the time she spent with the regiment is still available to readers in print and online.
- Taylor was the only Black woman to publish a Civil War memoir.

Choice D is the best answer. By indicating that Taylor's book was the only Civil War memoir published by a Black woman, this sentence emphasizes the uniqueness, or one-of-a-kind nature, of Taylor's accomplishment.

Figure 3. Example of GPT-4o correct answer on SAT Reading & Writing: Question (Top), GPT-4o response (Bottom Left), and correct answer (Bottom Right).



Solution:

Observing the graph, the line has a negative slope, which means as x increases, y decreases. The equation that matches this relationship is in the form of a linear equation that, when solved for y , has a negative coefficient for x .

By rearranging and comparing:

$$8x + 12y = 480 \implies y = \frac{480 - 8x}{12} = 40 - \frac{2}{3}x$$

$$12x + 8y = 480 \implies y = \frac{480 - 12x}{8} = 60 - 1.5x$$

The second equation matches the linear decrease observed in the graph.

Answer: D) $12x + 8y = 480$

The graph shows the relationship between the number of shares of stock from Company A, x , and the number of shares of stock from Company B, y , that Simone can purchase. Which equation could represent this relationship?

- $y = 8x + 12$
- $8x + 12y = 480$
- $y = 12x + 8$
- $12x + 8y = 480$

Choice B is correct. The graph shown is a line passing through the points $(0, 40)$ and $(60, 0)$. Since the relationship between x and y is linear, if two points on the graph make a linear equation true, then the equation represents the relationship. Substituting 0 for x and 40 for y in the equation in choice B, $8x + 12y = 480$, yields $8(0) + 12(40) = 480$, or $480 = 480$, which is true. Substituting 60 for x and 0 for y in the equation $8x + 12y = 480$ yields $8(60) + 12(0) = 480$, or $480 = 480$, which is true. Therefore, the equation $8x + 12y = 480$ represents the relationship between x and y .

Choice A is incorrect. The point $(0, 40)$ is not on the graph of this equation, since $40 = 8(0) + 12$, or $40 = 12$, is not true. **Choice C** is incorrect. The point $(0, 40)$ is not on the graph of this equation, since $40 = 12(0) + 8$, or $40 = 8$, is not true. **Choice D** is incorrect. The point $(0, 40)$ is not on the graph of this equation, since $12(0) + 8(40) = 480$, or $320 = 480$, is not true.

Figure 4. Example of GPT-4o incorrect answer on SAT Math: Question (Top), GPT-4o response (Bottom Left), and correct answer (Bottom Right).

The perimeter of an equilateral triangle is 624 centimeters. The height of this triangle is $k\sqrt{3}$ centimeters, where k is a constant. What is the value of k ?

Solution:

First, find the side length of the equilateral triangle:

$$\text{Perimeter} = 3 \times \text{side length}$$

$$624 = 3 \times \text{side length}$$

$$\text{side length} = \frac{624}{3} = 208 \text{ cm}$$

For an equilateral triangle with side length s , the height h is given by:

$$h = \frac{\sqrt{3}}{2} \times s$$

$$k\sqrt{3} = \frac{\sqrt{3}}{2} \times 208$$

$$k = \frac{208}{2} = 104$$

Answer: 104

The correct answer is 104. An equilateral triangle is a triangle in which all three sides have the same length and all three angles have a measure of 60° . The height of the triangle, $k\sqrt{3}$, is the length of the altitude from one vertex. The altitude divides the equilateral triangle into two congruent 30-60-90 right triangles, where the altitude is the side across from the 60° angle in each 30-60-90 right triangle. Since the altitude has a length of $k\sqrt{3}$, it follows from the properties of 30-60-90 right triangles that the side across from each 30° angle has a length of k and each hypotenuse has a length of $2k$. In this case, the hypotenuse of each 30-60-90 right triangle is a side of the equilateral triangle; therefore, each side length of the equilateral triangle is $2k$. The perimeter of a triangle is the sum of the lengths of each side. It's given that the perimeter of the equilateral triangle is 624; therefore, $2k + 2k + 2k = 624$, or $6k = 624$. Dividing both sides of this equation by 6 yields $k = 104$.

Figure 5. Example of GPT-4o correct answer on SAT Math: Question (Top), GPT-4o response (Bottom Left), and correct answer (Bottom Right).

2.1.4. Performance on MBE

The Multistate Bar Examination (MBE) is a standardized test that assesses the ability of prospective lawyers to apply fundamental legal principles and reasoning to analyze given fact patterns [36]. Developed and administered by the National Conference of Bar Examiners (NCBE), the MBE is a critical component of the bar examination in most U.S. jurisdictions [37]. The MBE includes 200 multiple-choice questions that cover a wide range of legal topics, including constitutional law, contracts, evidence, real property, and torts. The test evaluates the examinee's capacity to think like a lawyer and apply legal knowledge in a practical, problem-solving context.

The dataset for evaluating GPT-4o's performance includes sample test questions from the MBE sample booklet, updated in 2023 (https://www.ncbex.org/sites/default/files/2023-05/MBE_Sample_Test_Questions_New_2023%20.pdf, accessed on 15 June 2024). These questions represent the types and formats of questions that examinees will encounter on the actual MBE, providing a comprehensive overview of the subjects tested and the skills required. In this test, GPT-4o correctly answered 15 out of 20 questions, leading to 75% accuracy. The comparison with previous models is presented in Table 5, based on the results reported by Katz et al. [38]. We note that while this comparison offers valuable insights into the relative performance of GPT models, it may not be entirely fair (as the specific questions used might differ from those evaluated by Katz et al.). Despite this limitation, the comparison still provides a meaningful understanding of GPT-4o's capabilities in relation to earlier models.

Table 5. Performance comparison of GPT models on MBE.

Model	Accuracy
GPT-3.5	45.10%
GPT-4	75.70%
GPT-4o	75.00%

The evaluation results indicate that GPT-4o performs comparably to GPT-4 on the MBE, with a minor difference in accuracy. However, compared to GPT-3.5, which achieved an accuracy of 45.10%, GPT-4o demonstrates a significant improvement. Therefore, law students and bar examinees can benefit from using GPT-4o as an interactive learning tool that provides immediate feedback and explanations, helping them to understand complex legal principles and improve their problem-solving skills.

2.2. Reasoning

Human intellect is remarkably characterized by reasoning, which is described as an activity of methodically and logically thinking about a subject [39]. Reasoning enables humans to come to conclusions or make decisions by using previous experiences and data gathered, thus extending one's knowledge of the world and releasing the possibility for innovation and development. In recent times, AI has made significant advancements in narrowing the gap between human and machine intellect through the use of Natural Language Processing (NLP) and LLMs, which have established remarkable reasoning abilities.

In this section, the authors assess the reasoning capacity of the most recent GPT-4o model via a manual technical evaluation through a sequence of question-answering tasks. The model will answer a range of logical reasoning tasks in different types, including deductive, inductive, and abductive reasoning, as shown in Figure 6. Starting with a broad principle or assumption and applying it to produce predictions or draw conclusions, deductive reasoning takes a top-down method [40]. By contrast, inductive reasoning uses a bottom-up methodology to deduce broad principles or conclusions from observations or data [41]. In abductive reasoning, theories or explanations are developed from little, ambiguous, or incomplete data [42]. With all these assessments based on the model, this article stands to obtain an understanding of the reasoning capacity of the GPT-4o model in various settings.

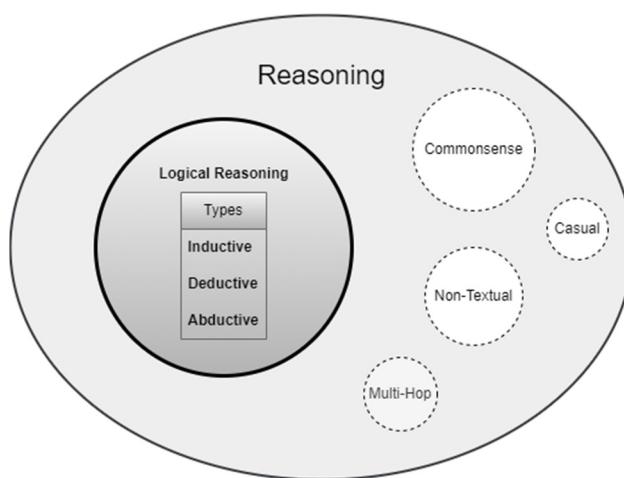


Figure 6. Types of reasoning.

In this subsection, we assess the performance of GPT-4o on five datasets that include all of the aforementioned types of reasoning, as illustrated in Figure 7. To evaluate the deductive reasoning ability, two datasets were utilized, namely EntailmentBank [43] and bAbI (task 15) [44]. Similarly, to assess the capability of inductive reasoning, we employed two datasets, CLUTRR [45] and bAbI (task 16) [44]. For abductive reasoning, we use the α NLI dataset [46]. Adhering to the methods of López Espejel et al. (2023), we randomly selected 30 samples from the test set for both bAbI (task 16) and CLUTRR datasets. The bAbI (task 16) dataset focuses on basic induction, where the model must derive a general rule or principle from a set of given premises. The CLUTRR dataset consists of semi-artificial stories describing hypothetical families, challenging the model to deduce relationships between family members that are not explicitly stated in the text. Using the methods of López Espejel et al. [47] again, our evaluation encompassed the same set of 30 randomly chosen samples from each evaluation dataset. The observations are selected from 10 samples from each of the training-easy, train-medium, and train-hard sets of the α NLI dataset. Concurrently, a total of 30 samples are drawn from the test set for the bAbI (task 15), bAbI (task 16), CLUTRR, and EntailmentBank datasets. In accordance with López Espejel et al. [47], we utilized the identical set of proven and effective prompts that were implemented in their assessment to evaluate the capabilities of the model.

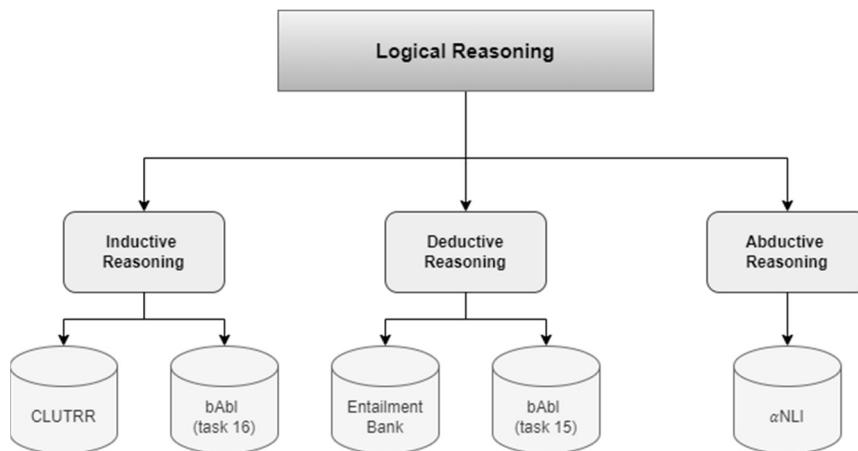


Figure 7. Logical reasoning categories and datasets.

The evaluation results showcase the remarkable reasoning abilities of GPT-4o in all three domains, as indicated in Table 6. GPT-4o demonstrated exceptional performance in deductive reasoning by achieving nearly flawless scores on both bAbI (task 15) and Entailment-Bank. It outperformed ChatGPT-3.5 and performed at the same level as ChatGPT-4 [47]. GPT-4o achieved flawless results in the inductive reasoning tasks, scoring perfectly on bAbI (task 16) and achieving a score of 17 out of 30 on CLUTRR. It outperformed both ChatGPT-3.5 and ChatGPT-4. GPT-4o achieved a score of 27 out of 30 on α NLI, surpassing the performance of its previous versions in abductive reasoning. To ensure a fair comparison, we applied the same evaluation criteria and prompts across all models (GPT-3.5, GPT-4, and GPT-4o), mitigating potential subjective biases in the assessment process. The results underscore GPT-4o's superior reasoning capabilities compared to its predecessors. The model's proficiency in deductive reasoning showcases its ability to derive valid conclusions from premises. Its success in inductive reasoning demonstrates the capacity to generalize from specific facts, while its performance in abductive reasoning highlights the ability to generate credible hypotheses with limited knowledge.

Table 6. Performance of GPT models on logical reasoning tasks.

Category/Model	Deductive Reasoning		Inductive Reasoning		Abductive Reasoning
	Entailment Bank	bAbI (Task 15)	CLUTRR	bAbI (Task 15)	α NLI
GPT-3.5	25/30	26/30	2/30	14/30	19/30
GPT-4	27/30	30/30	11/30	28/30	25/30
GPT-4o	29/30	30/30	17/30	30/30	27/30

Even with the remarkable reasoning powers of GPT-4o, this assessment points to a few drawbacks that need more rigorous investigation. One case was when the model gave different answers to the same topic in various chat sessions while evaluating the bAbI (task 16) dataset for inductive reasoning. Additionally, the model sometimes requested the end-user to choose between different answers. This implies that in some situations, GPT-4o could have trouble with ambiguity, resulting in varying responses. Furthermore, the model sometimes gave different responses when the same subject was posed repeatedly in the same chat session. In determining accuracy, only the first response was considered to maintain synchrony, although this inconsistency raises questions about the model's efficiency and dependability in certain areas. Additionally, we acknowledge that the performance on the bAbI and CLUTRR datasets may vary depending on the specific samples selected, despite our efforts to maintain consistency in the evaluation process. The model's sensitivity to question-wording, the information presentation sequence, or the

existence of unclear or contradictory information in the input may be the causes of these issues. To overcome these problems and enhance its capacity to manage ambiguity and resolve contradictions, future studies should concentrate on creating more reliable and consistent reasoning mechanisms and optimizing prompts for LLMs.

With the notable performance of GPT-4o, there might be more advancements in many AI applications with GPT-4o's improved reasoning skills. High-accuracy complex reasoning tasks performed by it can lead to advancements in information retrieval, decision support tools, and question-answering systems. Still, further study is required to determine how well the model performs on a larger variety of reasoning problems and how well it can manage more intricate and domain-specific reasoning situations. Future research should consider expanding the sample size for evaluations, particularly for datasets like bAbI and CLUTRR, to provide a more comprehensive assessment of model performance across a wider range of reasoning tasks and scenarios.

2.3. Language Translation

Language translation has become an increasingly important task in our globalized world, facilitating communication and understanding across diverse linguistic backgrounds. With the advent of LLMs, like GPT-3.5, GPT-4, Llama, Gemini, and now GPT-4o, the potential for accurate and efficient machine translation has grown significantly [48]. These models, which were trained on massive volumes of multilingual data, can produce translations accurately while capturing the subtle meanings and complexities of different languages. Therefore, in this section, we aim to evaluate the translation proficiency of GPT-4o in six of the most widely spoken languages: Spanish, Arabic, Hindi, French, Portuguese, and Russian.

The choice of these six languages is not arbitrary; they represent a diverse set of linguistic structures and cultural contexts, making them ideal for a comprehensive evaluation of translation capabilities. Spanish is commonly used across Europe and the Americas and is characterized by its straightforward structure and rich vocabulary. Arabic, known for its intricate script and complex word forms, poses distinct challenges for translation technology. Hindi, widely spoken in India, mixes local and foreign words, requiring careful handling to achieve accurate translation. French, spoken in many parts of the world, helps test the model's ability to handle grammatical rules and nuances. Portuguese, similar to Spanish but distinct in several key aspects, allows for an assessment of the model's precision in closely related languages. Lastly, Russian, with its Cyrillic script and case system, provides a test for the model's ability to manage non-Latin scripts and complex grammatical structures.

By focusing on these languages, this study aims to provide a robust and diverse evaluation of GPT-4o's translation performance. Given the widespread use and significant number of native speakers of these languages, improvements in translation accuracy can have a substantial impact on global communication and information dissemination. Hence in this section, we seek to verify GPT-4o's ability to translate across these six languages, providing insights into its potential for breaking down language barriers and facilitating communication among people from different linguistic backgrounds.

2.3.1. Data

The datasets for Spanish, Arabic, French, Portuguese, and Russian were sourced from the OPUS dataset, a well-known collection of texts used for training and evaluating machine translation models [49], and the Hindi dataset was obtained from the IIT Bombay English-Hindi Parallel Corpus, created by the Center for Indian Language Technology (CFILT) at IIT Bombay [50].

For this analysis, 500 data points were randomly sampled from each dataset. The selection of 500 data points is a good balance between feasibility and the need for sufficient data diversity. This sample size is large enough to encompass a wide variety of sentence structures, vocabulary, and translation challenges present in each language, ensuring that

the evaluation is comprehensive and representative. Random sampling was employed to mitigate selection bias and to ensure that the sampled data points provide an unbiased representation of the overall dataset. By using random sampling, this approach captures the natural variability and complexity of language, which is essential for a robust assessment of the GPT-4o model's translation performance across different linguistic contexts.

2.3.2. Evaluation Method

To measure how similar two sentences are in terms of their meaning, an advanced NLP, specifically focusing on sentence embeddings generated by a model called BERT (Bidirectional Encoder Representations from Transformers) [51] and a similarity measure called cosine similarity has been used. BERT is a powerful model that has greatly improved how well computers understand language. For our research, we use a pre-trained model from the sentence-transformers library called paraphrase-MiniLM-L6-v2. This model is specially tuned to understand the meanings and similarities between sentences. It works by turning each sentence into a vector, which is a list of numbers. These vectors or embeddings, encapsulate the semantic information of the sentences in a way that allows for a meaningful comparison between the actual translations and the translations generated by GPT-4o.

To find out how similar two sentences are, we compare their vectors using cosine similarity. Cosine similarity measures the angle between two vectors or embeddings. If the vectors point in the same direction, the sentences are very similar. If they point in completely different directions, the sentences are very different. The values are between -1 and 1 , where

- 1 indicates that the vectors are identical.
- 0 indicates that the vectors are orthogonal (i.e., no similarity).
- -1 indicates that the vectors are opposed.

By calculating the cosine similarity between the embeddings of two sentences, we can effectively measure their semantic similarity. The formula for cosine similarity is

$$\text{Cosine Similarity} = \mathbf{A} \cdot \mathbf{B} / \|\mathbf{A}\| \cdot \|\mathbf{B}\|$$

where \mathbf{A} and \mathbf{B} are the embeddings of the two sentences.

It is important to note that while this computational method provides an efficient and scalable approach to evaluating translation quality, it has limitations. Specifically, it may not fully capture contextual, cultural, and idiomatic aspects of language that human evaluators could identify. The absence of native language expert analysis in our study is a limitation that should be addressed in future research.

While the use of cosine similarity with BERT-based embeddings is an established method in NLP research, it is important to acknowledge the previous work that laid the foundation for this approach. The Sentence-BERT framework, which is utilized through the paraphrase-MiniLM-L6-v2 model, was introduced by Reimers and Gurevych [52].

The novelty of this work lies not in the evaluation method itself but in its comprehensive application across six linguistically diverse languages and the detailed analysis provided for each language pair. By applying this established method to evaluate GPT-4o's translation capabilities across Spanish, Arabic, Hindi, French, Portuguese, and Russian, this study offers unique insights into how this state-of-the-art language model handles the diverse linguistic challenges presented by each language pair.

2.3.3. Results

This study sought to evaluate the capabilities of GPT-4o in translating passages across six major languages: Spanish, Arabic, Hindi, French, Portuguese, and Russian. The results reveal a generally high level of translation accuracy, particularly in Spanish and Portuguese, which scored 88% and 86%, respectively. However, there were notable variations among the languages. Arabic and French, with scores of 78% and 75%, respectively, presented more challenges for the model due to their complex linguistic structures and nuances. Hindi and

Russian scored 82% and 80%, demonstrating the model's competence but also highlighting areas for improvement. The results are summarized in Table 7.

Table 7. GPT-4o translation accuracy across languages.

Language	Translation Accuracy (%)
Spanish	88
Arabic	78
Hindi	82
French	75
Portuguese	86
Russian	80

It is necessary to emphasize that these percentage scores are based on computational comparisons using cosine similarity values, rather than a comprehensive human evaluation. As such, they should be interpreted as indicators of potential performance rather than definitive measures of translation quality across all possible contexts and use cases.

The findings suggest that the line between human and machine translation is becoming increasingly narrow. GPT-4o's performance, though not specifically optimized for translation, approaches the quality of dedicated translation systems. This is particularly noteworthy given the diverse linguistic and structural characteristics of the evaluated languages. While the exact nature of the source translations in the datasets (whether human or machine-translated) is not confirmed, the high similarity scores indicate that GPT-4o is capable of producing translations with a quality that is comparable to the existing translations.

2.3.4. Limitations

However, several limitations must be considered. The random sampling of 500 data points from each dataset may not fully capture the linguistic diversity and complexity of each language. Different samples could yield varying results, suggesting that a larger and more representative dataset might provide a more accurate assessment. Additionally, the reliance on BERT-based embeddings and cosine similarity may not fully encapsulate the nuances of translation quality, particularly in capturing cultural and contextual subtleties. Expanding the dataset size and including more language pairs could yield more comprehensive insights. This research serves as a proof-of-concept for larger-scale studies that could further investigate the capabilities of AI in translation. Future research should focus on incorporating more extensive data, diverse language combinations, and advanced fine-tuning techniques.

Moreover, while the computational methods used in this study, such as cosine similarity with BERT-based embeddings, are efficient and scalable, they do not capture all the nuances of translation quality. This approach, though common in NLP research, may overlook the contextual, cultural, and idiomatic aspects of language that are crucial for accurate translation. The absence of native language expert analysis is another significant limitation. Human evaluation could provide valuable insights into subtle linguistic nuances, cultural appropriateness, and overall fluency that automated metrics might miss. Additionally, the study's focus on general language translation may not fully represent GPT-4o's performance in specialized domains or with more complex linguistic structures. These limitations highlight the need for caution in interpreting the results and underscore the complexity of evaluating machine translation systems across diverse languages.

3. Vision Capacity of GPT-4o

Vision capacity is foundational to developing intelligent models capable of understanding, interpreting, and interacting with visual content. This capacity encompasses a range of skills that enable models to process and produce coherent and contextually appropriate responses to visual inputs. The objective is to comprehensively assess the

vision performance of GPT-4o by testing it on various image-based tasks. Each of these tasks is significant for evaluating different aspects of the model's visual capabilities.

For each task, a dataset of approximately 100 representative images was curated. The model was provided with an image along with a text prompt specifying the desired output format. The prompts were designed to probe the model's ability to identify, classify, describe, and analyze visual content without additional context. For select tasks, we further investigated the model's few-shot learning capabilities by providing a small number of labeled examples before the query image.

Model outputs were compared against ground truth labels to compute standard performance metrics, such as accuracy. Qualitative analysis was also conducted on a subset of responses to identify common failure modes and strengths. The results across different tasks provide insights into GPT-4o's current visual understanding capabilities, areas for improvement, and potential as a foundation model for vision tasks. Subsequent sections discuss the specifics of each task, dataset, and findings, offering a comprehensive evaluation of GPT-4o's visual reasoning skills.

The following subsections evaluate GPT-4o's performance on various image-based tasks, focusing on image classification across different domains and collectively demonstrating the model's capabilities in visual understanding and classification. From fruit identification to medical image analysis, these tasks showcase GPT-4o's versatility in handling diverse visual inputs and classification challenges.

3.1. Image Classification: Fruits Classification

Fruit image classification is crucial for applications in agriculture, supply chain management, and food industry automation. The accurate identification of fruit types can enhance inventory tracking, quality control, and efficient sorting processes [53]. The fruit images dataset (<https://www.kaggle.com/datasets/afsananadia/fruits-images-dataset-object-detection>, accessed on 15 June 2024) consists of approximately 400 images spanning 10 different fruit classes, such as banana, jackfruit, and mango. Each fruit class has 40 labeled images, with the dataset split into 320 training images and 80 test images. The images were collected from various sources, such as Google Images and stock image websites, and were labeled by the dataset creators. For this evaluation, the model was provided with an image along with a prompt to identify the fruit class from the list of 10 classes in a specified format. Model predictions were compared against ground truth labels to assess performance.

The results indicate that GPT-4o performed exceptionally well on this task. The model achieved an average **precision of 0.98**, an average **recall of 0.98**, and an average **F1-score of 0.98**. These metrics suggest that GPT-4o is highly capable of accurately identifying and classifying different fruit images. Table 8 summarizes the performance for each class.

Table 8. GPT-4o performance on fruit classification.

Class	Precision	Recall	F1-Score
Banana	0.91	1.00	0.95
Papaya	1.00	1.00	1.00
Apple	1.00	1.00	1.00
Litchi	1.00	1.00	1.00
Jackfruit	1.00	0.90	0.95
Hog Plum	1.00	1.00	1.00
Grapes	1.00	1.00	1.00
Guava	1.00	1.00	1.00
Mango	0.91	1.00	0.95
Orange	1.00	0.90	0.95

The model demonstrated strong performance in classifying the 10 different fruit classes, achieving high precision, recall, and F1-scores across most classes. Several classes,

including Papaya, Apple, Litchi, Hog Plum, Grapes, and Guava, obtained perfect scores of 1.0 for precision, recall, and F1-score. The Banana and Mango classes had slightly lower but still impressive precision scores (0.91), with a perfect recall of 1.0. Figure 8 presents the confusion matrix and metric visualization for this dataset.

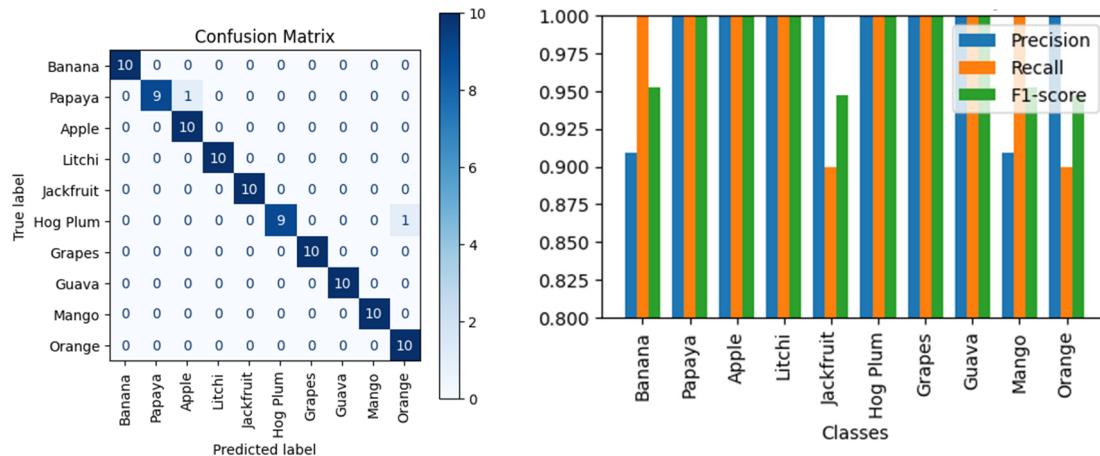


Figure 8. Confusion matrix (left) and performance comparison (right) for fruit classification.

3.2. Image Classification: Driver Drowsiness Detection

Detecting driver drowsiness is critical for enhancing road safety, as the timely identification of fatigue can prevent accidents and save lives. The drowsy detection dataset consists of images extracted from videos capturing drivers in three distinct states: natural, fatigued, and drowsy [54]. The dataset was curated by gathering relevant videos, converting them into image frames, and applying facial detection algorithms to isolate key facial regions like eyes, mouth, and cheeks, which are indicative of drowsiness.

The extracted images were converted to grayscale, resized to 48×48 pixels (reducing computational classification complexity), and accurately labeled based on the driver's state. This resolution was chosen to focus on the most essential visual cues associated with drowsiness, such as subtle changes in the eyes and mouth. While this resolution may seem limited, it is sufficient to capture the necessary facial features and allows for faster processing, making it practical for real-time applications where computational resources may be constrained.

The dataset comprises two classes: drowsy and natural, with a total of 100 labeled images sampled evenly from each class. For this evaluation, GPT-4o was provided with an image along with a prompt to classify it into one of the two classes in a specified JSON format. The model's predictions were compared against the ground truth labels to assess its performance in detecting driver drowsiness from facial features.

In this task, the model achieved an average **precision of 0.80**, an average **recall of 0.80**, and an average **F1-score of 0.80**. Table 9 summarizes the performance for each class.

Table 9. GPT-4o performance on drowsiness detection.

Class	Precision	Recall	F1-Score
Drowsy	0.8	0.8	0.8
Natural	0.8	0.8	0.8

The results indicate that GPT-4o, without fine-tuning, achieves an impressive precision, recall, and F1-score of 0.8. While lower than that of specialized deep learning models like VGG, ResNet, and CNN [54], the performance is impressive given GPT-4o's lack of training on this specific dataset. The notable performance despite no domain-specific training underscores its robustness and adaptability, implying that GPT-4o could be valuable in

scenarios where rapid deployment and flexibility across different tasks are crucial. Figure 9 presents the confusion matrix and metric visualization for this task.

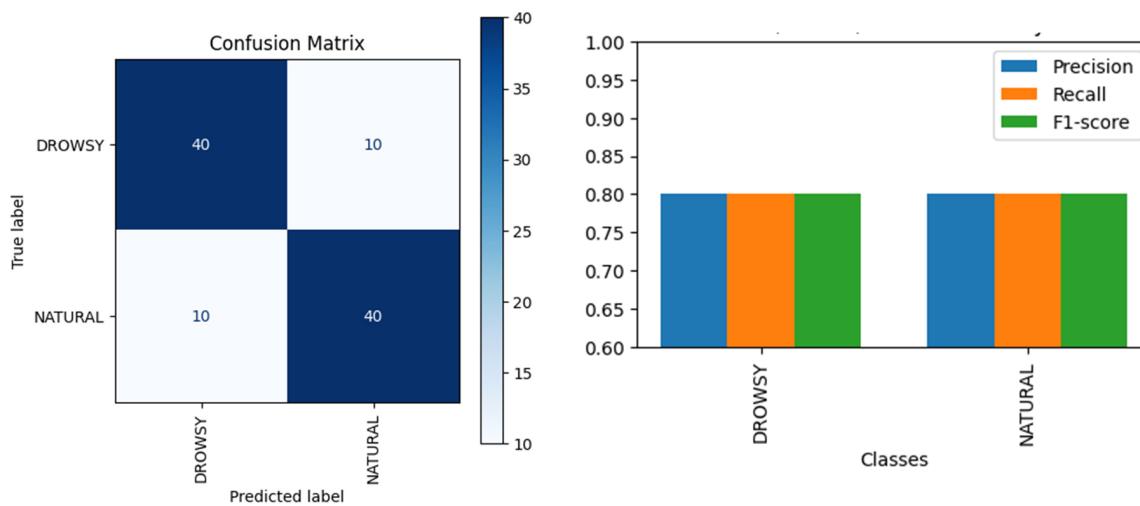


Figure 9. Confusion matrix (left) and performance comparison (right) for drowsiness detection.

3.3. Image Classification: Crop Disease Classification

The accurate identification of crop diseases is essential for ensuring agricultural productivity and preventing significant crop losses. The crop disease classification dataset is a comprehensive collection of images aimed at evaluating GPT-4o's capabilities in identifying crop diseases (<https://www.kaggle.com/datasets/sadikaljarif/crop-disease>, accessed on 15 June 2024). The dataset encompasses 20 distinct classes of common crop diseases, including blight, cedar apple rust, crown gall, and clubroot. For this evaluation, 100 images were randomly sampled from the dataset, with each class represented by approximately five images. GPT-4o was provided with these images along with a prompt to classify the crop disease depicted in each image. The model's predictions were compared against the ground truth labels to assess its performance in accurately identifying and distinguishing various crop diseases based solely on visual information.

The model achieved an average **precision of 0.77**, an average **recall of 0.71**, and an average **F1-score of 0.68** in this task. Table 10 summarizes the performance for each class.

Table 10. GPT-4o performance on crop disease detection.

Class	Precision	Recall	F1-Score
Anthracnose	0.60	0.60	0.60
Apple Scab	1.00	0.80	0.89
Black Spot	0.67	1.00	0.80
Blight	0.38	0.75	0.50
Blossom End Rot	1.00	1.08	1.00
Botrytis	1.00	0.20	0.33
Brown Rot	1.00	0.20	0.33
Canker	0.25	0.25	0.25
Cedar Apple Rust	0.83	1.00	0.91
Clubroot	1.00	1.00	1.00
Crown Gall	1.00	1.00	1.00
Downy Mildew	1.00	0.20	0.33
Fire Blight	0.80	0.80	0.80
Fusarium	1.00	0.60	0.75
Gray Mold	0.43	0.75	0.55
Leaf Spots	0.40	0.80	0.53
Mosaic Virus	0.67	0.80	0.72
Nematodes	1.00	1.00	1.00
Powdery Mildew	0.71	1.00	0.83
Vetricillium	0.67	0.40	0.50

Given the large number of classes (20 classes), this highlights GPT-4o's potential for accurate crop disease classification and adaptability, despite no prior training on this dataset. The limitations in specific classes like Botrytis, Brown rot, and Canker can be attributed to the need for specialized training in certain classes. The confusion matrix and metric visualization for this dataset are presented in Figure 10.

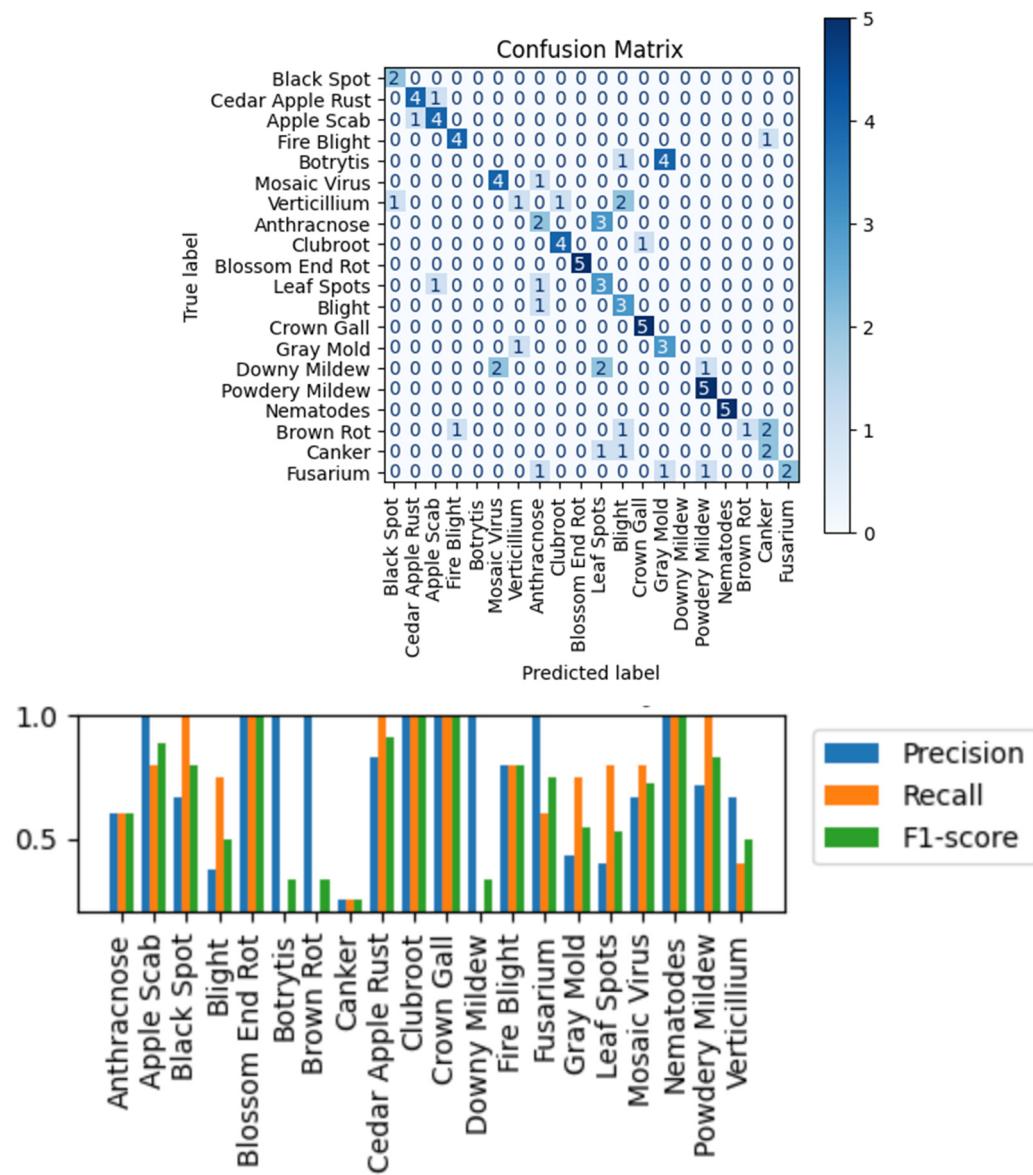


Figure 10. Confusion matrix (top) and performance comparison (bottom) for crop disease classification.

3.4. Image Classification: Glaucoma Detection

The early detection of glaucoma is critical for preventing vision loss and ensuring timely treatment. The glaucoma detection dataset used for this evaluation consisted of retinal fundus images from the ACRIMA database (<https://www.kaggle.com/datasets/chetanpediredla/glaucoma-dataset>, accessed on 15 June 2024). A subset of 100 images was sampled, evenly split between glaucomatous and normal cases. These images were collected at FISABIO Oftalmología Médica in Valencia, Spain, and were annotated by experienced glaucoma experts. GPT-4o was tasked with classifying each image into either glaucoma or normal based solely on the visual information provided. The model's predictions were compared against the expert-annotated ground truth labels to assess its performance in detecting glaucoma from retinal fundus imagery.

As shown in Table 11, GPT-4o achieved an average **precision of 0.65**, an average **recall of 0.62**, and an average **F1-score of 0.59**. For the glaucoma class, the model demonstrated a precision of 0.58, a recall of 0.86, and an F1-score of 0.69. In contrast, the normal class had a higher precision of 0.73 but a significantly lower recall of 0.38, resulting in an F1-score of 0.50.

Table 11. GPT-4o performance on glaucoma detection.

Class	Precision	Recall	F1-Score
Glaucoma	0.58	0.86	0.69
Normal	0.73	0.38	0.50

The confusion matrix in Figure 11 reveals that the model correctly identified 42 out of 49 glaucoma cases but struggled more with normal cases, correctly classifying only 19 out of 50. The plot shows the model's relatively balanced precision and recall for the glaucoma class but highlights a pronounced discrepancy for the normal class, with precision substantially higher than recall. GPT-4o is effective at identifying glaucomatous images but has difficulty in correctly classifying normal cases.

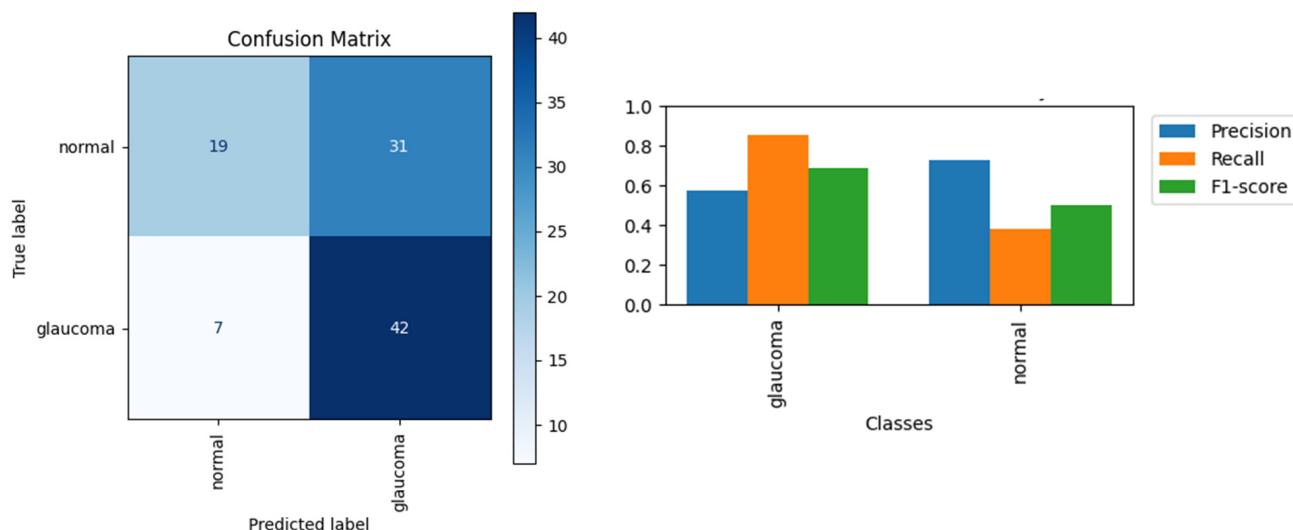


Figure 11. Confusion matrix (left) and performance comparison (right) for glaucoma detection.

Few-shot learning allows models to make accurate predictions with only a small number of training examples. This approach is particularly beneficial in scenarios where data are scarce. Figure 12 illustrates the F1 scores for both classes across different numbers of shots, indicating how the model's performance evolves with the number of examples provided during training. The glaucoma class maintains a relatively high F1 score across all shot levels, showing slight improvement with additional examples. This consistency suggests that GPT-4o effectively learns to identify glaucomatous features even with a limited number of examples. In contrast, the normal class exhibits significant improvement in the F1-score from zero shots to one shot but then plateaus. This indicates that while the initial provision of examples significantly enhances the model's ability to recognize normal cases, further increases in the number of examples yield diminishing returns.

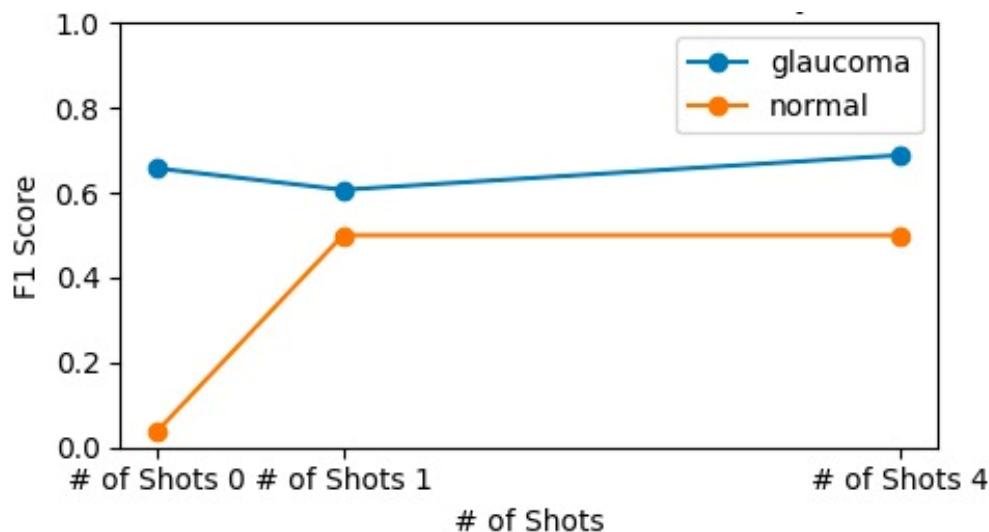


Figure 12. Performance evolution against number of shots for glaucoma detection.

3.5. Image Classification: Cancer, Tumor, and Aneurysm Detection

The accurate detection and classification of brain conditions, such as cancer, tumors, and aneurysms are crucial for timely diagnosis and treatment. The computed tomography (CT) brain scan dataset contains CT images of the brain aimed at detecting and classifying various conditions, such as cancer, tumors, and aneurysms (<https://www.kaggle.com/datasets/trainingdatapro/computed-tomography-ct-of-the-brain>, accessed on 15 June 2024). For this evaluation, a subset of 100 CT scan images was sampled from the dataset. GPT-4o was tasked with analyzing these images and classifying them into one of three categories: cancer, tumor, or aneurysm. The model's predictions were compared against the ground truth labels to assess its performance in identifying these medical conditions from CT brain imagery.

GPT-4o achieved an average **precision of 0.21**, an average **recall of 0.32**, and an average **F1-score of 0.26**. Table 12 summarizes the performance for each class.

Table 12. GPT-4o performance on cancer, tumor, and aneurysm detection.

Class	Precision	Recall	F1-Score
Aneurysm	0.3725	0.5758	0.4524
Cancer	0.0000	0.0000	0.0000
Tumor	0.2708	0.3939	0.3210

The confusion matrix in Figure 13 reveals that the model completely failed to predict the “cancer” class, potentially due to a lack of representative training data or inherent similarities with other classes. Additionally, it struggled to distinguish between “aneurysm” and “tumor” classes, with significant misclassifications in both directions, suggesting a need for further fine-tuning or the incorporation of additional relevant features.

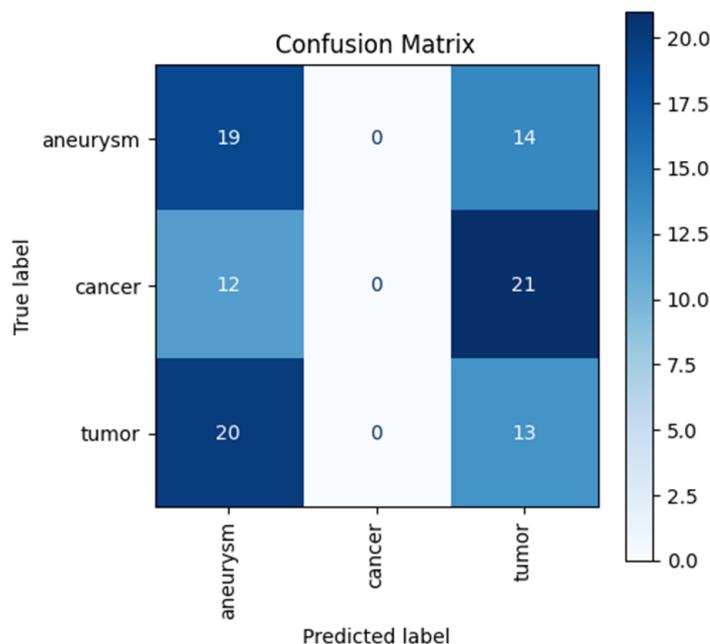


Figure 13. Confusion matrix for cancer, tumor, and aneurysm detection task.

It is important to note that this evaluation was conducted to assess the model's technical ability to process and classify medical images, not to provide clinical diagnoses. The model's predictions were compared against the ground truth labels to assess its performance in identifying these medical conditions from CT brain imagery. However, we emphasize that image detection alone is not sufficient for diagnosis, and any results from this evaluation should be interpreted with caution and in conjunction with an expert medical evaluation. This study does not suggest that the model is suitable for clinical use, and further comprehensive research would be required before considering such an application.

3.6. Image Captioning

The Flickr8k captions dataset, consisting of 8000 images from Flickr with multiple human-annotated captions, was used for this task (<https://www.kaggle.com/datasets/aladdinpersson/flickr8kimagescaptions>, accessed on 15 June 2024). We randomly sampled 100 images from this dataset. GPT-4o was tasked with generating a single-line caption for each image using a zero-shot learning approach, meaning no examples were provided to the model. The prompt instructed the model to "write a short caption for it in a very short single line" and to format the output as a JSON object. This zero-shot method tests the model's ability to comprehend visual scenes and translate them into natural language descriptions without prior examples. To evaluate performance, we compared the model-generated captions against the ground truth human captions using the BLEU score, a standard metric for measuring similarity between machine- and human-generated texts. This approach assesses GPT-4o's capacity to produce accurate and coherent textual descriptions of visual content without specific training on the task. The resulting BLEU scores are summarized in Table 13.

Table 13. GPT-4o performance on image captioning.

Metric	Score
BLEU-1	0.193
BLEU-2	0.095
BLEU-3	0.058
BLEU-4	0.031

With a BLEU-1 score of 0.193, the model demonstrates a moderate ability to capture the essence of the captions with a reasonable degree of similarity in individual words. However, as the n-gram length increases, the scores decline significantly (BLEU-2: 0.095, BLEU-3: 0.058, BLEU-4: 0.031), indicating that the model struggles with maintaining coherence and context in longer sequences. This highlights the challenges GPT-4o faces in generating more complex and accurate descriptions. The results show that GPT-4o has a foundational understanding of visual scenes, but there is room for improvement in generating detailed and contextually rich captions.

4. Speech Capacity of GPT-4o

Speech capacity evaluates the ability of intelligent models to understand, interpret, and interact with auditory content. This encompasses a range of skills that enable models to process and produce coherent and contextually appropriate responses to audio inputs. The objective is to assess the audio performance of GPT-4o by testing it on various audio-based tasks. Each of these tasks is significant for evaluating different aspects of the model's auditory capabilities.

Unlike text-based evaluations, which are well-established and extensively explored in works, such as Hung and Alias [55], the evaluation of audio-based tasks in models like GPT-4o presents unique challenges due to the lack of standardized methodologies and comprehensive datasets. This work extends the evaluation of GPT-4o beyond traditional language tasks. However, we also acknowledge that this area of evaluation is still in its early stages, and the results should be viewed as a preliminary step towards more comprehensive assessments of speech capabilities in large language models.

4.1. Emotion Detection

Emotion detection is a critical aspect of understanding human communication, as the same speech can convey different meanings depending on the emotional tone in which it is expressed [56]. Recognizing emotions in speech is essential for applications ranging from customer service to mental health monitoring. For this evaluation, we used the Arabic natural audio dataset (ANAD) from Kaggle, designed to detect discrete emotions in Arabic speech (<https://www.kaggle.com/datasets/suso172/arabic-natural-audio-dataset/data>, accessed on 15 June 2024). The ANAD consists of 1384 audio recordings, each labeled with one of three emotions: happy, angry, or surprised. These recordings were sourced from live Arabic talk shows, where each video was labeled by 18 listeners to determine the perceived emotion. To evaluate the emotion-detection capabilities of GPT-4o, we randomly sampled 100 audio files from the ANAD dataset. Each audio file was fed to the model along with a prompt to predict the emotion class. The model's predictions were then compared against the ground truth labels to assess its performance.

The results of the emotion detection task, as illustrated in Figure 14, reveal that GPT-4o demonstrates variable performance across different emotion classes. The confusion matrix shows that the model performs best for the "surprised" class, correctly predicting 21 instances, but it frequently misclassifies "happy" as "surprised" (19 times). The "angry" class has the lowest true positive rate with only two correct predictions, often being mistaken for "happy" or "surprised." The model has the highest recall for the "surprised" class, indicating it correctly identifies "surprised" emotions more frequently than others. The precision for "angry" is reasonably high, but the recall is very low, meaning that while it predicts "angry" correctly when it does so, it rarely predicts "angry" overall. The "happy" class has moderate precision and recall, suggesting a balanced but moderate performance in predicting this class.

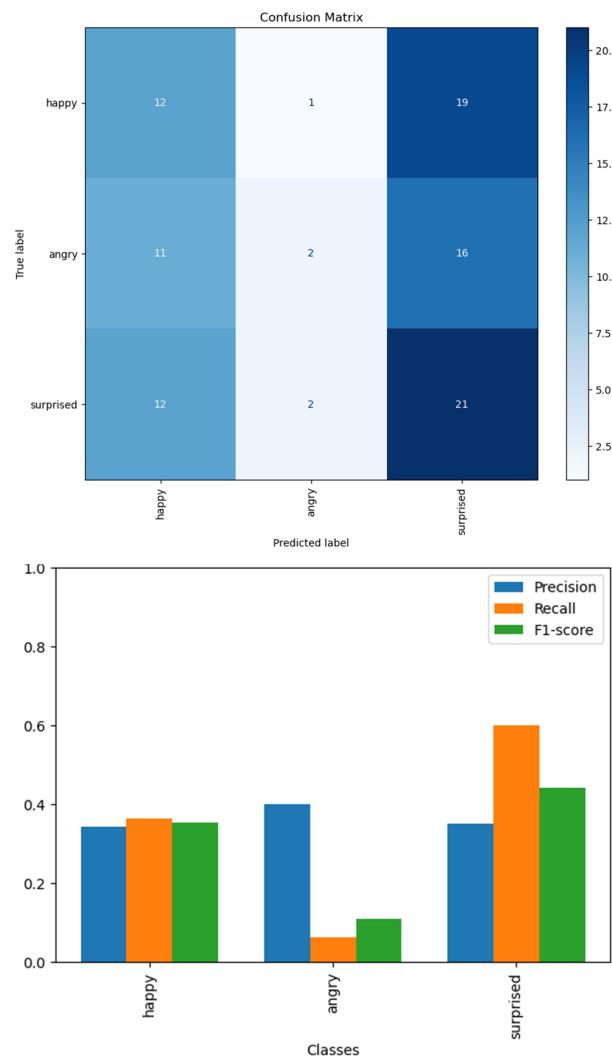


Figure 14. Confusion matrix (**top**) and performance comparison (**bottom**) for audio emotion detection.

4.2. Accent Detection

Accents play a crucial role in speech recognition, affecting the accuracy and efficiency of automatic speech recognition (ASR) systems. Understanding and detecting accents is essential for developing robust ASR systems that can handle diverse linguistic backgrounds [57]. For this evaluation, we utilized the AccentDB dataset, a comprehensive collection of non-native English accents designed to assist neural speech-recognition tasks [58].

The AccentDB dataset includes samples from various Indian-English accents and native English accents, providing a diverse range of phonetic and prosodic variations. It contains speech recordings from speakers with distinct linguistic backgrounds, such as Bangla, Malayalam, Odiya, and Telugu, alongside metropolitan Indian accents and native accents from American, Australian, British, and Welsh English. The dataset is structured to meet key requirements for ASR development, including a variety of speakers, uniformity of content, and well-labeled data for training and testing models. To assess the accent-detection capabilities of GPT-4o, we randomly selected 100 audio files from the AccentDB dataset. Each file was presented to the model with a prompt to identify the speaker's accent. The predictions made by GPT-4o were then compared to the ground truth labels to evaluate their performance.

The confusion matrix in Figure 15 highlights significant misclassifications, particularly with the Malayalam accent, which is frequently misclassified as Telugu. This misclassifica-

tion suggests that the acoustic features of Malayalam and Telugu might be similar enough to confuse the model, indicating a need for more distinctive feature extraction and training data augmentation. Bangla and Telugu also exhibit substantial misclassification errors, particularly in Malayalam. This pattern suggests a broader challenge in differentiating between the phonetic characteristics of these languages, necessitating further refinement in the model's training process. The precision, recall, and F1-score metrics provide additional insights into the model's performance across different classes. The model demonstrates the highest precision for Odiya, indicating that when it predicts Odiya, it is often correct. However, the low recall for Odiya means that many Odiya instances are not being correctly identified. Malayalam shows a more balanced performance with relatively higher recall and F1-scores, suggesting that the model can correctly identify Malayalam instances more frequently. Both Bangla and Telugu have consistently low precision, recall, and F1-scores, indicating significant challenges in accurately detecting these accents. This demonstrates GPT-4o's limited ability to recognize and differentiate between various English accents, which is essential for enhancing the usability of ASR systems in multilingual and multicultural environments.

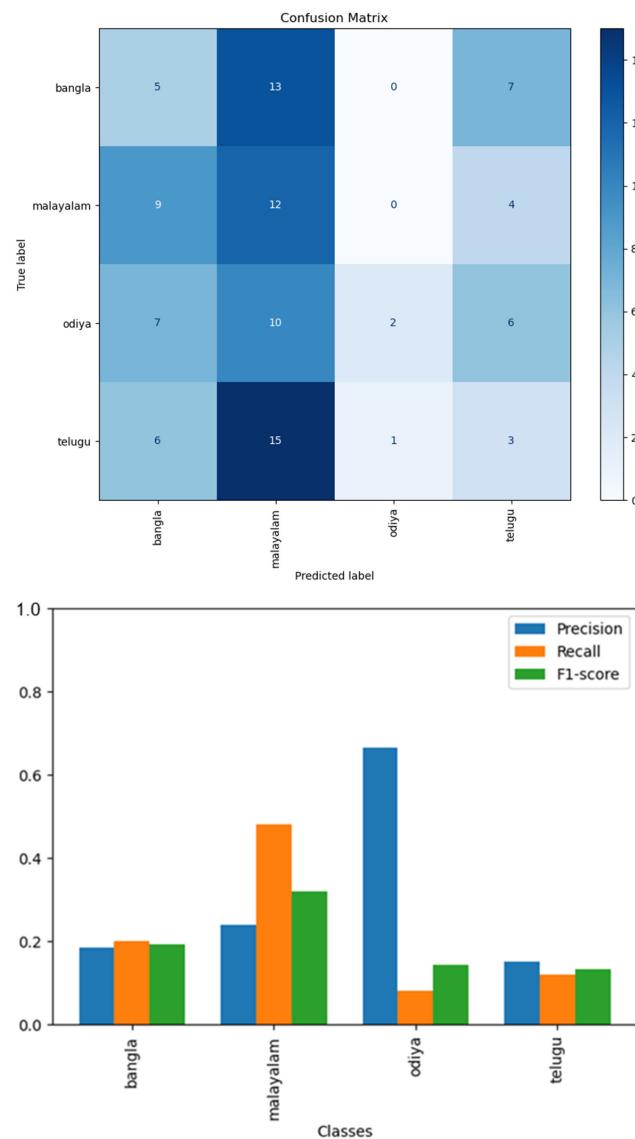


Figure 15. Confusion matrix (Top) and performance comparison (Bottom) for accent detection.

5. Multimodal Capacity of GPT-4o

The ability to integrate and interpret information from multiple modalities is crucial for developing advanced intelligent systems. Multimodal capacity refers to the capability of a model to understand and synthesize information from various sources, such as text, images, and audio. This enables the model to generate more comprehensive and contextually enriched responses. The objective of assessing GPT-4o's multimodal capacity is to evaluate its performance across tasks that require the integration of different types of data.

5.1. Visual Question Answering

The Visual Question Answering (VQA) dataset is a multimodal benchmark that combines computer vision and NLP tasks. It consists of images paired with natural language questions related to the visual content (<https://www.kaggle.com/datasets/bhavikardeshna/visual-question-answering-computer-vision-nlp>, accessed on 15 June 2024). The goal is to produce accurate natural language answers by comprehending the semantics of both the image and the question. For this evaluation, a subset of 100 image-question pairs was sampled from the dataset. GPT-4o was tasked with analyzing the provided image and the corresponding question and generating an appropriate answer chosen from a predefined list of possible answers. The model's generated answers were compared against the ground truth answers to assess its performance in this AI-complete task, which involves a wide range of sub-problems, such as object detection, scene classification, and multimodal reasoning. The maximum accuracy was 0.36, as shown in Figure 16.

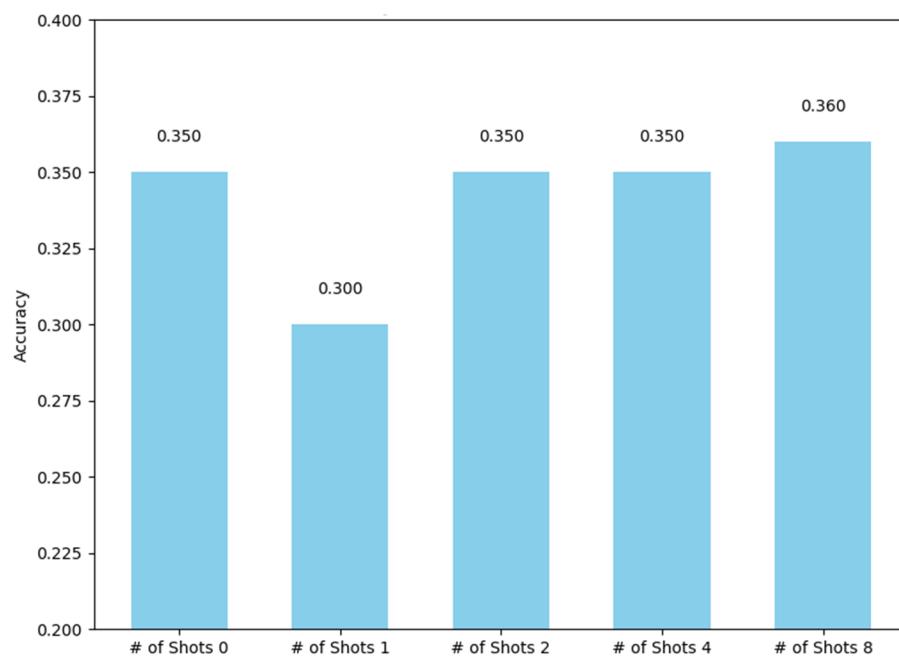


Figure 16. Accuracy by numbers of shots in visual question answering (VQA).

The performance shows some variability with different shot numbers, peaking at 0.36 accuracy with eight shots. Interestingly, the model's performance decreases with just one example, suggesting that providing few examples in a task with many options may not always be beneficial. This decrease in performance could be due to the distribution of answers becoming skewed by the unrelated task, given the diverse possibilities in VQA.

5.2. Vision-Language Capabilities

Vision-language (VL) capabilities represent a critical advancement in the development of AI models that can understand and interpret multimodal data, integrating both visual and linguistic information to perform complex tasks. The ability to combine these two

types of data allows for a more nuanced understanding of content, which is essential for applications ranging from image captioning to more sophisticated tasks like explaining visual jokes or reasoning about events depicted in images.

To evaluate the vision-language capabilities of GPT-4o, we employed the MM-Vet benchmark [59]. MM-Vet is designed to systematically assess large multimodal models (LMMs) on a variety of integrated tasks that require a combination of core VL capabilities, including recognition, optical character recognition (OCR), knowledge, language generation, spatial awareness, and math. This evaluation framework ensures a comparison across diverse question types and answer styles and provides insights beyond simple performance rankings.

The MM-Vet benchmark includes tasks that necessitate the integration of these capabilities to solve complex problems. For instance, a task might involve recognizing objects in an image, understanding the spatial relationships between them, reading and interpreting text within the image, and generating a coherent textual response that incorporates external knowledge. The evaluation metrics employed by MM-Vet are based on an LLM-based evaluator that uses few-shot learning to provide scores for open-ended model outputs. This approach allows for a consistent and comprehensive evaluation across different answer styles and question types. We compare the performance of GPT-4o with its predecessors in Table 14.

Table 14. Performance comparison of GPT models on VL capabilities.

Model	Recognition	OCR	Knowledge	Language Generation	Spatial Awareness	Math	Total
GPT-3.5	22.3	31.4	15.6	16.6	32.9	24.0	27.6
GPT-4	34.3	66.3	25.6	36.6	60.6	72.0	48.1
GPT-4V	67.5	68.3	56.2	60.7	69.4	58.6	67.7
GPT-4o	80.6	87.5	75.7	77.1	84.7	88.1	83.9

The results from the MM-Vet benchmark highlight the advancements made with GPT-4o in VL capabilities compared to its predecessors. As summarized in Table 14, GPT-4o outperforms previous models across all evaluated metrics. GPT-4o's notable performance across all metrics highlights its advanced VL capabilities, setting a new benchmark for multimodal models. The high scores in knowledge, spatial awareness, and language-generation tasks, in particular, highlight GPT-4o's ability to understand and produce contextually relevant responses based on visual inputs, making it versatile in various applications. The high performance noted is also consistent with that reported by Zhu et al. [60]. Table 15 provides examples of GPT-4o responses for several images and prompts.

Table 15. Instances of GPT-4o responses with visual inputs (incorrect answers highlighted in red).

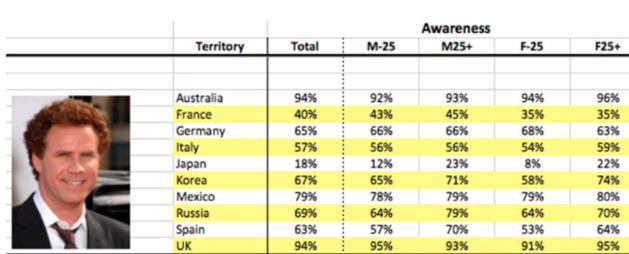
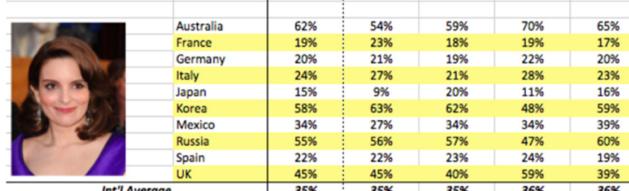
Input Image	Questions Assessed Capabilities	GPT-4o Responses
	(a) Q: How many gallons of supreme gasoline can I get with \$50? (b) Q: What is the price difference between diesel and regular gasoline?	GPT-4o: (a) With \$50, you can get approximately 13.70 gallons of supreme gasoline. GPT-4o: The price difference between diesel and regular gasoline is \$0.11 per gallon.
	Required capabilities: OCR, Math	
	(a) Q: Based on the image, what is the total international awareness of Tina Fey? (b) Q: Based on the image, in which country does Tina Fey have the highest level of total awareness?	GPT-4o: Based on the image, the total international awareness of Tina Fey is 35%. This percentage is located at the bottom row under the column labeled "Total" for Tina Fey.
	Required capabilities: OCR, Spatial awareness	GPT-4o: Based on the image, Tina Fey has the highest level of total awareness in Australia, with an awareness level of 62%.

Table 15. Cont.

Input Image	Questions Assessed Capabilities	GPT-4o Responses
	<p>(a) Q: What is the price for tomatoes? Required capabilities: Recognition, Spatial awareness</p> <p>(b) Q: What fruit is the right of plums?</p>	<p>GPT-4o: The price for tomatoes is 8.00, as indicated by the sign placed above the pile of tomatoes in the image.</p> <p>GPT-4o: The fruit to the right of the plums in the image is pears.</p>
	<p>(a) Q: Which car is on the parking spot 33? Required capabilities: OCR, Recognition, Spatial awareness</p> <p>(b) Q: Which spot is the red car parking on?</p>	<p>GPT-4o: The parking spot labeled 33 is empty.</p> <p>GPT-4o: The red car is parked in spot number 30.</p>

Table 15. Cont.

Input Image	Questions Assessed Capabilities	GPT-4o Responses																														
<p>Plastic Waste is Out of Control In 2015, humans threw out almost as much plastic as they made.</p> <table border="1"> <thead> <tr> <th>Category</th> <th>Plastic made (metric tonnes)</th> <th>Plastic thrown out (metric tonnes)</th> </tr> </thead> <tbody> <tr><td>Industrial Machines</td><td>3M</td><td>1M</td></tr> <tr><td>Electronics</td><td>18M</td><td>13M</td></tr> <tr><td>Transportation</td><td>27M</td><td>17M</td></tr> <tr><td>Consumer Products</td><td>42M</td><td>37M</td></tr> <tr><td>Other</td><td>47M</td><td>38M</td></tr> <tr><td>Textiles</td><td>59M</td><td>42M</td></tr> <tr><td>Construction</td><td>65M</td><td>13M</td></tr> <tr><td>Packaging</td><td>146M</td><td>141M</td></tr> <tr><td>Total</td><td>407M</td><td>302M</td></tr> </tbody> </table> <p>In 2015, humans produced 407 million metric tons of new plastic, continuing the trend of making more and more plastic each year. Meanwhile, 302 million metric tons was discarded as waste, meaning it wasn't recycled or incinerated; it's just sitting in a dump or the ocean somewhere.</p>	Category	Plastic made (metric tonnes)	Plastic thrown out (metric tonnes)	Industrial Machines	3M	1M	Electronics	18M	13M	Transportation	27M	17M	Consumer Products	42M	37M	Other	47M	38M	Textiles	59M	42M	Construction	65M	13M	Packaging	146M	141M	Total	407M	302M	<p>(a) Q: Which category did the worst job in plastic recycling, considering the amount of plastic made and thrown out?</p> <p>(b) Q: What is the difference in metric tonnes between the amount of plastic produced and the amount discarded?</p> <p>Required capabilities: OCR, Spatial awareness</p>	<p>GPT-4o: The category that did the worst job in plastic recycling, considering the amount of plastic made and thrown out, is "Packaging". In 2015, 146 million metric tonnes of plastic were made for packaging, and 141 million metric tonnes were thrown out. This shows that almost all the plastic made for packaging was discarded as waste.</p> <p>The difference between the amount of plastic produced (407 million metric tonnes) and the amount discarded (302 million metric tonnes) is:</p> $407 \text{ million metric tonnes} - 302 \text{ million metric tonnes} = 105 \text{ million metric tonnes}$ <p>So, the difference is 105 million metric tonnes.</p>
Category	Plastic made (metric tonnes)	Plastic thrown out (metric tonnes)																														
Industrial Machines	3M	1M																														
Electronics	18M	13M																														
Transportation	27M	17M																														
Consumer Products	42M	37M																														
Other	47M	38M																														
Textiles	59M	42M																														
Construction	65M	13M																														
Packaging	146M	141M																														
Total	407M	302M																														
	<p>Q: What does this factory make?</p> <p>Required capabilities: OCR</p>	<p>GPT-4o: The factory in the image belongs to Leica, which is a company known for manufacturing cameras, lenses, and other optical equipment.</p>																														

Table 15. Cont.

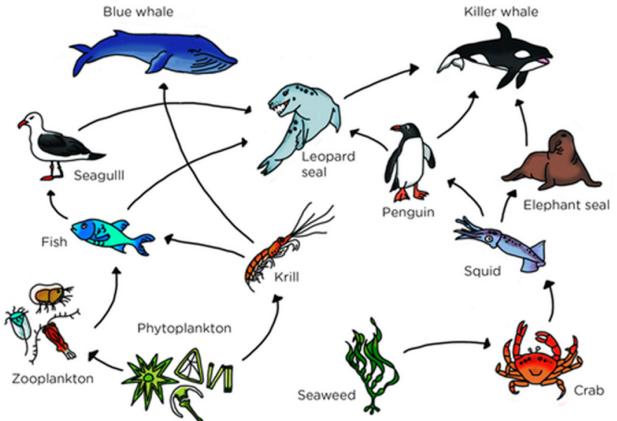
Input Image	Questions Assessed Capabilities	GPT-4o Responses
	<p>Q: In which country was this photo taken?</p> <p>Required capabilities: Recognition, Knowledge</p>	<p>GPT-4o: The photo shows Novak Djokovic kissing a trophy that appears to be the Australian Open trophy. Therefore, this photo was likely taken in Australia.</p>
	<p>Q: Which are producers in this food web?</p> <p>Required capabilities: OCR, Knowledge, Spatial Awareness</p>	<p>GPT-4o: In this food web, the producers are:</p> <p>Phytoplankton Seaweed</p> <p>These are the organisms that produce energy through photosynthesis and serve as the base of the food web.</p>

Table 15. Cont.

Input Image	Questions Assessed Capabilities	GPT-4o Responses
	<p>Q: Is it summertime in the scene? Answer it and give the rationale.</p> <p>Required capabilities: Recognition, Knowledge, Language generation</p>	<p>GPT-4o: No, it is not summertime in the scene. The presence of large pumpkins suggests that it is likely fall, as pumpkins are typically associated with the autumn harvest season. Additionally, the characters are wearing long-sleeved clothing, which is more common in cooler weather.</p>
<p>Me: I'll do it at 8 Time: 8.05 Me: looks like I gotta wait till 9 now</p>  <p>boredpanda.com</p>	<p>Q: Can you explain this meme?</p> <p>Required capabilities: Recognition, OCR, Knowledge, Language generation</p>	<p>GPT-4o: This meme humorously illustrates the common tendency to procrastinate. The text shows an intention to start a task at a specific time (8:00), but when the time is slightly missed (8:05), the person decides to delay the task further, waiting until the next hour (9:00) to start. The image of Kermit the Frog relaxing on a bed adds to the humor by depicting the relaxed attitude towards procrastination. The meme pokes fun at the irrational yet relatable habit of postponing tasks due to missing an arbitrary starting time by a few minutes.</p>

Table 15. Cont.

Input Image	Questions Assessed Capabilities	GPT-4o Responses
	<p>Q: What is funny about this image?</p> <p>Required capabilities: Recognition, Knowledge, Language generation</p>	<p>GPT-4o: This image is a popular meme that combines two unrelated photos to create a humorous juxtaposition. On the left side, there is an image of a woman (Taylor Armstrong from "The Real Housewives of Beverly Hills") angrily pointing and shouting, while another woman (Kyle Richards) tries to calm her down. On the right side, there is a photo of a white cat (known as "Smudge the Cat") sitting at a dinner table with a confused and displeased expression in front of a plate of vegetables. The humor comes from the contrast between the woman's intense emotion and the cat's indifferent and bemused reaction, as if the cat is unbothered by the woman's outburst. This combination creates a relatable and funny depiction of exaggerated reactions and misunderstandings.</p>
	<p>Q: How to cook this dish?</p> <p>Required capabilities: Recognition, Knowledge, Language generation</p>	<p>GPT-4o: This image shows a pie with a lattice crust. Here is a basic recipe to make a lattice-top pie, such as a strawberry or mixed berry pie:</p> <p>Ingredients: For the Filling: 4 cups fresh strawberries or mixed berries, 3/4 cup granulated sugar, 1/4 cup cornstarch, 1 tablespoon lemon juice, 1 teaspoon vanilla extract, A pinch of salt For the Crust: 2 1/2 cups all-purpose flour, 1 teaspoon salt, 1 teaspoon sugar, 1 cup unsalted butter, chilled and cut into small pieces, 1/4 to 1/2 cup ice water</p> <p>Instructions: In a large bowl, combine the flour, salt, and sugar. Add the chilled butter pieces and cut them into the flour mixture using a pastry blender or your fingers until the mixture resembles coarse crumbs. Gradually add ice water, a tablespoon at a time, mixing until the dough begins to come together. Do not overwork the dough. Divide the dough into two equal parts, shape them into disks, wrap in plastic wrap, and refrigerate for at least 1 h. In a large bowl, combine the berries, sugar, cornstarch, lemon juice, vanilla extract, and salt. Stir gently to combine. On a lightly floured surface, roll out one disk of dough to fit a 9-inch pie pan. Place the rolled-out dough into the pie pan and trim the edges. Pour the berry mixture into the crust-lined pie pan. Roll out the second disk of dough and cut it into 1-inch-wide strips. Lay 4 to 5 strips over the filling, parallel to each other. Lay another 4 to 5 strips perpendicular to the first set, weaving them to create a lattice pattern. Trim and crimp the edges to seal. Preheat the oven to 375°F (190 °C). Bake: Place the pie on a baking sheet to catch any drips and bake for 45–55 min, or until the crust is golden brown and the filling is bubbly. Cool: Let the pie cool completely before serving to allow the filling to set. Enjoy your homemade lattice-top pie!</p>

Table 15. *Cont.*

Input Image	Questions Assessed Capabilities	GPT-4o Responses
 A historical Chinese painting titled "Emperor Taizong Receiving the Tibetan Envoy". It depicts Emperor Taizong of the Tang Dynasty standing on the left, wearing a red robe and a black hat. He is facing a group of figures on the right who are dressed in traditional Tibetan attire, including a man in a white robe and a woman in a red robe. In the background, there is a large green and gold canopy. Several red seals are visible on the top left of the painting.	<p>Q: Can you give a short introduction to this painting?</p> <p>Required capabilities: Recognition, Knowledge, Language generation</p>	<p>GPT-4o: This painting is a historical Chinese artwork titled "Emperor Taizong Receiving the Tibetan Envoy", created by Yan Liben during the Tang Dynasty. The piece illustrates a significant diplomatic event where Emperor Taizong of the Tang Dynasty meets a Tibetan envoy. This event highlights the political and cultural exchanges between the Tang Dynasty and Tibet. The painting is notable for its detailed depiction of the figures, the hierarchical arrangement, and the ceremonial aspects of the meeting. Yan Liben, a renowned court artist of the Tang Dynasty, is celebrated for his ability to capture the grandeur and solemnity of imperial events. This artwork not only serves as a historical record but also exemplifies the artistic sophistication of the Tang Dynasty.</p>

6. Implications, Limitations, and Future Work

This section summarizes the key implications of our findings, acknowledges the limitations of the study, and outlines potential directions for future research.

6.1. Implications

The findings from this research have significant implications for the development and application of LLMs in various fields. GPT-4o's high performance in tasks like medical exam question answering and financial analysis suggests its potential utility in educational and professional training environments. The model's ability to integrate vision and language data effectively positions it as a valuable tool in fields requiring multimodal analysis, such as healthcare, finance, and customer service. The demonstrated proficiency in few-shot learning highlights the model's potential for applications where data are scarce or expensive. This could lead to more accessible AI-driven solutions in underrepresented languages and domains, offering inclusivity and the broader application of AI technologies.

Furthermore, enhancing the integration of language, vision, and speech capabilities in GPT-4o can significantly improve its performance in complex multimodal tasks and cross-domain interactions. Current methods of combining these modalities can be refined through the use of advanced multimodal fusion techniques. For instance, attention-based mechanisms [61] can dynamically weigh the importance of each modality depending on the task context, helping the model better respond to complex scenarios. Additionally, cross-modal training [62], where the model is exposed to tasks that require the simultaneous processing of text, images, and audio, can further enhance its multimodal interactions.

Moreover, the need to evaluate newer models on comprehensive and diverse sets of data and tasks is underscored by this research. The gap in robust and extensive evaluations has been a notable limitation in understanding the full capabilities and potential weaknesses of advanced models like GPT-4o. This calls for the development and adoption of more comprehensive benchmarks that can rigorously test models across a wider array of real-world scenarios. The findings also suggest implications for policy and regulatory frameworks. As AI models become increasingly integrated into critical sectors, such as healthcare and finance, ensuring their reliability, transparency, and fairness becomes necessary [63]. This necessitates continuous monitoring, rigorous testing, and the establishment of standards to guide the ethical deployment of AI technologies.

6.2. Limitations

Despite the promising results presented in this study, several limitations must be acknowledged. Firstly, the evaluation datasets used in various tasks, particularly in image and audio data, were relatively small and not exhaustive. This limited sample size may not fully capture the model's performance across all potential scenarios. It may also result in some limitations in standardized comparisons among models. While we aimed for a comprehensive evaluation across data types and multimodal (breadth), the categories within each are not exhaustive (depth). For example, we did not evaluate image and audio generation as it was beyond the scope of this study.

Moreover, qualitative or human judgment was not used as a criterion to assess performance. Incorporating human judgment is crucial for evaluating the practical usability and contextual accuracy of model outputs, as it provides insights that quantitative metrics alone may not reveal [64]. The model also exhibited inconsistencies in handling ambiguous or complex inputs, as seen in the varying accuracy rates across different tasks. Furthermore, the few-shot learning approach, although beneficial in some contexts, showed limitations in tasks with a high degree of variability, such as VQA. The potential for overfitting to specific examples in these cases remains a concern. Additionally, the lack of real-time and longitudinal data evaluation poses a constraint on understanding the model's adaptability and robustness over time. For example, evaluating the model's performance in real-time applications, such as continuously monitoring driver drowsiness or detecting sudden

changes in patient health through medical imaging, would provide valuable insights into its practical effectiveness and reliability under dynamic conditions.

GPT-4o exhibits strong few-shot learning capabilities across various tasks, as evidenced by its performance in glaucoma detection. However, for tasks with high variability, such as VQA, the model's performance can be inconsistent. In some cases, providing just one example led to decreased accuracy, suggesting that few-shot learning may not always be beneficial in highly variable contexts. To enhance GPT-4o's adaptability and generalization, future research should focus on developing more sophisticated few-shot learning techniques that can better handle diverse and complex scenarios. This may include exploring dynamic prompt engineering, meta-learning approaches, or incorporating task-specific knowledge to guide the model's few-shot learning process more effectively.

6.3. Future Work

Building on the existing research, this paper highlights several avenues for future research directions. Expanding the evaluation datasets to include a more diverse and comprehensive range of tasks will provide a deeper understanding of the model's capabilities and limitations. Integrating real-time and longitudinal data assessments can offer insights into the model's adaptability and performance stability over extended periods. Further refinement of the few-shot learning techniques is essential, especially for tasks with high variability. Exploring advanced prompting strategies and incorporating a more contextual understanding could enhance performance in these areas [65]. It is thus important to also investigate the impact of prompt quality on model performance. Additionally, understanding the reasons behind the model's low performance and conducting thorough error analysis are crucial. This involves examining how and why the model failed in specific tasks to inform targeted training and fine-tuning efforts. Such analysis will provide valuable insights into the model's limitations and guide improvements to enhance its utility in nuanced language-understanding tasks.

Future work should also prioritize creating and adopting new, comprehensive benchmarks that evaluate models across diverse tasks and datasets, addressing the current gap in robust model evaluation. This approach will ensure a holistic understanding of the model's performance, guiding improvements and encouraging the development of more reliable AI systems. The current multimodal evaluation only investigated image and text inputs, highlighting the necessity to explore other inputs and their combinations. For instance, incorporating audio, image, and text together could significantly contribute to cross-domain applications and arts [66], enhancing the model's utility in various fields. Lastly, incorporating qualitative assessments and human judgment in the evaluation process will provide a more nuanced understanding of the model's practical applicability and contextual performance. This can help identify areas where the model performs well in real-world scenarios and where it may require further enhancement.

7. Conclusions

In this comprehensive evaluation of GPT-4o's capabilities across language, vision, speech, and multimodal tasks, we have demonstrated the model's substantial improvements over its predecessors, particularly in multimodal integration and few-shot learning. The results reveal that GPT-4o excels in structured and well-defined tasks, such as medical exam question answering and financial analysis, showcasing its potential in professional and educational applications. However, the study also highlights limitations, particularly in tasks requiring high variability and real-time processing, such as visual question answering and certain audio-based evaluations. The findings prove the need for further research and development, particularly in expanding dataset diversity, refining few-shot learning techniques, and incorporating real-time and longitudinal assessments.

Author Contributions: Conceptualization, S.S.; methodology, all authors.; investigation, all authors; writing—original draft preparation, all authors; writing—review and editing, B.D.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gemini Team; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A.M.; Hauth, A.; Millican, K.; et al. Gemini: A Family of Highly Capable Multimodal Models. *arXiv* **2024**, arXiv:2312.11805.
2. Korinek, A. *Language Models and Cognitive Automation for Economic Research*; National Bureau of Economic Research: Cambridge, MA, USA, 2023.
3. Floridi, L.; Chiriaci, M. GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds Mach.* **2020**, *30*, 681–694. [[CrossRef](#)]
4. Dillion, D.; Mondal, D.; Tandon, N.; Gray, K. Large Language Models as Moral Experts? GPT-4o Outperforms Expert Ethicist in Providing Moral Guidance; OSF: Peoria, IL, USA, 2024. [[CrossRef](#)]
5. Ray, S. Google CEO Says Gemini AI's 'Unacceptable' Responses Offended Users and Showed Bias. 2024. Available online: <https://www.forbes.com/sites/siladityaray/2024/02/28/google-ceo-says-gemini-ais-unacceptable-responses-offended-users-and-showed-bias/?sh=250e1a1b1103> (accessed on 15 June 2024).
6. Ongsulee, P. Artificial intelligence, machine learning and deep learning. In Proceedings of the 2017 15th International Conference on ICT and Knowledge Engineering (ICT&KE), Bangkok, Thailand, 22–24 November 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–6.
7. Thirunavukarasu, A.J.; Ting, D.S.J.; Elangovan, K.; Gutierrez, L.; Tan, T.F.; Ting, D.S.W. Large language models in medicine. *Nat. Med.* **2023**, *29*, 1930–1940. [[CrossRef](#)]
8. Hayawi, K.; Shahriar, S. AI Agents from Copilots to Coworkers: Historical Context, Challenges, Limitations, Implications, and Practical Guidelines. *Preprints* **2024**. [[CrossRef](#)]
9. Aher, G.V.; Arriaga, R.I.; Kalai, A.T. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. In Proceedings of the 40th International Conference on Machine Learning, PMLR, Honolulu, HI, USA, 23–29 July 2023; pp. 337–371. Available online: <https://proceedings.mlr.press/v202/aher23a.html> (accessed on 15 June 2024).
10. Mannuru, N.R.; Shahriar, S.; Teel, Z.A.; Wang, T.; Lund, B.D.; Tijani, S.; Pohboon, C.O.; Agbaji, D.; Alhassan, J.; Galley, J.; et al. Artificial intelligence in developing countries: The impact of generative artificial intelligence (AI) technologies for development. *Inf. Dev.* **2023**, 0266669231200628. [[CrossRef](#)]
11. Lund, B.D.; Wang, T.; Mannuru, N.R.; Nie, B.; Shimray, S.; Wang, Z. ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. *J. Assoc. Inf. Sci. Technol.* **2023**, *74*, 570–581. [[CrossRef](#)]
12. Hu, B.; Sheng, Q.; Cao, J.; Shi, Y.; Li, Y.; Wang, D.; Qi, P. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 26–27 February 2024; Volume 38, pp. 22105–22113.
13. Koubaa, A. GPT-4 vs. GPT-3.5: A Concise Showdown. *Preprints* **2023**. [[CrossRef](#)]
14. Coyne, S.; Sakaguchi, K.; Galvan-Sosa, D.; Zock, M.; Inui, K. Analyzing the Performance of GPT-3.5 and GPT-4 in Grammatical Error Correction. *arXiv* **2023**, arXiv:2303.14342.
15. Salman, S.; Liu, X. Overfitting Mechanism and Avoidance in Deep Neural Networks. *arXiv* **2019**, arXiv:1901.06566.
16. Shen, X.; Wu, Y.; Backes, M.; Zhang, Y. Voice Jailbreak Attacks Against GPT-4o. *arXiv* **2024**, arXiv:2405.19103.
17. Ying, Z.; Liu, A.; Liu, X.; Tao, D. Unveiling the Safety of GPT-4o: An Empirical Study using Jailbreak Attacks. *arXiv* **2024**, arXiv:2406.06302.
18. Kalyanpur, A.; Saravanakumar, K.; Barres, V.; Chu-Carroll, J.; Melville, D.; Ferrucci, D. LLM-ARC: Enhancing LLMs with an Automated Reasoning Critic. *arXiv* **2024**, arXiv:2406.17663.
19. Zhang, N.; Sun, Z.; Xie, Y.; Wu, H.; Li, C. The latest version ChatGPT powered by GPT-4o: What will it bring to the medical field? *Int. J. Surg.* **2024**. [[CrossRef](#)]
20. Wang, H.; Xu, J.; Xie, S.; Wang, R.; Li, J.; Xie, Z.; Zhang, B.; Xiong, C.; Chen, X. M4U: Evaluating Multilingual Understanding and Reasoning for Large Multimodal Models. *arXiv* **2024**, arXiv:2405.15638.
21. Sonoda, Y.; Kurokawa, R.; Nakamura, Y.; Kanzawa, J.; Kurokawa, M.; Ohizumi, Y.; Gonoi, W.; Abe, O. Diagnostic Performances of GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro in "Diagnosis Please" Cases. *medRxiv* **2024**, 2024.05.26.24307915. [[CrossRef](#)]
22. Singgalen, Y.A. Analyzing an Interest in GPT 4o through Sentiment Analysis using CRISP-DM. *J. Inf. Syst. Inform.* **2024**, *6*, 882–898. [[CrossRef](#)]
23. Pang, S.; Nol, E.; Heng, K. ChatGPT-4o for English language teaching and learning: Features, applications, and future prospects. *SSRN Sch. Pap.* **2024**, 4837988. [[CrossRef](#)]
24. Xu, S.; Wang, Y.; Liu, D.; Xu, C. Collage Prompting: Budget-Friendly Visual Recognition with GPT-4V. *arXiv* **2024**, arXiv:2403.11468.

25. Zhou, Y.; Ong, H.; Kennedy, P.; Wu, C.C.; Kazam, J.; Hentel, K.; Flanders, A.; Shih, G.; Peng, Y.; Moy, L.; et al. Evaluating GPT-4V (GPT-4 with Vision) on Detection of Radiologic Findings on Chest Radiographs. *Radiology* **2024**, *311*, e233270. [CrossRef] [PubMed]
26. Allyn, B. Scarlett Johansson Says She Is “Shocked, Angered” over New ChatGPT Voice. Available online: <https://www.npr.org/2024/05/20/1252495087/openai-pulls-ai-voice-that-was-compared-to-scarlett-johansson-in-the-movie-her> (accessed on 15 June 2024).
27. Li, H.; Ding, W.; Kang, Y.; Liu, T.; Wu, Z.; Liu, Z. CTAL: Pre-training cross-modal transformer for audio-and-language representations. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 3966–3977.
28. Federation of State Medical Boards and National Board of Medical Examiners. USMLE Step 1 Content Description and General Information. 2024. Available online: <https://www.usmle.org> (accessed on 15 June 2024).
29. Kung, T.H.; Cheatham, M.; Medenilla, A.; Sillos, C.; Leon, L.D.; Elepafio, C.; Madriaga, M.; Aggabao, R.; Diaz-Candido, G.; Maningo, J.; et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digit. Health* **2023**, *2*, e0000198. [CrossRef] [PubMed]
30. Gilson, A.; Safranek, C.W.; Huang, T.; Socrates, V.; Chi, L.; Taylor, R.A.; Chartash, D. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Med. Educ.* **2023**, *9*, e45312. [CrossRef] [PubMed]
31. Brin, D.; Sorin, V.; Vaid, A.; Soroush, A.; Glicksberg, B.S.; Charney, A.W.; Nadkarni, G.; Klang, E. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Sci. Rep.* **2023**, *13*, 16492. [CrossRef]
32. Haleem, A.; Javaid, M.; Qadri, M.A.; Suman, R. Understanding the role of digital technologies in education: A review. *Sustain. Oper. Comput.* **2022**, *3*, 275–285. [CrossRef]
33. Callanan, E.; Mbakwe, A.; Papadimitriou, A.; Pei, Y.; Sibue, M.; Zhu, X.; Ma, Z.; Liu, X.; Shah, S. Can GPT models be Financial Analysts? An Evaluation of ChatGPT and GPT-4 on mock CFA Exams. *arXiv* **2023**, arXiv:2310.08678.
34. College Board. The SAT Suite of Assessments. Available online: <https://www.collegeboard.org> (accessed on 15 June 2024).
35. OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; et al. GPT-4 Technical Report. *arXiv* **2024**, arXiv:2303.08774.
36. National Conference of Bar Examiners. MBE Sample Test Questions. Available online: <https://www.ncbex.org> (accessed on 15 June 2024).
37. Griggs, M. Building a Better Bar Exam. *Tex. A&M Law Rev.* **2019**, *7*, 1.
38. Katz, D.M.; Bommarito, M.J.; Gao, S.; Arredondo, P. Gpt-4 passes the bar exam. *Philos. Trans. R. Soc. A* **2024**, *382*, 20230254. [CrossRef]
39. Huang, J.; Chang, K.C.-C. Towards Reasoning in Large Language Models: A Survey. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023, Toronto, ON, Canada, 9–14 July 2023; Rogers, A., Boyd-Graber, J., Okazaki, N., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 1049–1065.
40. Johnson-Laird, P. Deductive reasoning. *WIREs Cogn. Sci.* **2010**, *1*, 8–17. [CrossRef]
41. Hayes, B.K.; Heit, E.; Swendsen, H. Inductive reasoning. *WIREs Cogn. Sci.* **2010**, *1*, 278–292. [CrossRef]
42. Walton, D. *Abductive Reasoning*; University of Alabama Press: Tuscaloosa, AL, USA, 2014.
43. Dalvi, B.; Jansen, P.; Tafjord, O.; Xie, Z.; Smith, H.; Pipatanangkura, L.; Clark, P. Explaining Answers with Entailment Trees. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, 7–11 November 2021; Moens, M.-F., Huang, X., Specia, L., Yih, S.W., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 7358–7370. [CrossRef]
44. Weston, J.; Bordes, A.; Chopra, S.; Rush, A.M.; van Merriënboer, B.; Joulin, A.; Mikolov, T. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. *arXiv* **2015**, arXiv:1502.05698.
45. Sinha, K.; Sodhani, S.; Dong, J.; Pineau, J.; Hamilton, W.L. CLUTRR: A Diagnostic Benchmark for Inductive Reasoning from Text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; Inui, K., Jiang, J., Ng, V., Wan, X., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 4506–4515. [CrossRef]
46. Bhagavatula, C.; Bras, R.L.; Malaviya, C.; Sakaguchi, K.; Holtzman, A.; Rashkin, H.; Downey, D.; Yih, W.; Choi, Y. Abductive Commonsense Reasoning. International Conference on Learning Representations. 2019. Available online: <https://openreview.net/forum?id=Byg1v1HKDB> (accessed on 15 June 2024).
47. López Espejel, J.; Ettifouri, E.H.; Yahaya Alassan, M.S.; Chouham, E.M.; Dahhane, W. GPT-3.5, GPT-4, or BARD? Evaluating LLMs reasoning ability in zero-shot setting and performance boosting through prompts. *Nat. Lang. Process. J.* **2023**, *5*, 100032. [CrossRef]
48. Khoshafah, F. ChatGPT for Arabic-English Translation: Evaluating the Accuracy. *Res. Sq.* **2023**. [CrossRef]
49. Tiedemann, J. Parallel data, tools and interfaces in OPUS. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12), Istanbul, Turkey, 23–25 May 2012; pp. 2214–2218.
50. Kunchukuttan, A.; Mehta, P.; Bhattacharyya, P. The IIT bombay english-hindi parallel corpus. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018.
51. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Volume 1 (long and short papers), pp. 4171–4186.

52. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; Inui, K., Jiang, J., Ng, V., Wan, X., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 3982–3992. [CrossRef]
53. Cubero, S.; Aleixos, N.; Moltó, E.; Gómez-Sanchis, J.; Blasco, J. Advances in machine vision applications for automatic inspection and quality evaluation of fruits and vegetables. *Food Bioprocess Technol.* **2011**, *4*, 487–504. [CrossRef]
54. Jebraeily, Y.; Sharafi, Y.; Teshnehlab, M. Driver drowsiness detection based on convolutional neural network architecture optimization using genetic algorithm. *IEEE Access* **2024**, *12*, 45709–45726. [CrossRef]
55. Hung, L.P.; Alias, S. Beyond Sentiment Analysis: A Review of Recent Trends in Text Based Sentiment Analysis and Emotion Detection. *J. Adv. Comput. Intell. Intell. Inform.* **2023**, *27*, 84–95. [CrossRef]
56. Shahriar, S. GAN computers generate arts? A survey on visual arts, music, and literary text generation using generative adversarial network. *Displays* **2022**, *73*, 102237. [CrossRef]
57. Graham, C.; Roll, N. Evaluating OpenAI’s Whisper ASR: Performance analysis across diverse accents and speaker traits. *JASA Express Lett.* **2024**, *4*, 025206. [CrossRef] [PubMed]
58. Ahamed, A.; Anand, A.; Bhargava, P. AccentDB: A database of non-native english accents to assist neural speech recognition. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; European Language Resources Association: Marseille, France, 2020; pp. 5351–5358. Available online: <https://www.aclweb.org/anthology/2020.lrec-1.659> (accessed on 15 June 2024).
59. Yu, W.; Yang, Z.; Li, L.; Wang, J.; Lin, K.; Liu, Z.; Wang, X.; Wang, L. MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities. *arXiv* **2023**, arXiv:2308.02490.
60. Zhu, N.; Zhang, N.; Shao, Q.; Cheng, K.; Wu, H. OpenAI’s GPT-4o in surgical oncology: Revolutionary advances in generative artificial intelligence. *Eur. J. Cancer* **2024**, *206*, 114132. [CrossRef]
61. Zhu, H.; Wang, Z.; Shi, Y.; Hua, Y.; Xu, G.; Deng, L. Multimodal Fusion Method Based on Self-Attention Mechanism. *Wirel. Commun. Mob. Comput.* **2020**, *2020*, 8843186. [CrossRef]
62. Zhou, K.; Hassan, F.H.; Hoon, G.K. The State of the Art for Cross-Modal Retrieval: A Survey. *IEEE Access* **2023**, *11*, 138568–138589. [CrossRef]
63. Hayawi, K.; Shahriar, S.; Mathew, S.S. The imitation game: Detecting human and AI-generated texts in the era of ChatGPT and BARD. *J. Inf. Sci.* **2024**, *01655515241227531*. [CrossRef]
64. Shahriar, S.; Al Roken, N.; Zualkernan, I. Classification of Arabic poetry emotions using deep learning. *Computers* **2023**, *12*, 89. [CrossRef]
65. Sivarajkumar, S.; Kelley, M.; Samolyk-Mazzanti, A.; Visweswaran, S.; Wang, Y. An Empirical Evaluation of Prompting Strategies for Large Language Models in Zero-Shot Clinical Natural Language Processing: Algorithm Development and Validation Study. *JMIR Med. Inform.* **2024**, *12*, e55318. [CrossRef]
66. Shahriar, S.; Al Roken, N. How can generative adversarial networks impact computer generated art? Insights from poetry to melody conversion. *Int. J. Inf. Manag. Data Insights* **2022**, *2*, 100066. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.