

Brief Report

Not peer-reviewed version

---

# A Cross-Domain Performance Report of Open AI ChatGPT o1 Model

---

[Kadhim Hayawi](#)<sup>\*</sup> and [Sakib Shahriar](#)<sup>\*</sup>

Posted Date: 23 December 2024

doi: 10.20944/preprints202412.1930.v1

Keywords: Artificial intelligence; large language models; model evaluation; LLM benchmarks; OpenAI o1; AGI



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

*Brief Report*

# A Cross-Domain Performance Report of Open AI ChatGPT o1 Model

Kadhim Hayawi <sup>1,\*</sup> and Sakib Shahriar <sup>\*</sup>

<sup>1</sup> AICEO, Arjaan Office Tower 302, AL Sufouh Complex, Tecom, Dubai, United Arab Emirates

<sup>\*</sup> Correspondence: kadhim.hayawi@ceoai.ai | Abdul.Hayawi@zu.ac.ae

**Abstract:** Large language models (LLMs) represent a leap in the capabilities of artificial intelligence (AI) in natural language understanding, problem-solving, and domain-specific reasoning. Comparative and cross-domain evaluations of LLMs can help us understand their versatility and limitations, including real-world applicability. The o1 model developed by OpenAI represents a notable milestone in terms of state-of-the-art integration into the aspects of language processing and task execution. This report investigates the o1 (o1-preview) model on various tasks, including but not limited to mathematics, clinical knowledge, professional ethics, and the humanities. The results revealed that the o1 excels in certain areas, particularly in fields requiring specialized knowledge, such as college biology (98%) and clinical knowledge (93%). In comparison, it shows lower performance in areas like professional law (54%) and business ethics (81%).

**Keywords:** Artificial intelligence; large language models; model evaluation; LLM benchmarks; OpenAI o1; AGI

## 1. Introduction

In recent years, large language models (LLMs) have emerged as a transformative technology in artificial intelligence, marked by breakthroughs such as BERT, GPT-3, and GPT-4 ([1–3]). These models have surpassed traditional applications and demonstrated competency in domains ranging from academic research to professional assistance. GPT-4, for example, obtained human-level performance on professional benchmarks, such as language translation and standardized tests like SATs [3]. With over 175 billion parameters, GPT-4 can easily handle various tasks; consequently, LLMs offer exciting possibilities, such as enabling individuals to interact more easily with computers and revolutionizing fields such as medicine and finance.

Generative AI (Gen AI) represents a transformative paradigm shift of our generation as it unlocks unprecedented opportunities for innovation. Gen AI can revolutionize industries by equipping individuals with information and specialized skills that were once accessible to only a few, thus helping to maximize human potential. However, the need for responsible AI governance becomes important. From a risk and control perspective, it is essential to address potential challenges related to privacy, safety, and security, ensuring that the benefits of Gen AI are realized ethically and responsibly [4]. At the heart of Gen AI are language models, sophisticated computational systems designed to understand and generate human language. These models form the backbone of many natural language processing (NLP) applications, enabling machines to interpret and produce text in ways that closely mimic human communication. Language models predict the likelihood of word sequences to generate coherent text based on patterns learned from vast datasets; these models have become instrumental in advancing technologies across education, healthcare, creative industries, and beyond ([5–7]).

The advancements in Gen AI and LLMs represent a leap toward the broader vision of Artificial General Intelligence (AGI). AGI would constitute a computing system capable of understanding and performing tasks across diverse domains with human-like adaptability [8]. The transition from LLMs and Gen AI to AGI hinges on the ability of these models to demonstrate comprehensive, cross-

domain expertise, enabling them to handle complex, interdisciplinary challenges seamlessly [9]. Achieving AGI requires a rigorous evaluation that assesses not only specialized capabilities but also the ability to generalize across fields. In this context, cross-domain evaluation becomes essential for understanding the readiness and limitations of current models. This study addresses this gap by comprehensively evaluating the OpenAI o1 model across diverse domains, providing critical insights into its performance, and guiding its trajectory toward AGI.

## 2. Open AI O1: Overview and Capabilities

OpenAI o1 models are new LLMs trained with reinforcement learning to perform complex reasoning [10]. O1 models "think" before they answer and can generate a detailed internal chain of thought before responding to the user. These models excel in scientific reasoning, ranking in the 89th percentile on competitive programming questions (Codeforces), placing among the top 500 students in the US in a qualifier for the USA Math Olympiad (AIME), and exceeding human PhD-level accuracy on a benchmark covering physics, biology, and chemistry problems (GPQA) [11]. While o1 models offer significant advancements in reasoning, they are not intended to replace GPT-4o in all use cases. O1 builds on this legacy but introduces improvements in scalability, reasoning, and domain-specific knowledge. Unlike its predecessors, o1 is designed to excel across a broader range of tasks while maintaining alignment with human intent. Key differentiators include its refined training methodology, optimized architectures, and an emphasis on domain diversity in evaluation [10].

### 2.1. Reasoning Capabilities of O1

The o1 models introduce reasoning tokens, which enable the models to "think" by breaking down their understanding of a prompt and considering multiple approaches to generating a response. These reasoning tokens are generated internally and serve as a tool for the model to analyze the input thoroughly. Once the reasoning process is complete, the model produces an answer in the form of visible completion tokens while discarding the reasoning tokens from its context to optimize the use of the context window [10]. Figure 1 illustrates this process through a multi-step conversation between a user and an assistant. In the first turn, the user provides an input, which the model processes using reasoning tokens to generate an output. This output (along with the input) is carried over to the next turn. In subsequent turns, the model incorporates both the current input and the visible tokens from previous turns while discarding the reasoning tokens after each step. This mechanism ensures that only the relevant input and output tokens are contextually preserved, maintaining efficiency within the fixed 128k-token context window. If the conversation extends beyond this limit, older tokens are truncated to make space for new ones, as shown in the third turn. This mechanism helps the model balance reasoning and context management in extended conversations.

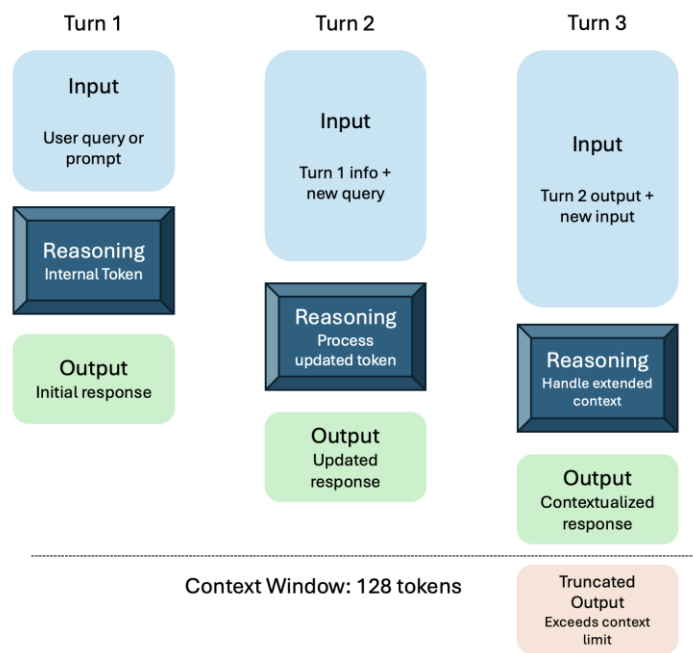


Figure 1. Open AI o1 reasoning: multi-turn conversation flow.

2.2. Prompting Strategies

Reasoning-based models like o1 perform best with straightforward prompts, as their design favors simplicity and clarity. Techniques like few-shot prompting or instructing the model to "think step by step" often fail to enhance performance and can sometimes hinder it. Prompts should be simple and direct to optimize interaction, as the models excel at processing brief and unambiguous instructions without requiring extensive guidance. Chain-of-thought prompts, such as asking the model to "explain your reasoning," are unnecessary since reasoning is already performed internally. For clarity, delimiters like triple quotation marks, XML tags, or section titles can separate distinct input parts, helping the model interpret them appropriately. Additionally, in retrieval-augmented generation (RAG) tasks ([12–15]), it is best to limit the additional context to only the most relevant information to avoid overcomplicating the model's response.

2.3. Comparison of O1-Preview and O1-Mini

The o1 model, like most LLMs, is trained using reinforcement learning to excel in complex reasoning tasks ([16,17]). These models are designed to "think" before answering, generating a detailed internal chain of thought before responding to the user. Currently, two types of o1 models are available: the o1-preview, a reasoning-focused model designed to tackle challenging problems across diverse domains, and the o1-mini, a faster and more cost-effective option specializing in coding, math, and science [10]. Both models offer unique strengths tailored to different use cases. Table 1 provides a comparative summary between the two o1 variants.

Table 1. Comparison of o1-preview and o1-mini.

Model	Context window	Max output tokens	Training data
o1-preview (Points to the most recent snapshot of the o1 model: o1-preview-2024-09-12)	128,000 tokens	32,768 tokens	Up to Oct 2023
o1-mini (Points to the most recent o1-mini snapshot: o1-mini-2024-09-12)	128,000 tokens	65,536 tokens	Up to Oct 2023

2.4. Key O1 Features and Improvements

The o1 series introduces several key features and enhancements. First, it incorporates advanced scaling laws to optimize training, enabling efficient computation across diverse datasets while ensuring reliable performance predictions even with limited training. Second, o1 boasts multidomain expertise, excelling across several fields, including STEM, humanities, and professional domains, unlike earlier models that specialized in narrower tasks. Third, its refined architecture includes improved token embeddings, context processing, and specialized modules to handle diverse input formats such as structured data and complex text. Additionally, o1 is better aligned with human intent, thanks to Reinforcement Learning from Human Feedback (RLHF) ([18,19]), which enhances reasoning, reduces biases, and ensures ethical outputs. Safety and reliability have also been significantly improved, with enhanced mechanisms to mitigate risks like hallucinations, disinformation, and context misinterpretation, supported by adversarial testing and safety pipelines for high-risk domains like law and medicine.

2.5. Comparison with Predecessor Model

Table 2 highlights the key advancements in the OpenAI o1 model compared to GPT-4. O1 demonstrates improvements in training efficiency through optimized scaling laws, achieving more with fewer compute resources. It offers balanced performance across STEM, professional, and humanities domains, unlike GPT-4, which excels primarily in STEM and general NLP tasks. O1's enhanced safety mechanisms, including adversarial testing, make it more reliable in high-stakes applications. It also supports extended context windows of ~16,000+ tokens, doubling GPT-4 capacity. The model introduces improved multimodal capabilities, seamlessly integrating structured data and text, and achieves consistently high multilingual performance, even in low-resource languages. Lastly, O1 features a more refined alignment with human intent by addressing biases through iterative feedback and training.

Table 2. Comparison of o1 with GPT-4 predecessor model.

Feature/Capability	GPT-4	Open AI O1
Training Efficiency	Predictable scaling; requires large compute resources	Optimized scaling laws with efficient compute usage
Domain Generalization	Excels in STEM and general NLP tasks	Balanced performance across STEM, professional, and humanities domains
Safety Mechanisms	RLHF reduces risks but limited in high-stakes applications	Enhanced safety pipeline and adversarial testing for critical domains
Context Window	Limited to shorter contexts (~8,000 tokens)	Supports extended context windows (~16,000+ tokens)
Multimodal Inputs	Limited to text and visual input	Improved multimodal capabilities for structured data and text integration
Multilingual Performance	High performance in English, moderate in low-resource languages	Consistently high performance across multiple languages, including low-resource ones
Alignment	Aligned through RLHF, some biases remain	More refined alignment through iterative feedback and training

2.6. Application Areas

The enhancements in O1 make it a versatile tool for a wide range of applications, especially in high-stakes scenarios. In education, it supports personalized learning by providing detailed and accurate explanations across diverse subjects. In healthcare, its strong performance in clinical knowledge tasks enables it to assist in diagnostics and medical training. For professional services, O1 offers advanced support in fields like law, accounting, and business ethics by addressing complex queries with domain-specific reasoning. In research and development, it facilitates innovation in



science and engineering through its deep understanding of mathematical and technical concepts. Table 3 highlights O1's consistent advancements over GPT-4, showcasing improved performance across diverse benchmarks. From excelling in coding and clinical knowledge tasks to achieving near-perfect scores in mathematics, O1 demonstrates superior reasoning and domain-specific capabilities.

Table 3. Comparison of O1 and GPT-4 performance across benchmarks.

Benchmark	Metric	GPT-4 Accuracy (%)	O1 Accuracy (%)
Multitask Language Understanding (MMLU)	Average Accuracy	86.4	89.2
HumanEval (Python Coding Tasks)	Pass Rate	67.0	73.5
Clinical Knowledge Test	Accuracy	90.5	93.0
Professional Law Test	Accuracy	50.2	54.0
High School Mathematics Test	Accuracy	95.0	98.0

While existing literature has begun to explore O1's capabilities in specific benchmarks, as illustrated in Table 3, our work aims to present a broader evaluation that spans multiple domains. This report seeks to identify o1's strengths, limitations, and potential applications by examining its performance across diverse tasks. This contribution complements existing evaluations and provides a foundational reference for understanding O1's real-world applicability across varied fields.

3. Materials & Methods

In our evaluation of the OpenAI O1 model (o1-preview), we utilized a variety of standardized tests to assess its performance across multiple domains. Table 4 summarizes the cross-domain evaluation of o1. The benchmark data is sourced from the massive multitask language understanding (MMLU) [20] benchmark and LLM Benchmarks on GitHub [21].

Table 4. Summary of Evaluation Data.

Domain	Test Description	Benchmark
Abstract Algebra	Evaluates the model's understanding of algebraic structures such as groups, rings, and fields, fundamental in advanced mathematics	MMLU
Anatomy	Assesses knowledge of human anatomical structures and systems, essential for medical and biological sciences	MMLU
Astronomy	Tests comprehension of celestial objects, phenomena, and the universe's structure, crucial for astrophysics studies	MMLU
Business Ethics	Evaluates understanding of moral principles in business contexts, important for corporate governance and ethical decision-making	MMLU
Clinical Knowledge	Measures proficiency in medical knowledge and clinical practices, vital for healthcare professionals. Included in the MMLU benchmark	MMLU
College Biology	Assesses understanding of biological concepts at the college level, covering topics like genetics, ecology, and physiology	MMLU
College Chemistry	Tests knowledge of chemical principles, reactions, and laboratory practices at the collegiate level	MMLU
College Mathematics	Evaluates proficiency in higher-level mathematics, including calculus, linear algebra, and differential equations	MMLU
Econometrics	Assesses the ability to apply statistical methods to economic data, essential for economic analysis and forecasting	MMLU
Electrical Engineering	Tests knowledge of electrical circuits, systems, and signal processing, fundamental for engineering disciplines	MMLU

World Religions	Evaluates understanding of major world religions, their histories, beliefs, and cultural impacts	MMLU
Elementary Mathematics	Assesses basic mathematical skills, including arithmetic and elementary problem-solving	LLM Benchmarks
Global Facts	Tests general knowledge about world geography, politics, and global events	LLM Benchmarks
High School Chemistry	Evaluates understanding of chemical principles taught at the high school level	LLM Benchmarks
High School Computer Science	Assesses knowledge of basic computer science concepts, including programming and algorithms	LLM Benchmarks
High School Geography	Tests understanding of physical and human geography topics covered in high school curricula	LLM Benchmarks
High School Macroeconomics	Evaluates knowledge of economic principles related to the economy, such as inflation and GDP	LLM Benchmarks
High School Physics	Assesses understanding of fundamental physics concepts taught at the high school level	LLM Benchmarks
High School Psychology	Tests knowledge of psychological theories and practices covered in high school courses	LLM Benchmarks
Medical Genetics	Evaluates understanding of genetic principles and their medical applications, crucial for healthcare and research	MMLU
Nutrition	Assesses knowledge of dietary principles, human nutrition, and health implications	MMLU
Professional Accounting	Tests proficiency in accounting principles, financial reporting, and auditing practices	MMLU
Professional Law	Evaluates understanding of legal concepts, case law, and legal reasoning, essential for legal professionals	MMLU
Professional Medicine	Assesses clinical knowledge and medical practices required for healthcare providers	MMLU
Public Relations	Tests understanding of communication strategies, media relations, and public perception management	MMLU

We evaluated the model using publicly available resources and repositories to ensure transparency and reproducibility. These standardized tests covered a wide range of academic and professional subjects. Model outputs were compared against verified ground truth answers to ensure accurate assessment and random sampling was employed to maintain unbiased evaluation across experiments. Experiment 1 involved selecting 100 questions randomly sampled from 11 key tests representing diverse domains. These included STEM subjects such as Abstract Algebra, College Mathematics, Clinical Knowledge, and Electrical Engineering, as well as professional domains like Business Ethics and World Religions. Scientific knowledge areas such as College Biology, College Chemistry, and Astronomy were also covered. Experiment 2 expanded the evaluation to include 25 randomly sampled questions from 14 additional tests, covering a further range of subjects like Professional Law, Public Relations, and Sociology. Accuracy served as the primary evaluation metric, calculated as the percentage of correct answers produced by the model.

4. Results

4.1. Experiment Set 1

The o1 model demonstrated strong performance across a range of tasks, particularly excelling in STEM fields such as abstract algebra, astronomy, and college mathematics, as well as in clinical knowledge and anatomy, as seen in Table 5 and Figures 2 and 3. These results highlight the model's proficiency in domains requiring technical precision and specialized reasoning. However, its performance in business ethics and chemistry-related tasks was relatively moderate, indicating areas

where further refinement may be necessary. Additionally, as depicted in Figures 2 and 3, the model exhibited consistently high accuracy across most tests, with minor fluctuations, reflecting its balanced cross-domain capabilities. The trends emphasize the model's ability to handle diverse subject areas while maintaining robust accuracy in the domains it is most competent in.

Table 5. Performance of o1 across several tests in experiment 1.

Test	Accuracy (%)
Abstract Algebra	91
Anatomy	92
Astronomy	97
Business Ethics	81
Clinical Knowledge	93
College Biology	98
College Chemistry	77
College Mathematics	95
Econometrics	84
Electrical Engineering	90
World Religions	90

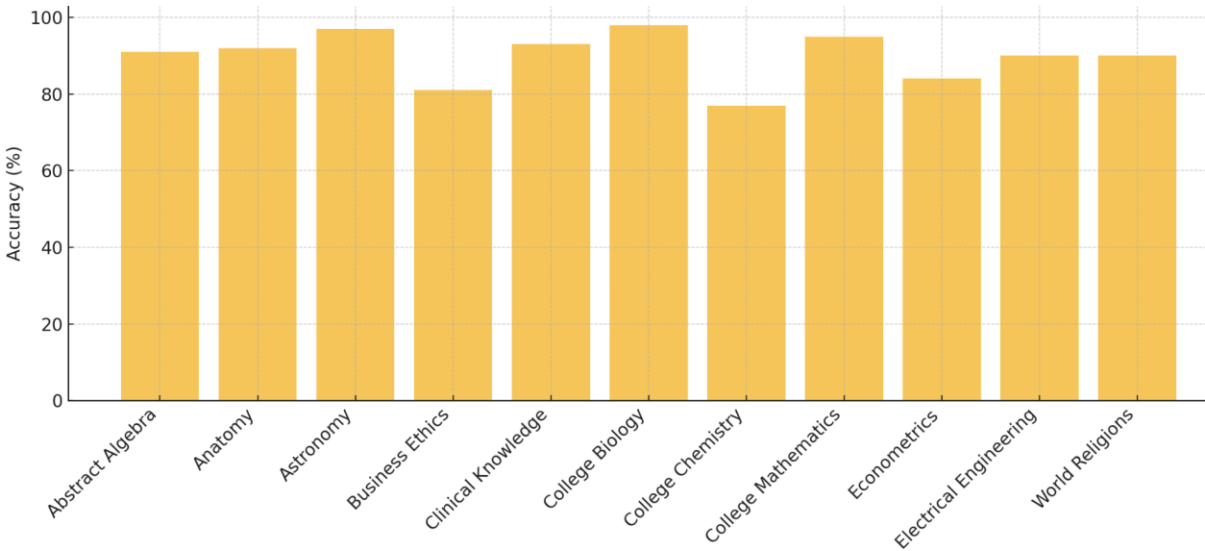


Figure 2. Performance of o1 across several tests in experiment 1.



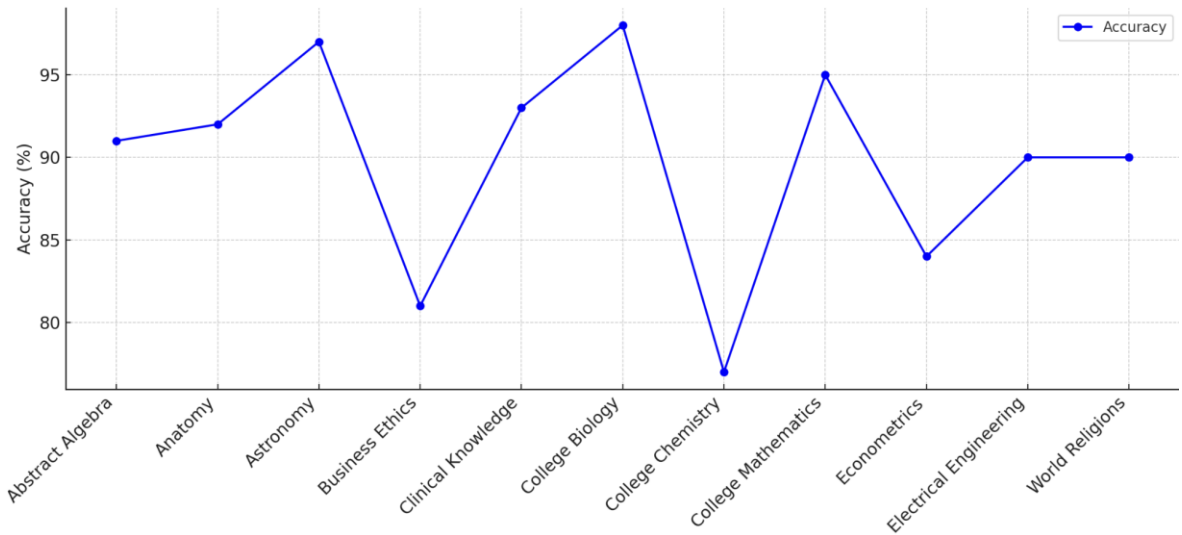


Figure 3. O1 Accuracy of o1 across several tests in experiment 1.

4.2. Experiment Set 2

In Experiment 2, o1 exhibited notable strengths across a variety of test domains, particularly excelling in subjects such as elementary mathematics, high school computer science, psychology, and medical genetics, achieving near-perfect accuracy. The results summarized in Table 6 show the model's proficiency in both foundational and advanced technical domains. However, as shown in Figure 4, the model demonstrated relatively lower performance in professional fields such as law and accounting, indicating challenges in applying its reasoning capabilities to tasks requiring nuanced contextual understanding. The performance trend across Figure 5 reveals a clear consistency in high school-level subjects and a sharp drop in professional law, emphasizing the model's domain-specific strengths and areas needing refinement.

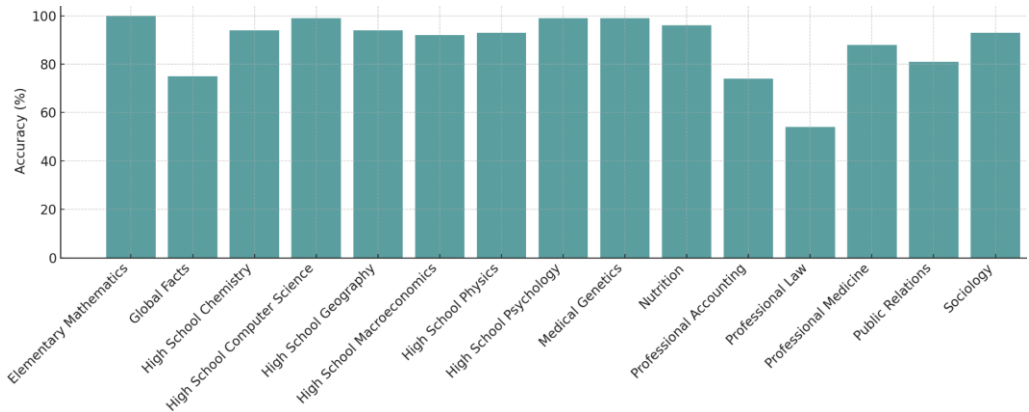


Figure 4. Performance of o1 across tests in experiment 2.

Table 6. Performance of o1 across tests in experiment 2.

Test	Accuracy (%)
Elementary Mathematics	100
Global Facts	75
High School Chemistry	94
High School Computer Science	99
High School Geography	94
High School Macroeconomics	92

High School Physics	93
High School Psychology	99
Medical Genetics	99
Nutrition	96
Professional Accounting	74
Professional Law	54
Professional Medicine	88
Public Relations	81
Sociology	93

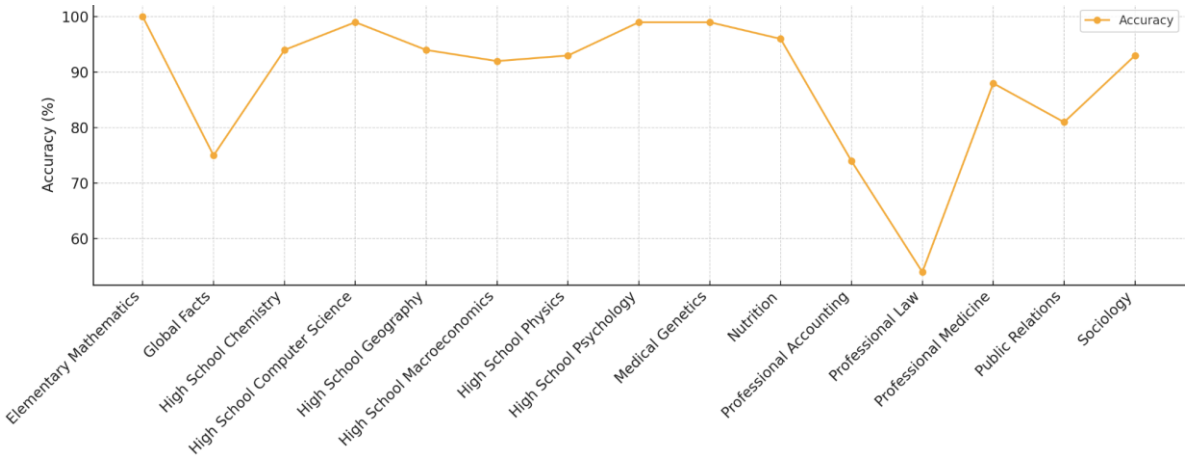


Figure 5. Accuracy of o1 across tests in experiment 2.

4.3. Sample Responses

In Table 7, we provide several examples of o1 responses across several tests; we outline the domain, correct answer (GT), o1 answer, and reasoning.

Table 7. Sample o1 responses.

<p><b>Question:</b> Primary motor cortex activity results in:</p> <ul style="list-style-type: none"><li>• <b>Choice A:</b> Bilateral contraction of antigravity limb muscles.</li><li>• <b>Choice B:</b> Bilateral contraction of limb musculature.</li><li>• <b>Choice C:</b> Contraction of ipsilateral limb musculature.</li><li>• <b>Choice D:</b> Contraction of contralateral limb musculature.</li></ul>	<p><b>Domain:</b></p> <p><b>Anatomy</b></p> <p><b>GT:</b> D</p> <p><b>O1:</b> D</p>	<p><b>O1 Reasoning:</b></p> <p><b>Answer:</b> Choice D: Contraction of contralateral limb musculature.</p> <p><b>Explanation:</b></p> <p>The <b>primary motor cortex</b>, located in the precentral gyrus of the frontal lobe, controls voluntary movements of the body. Neurons from the primary motor cortex project to the opposite (contralateral) side of the body via the corticospinal tract. This means that activity in the left motor cortex controls muscles on the right side of the body and vice versa.</p>
<p><b>Question:</b> Statement 1   A ring homomorphism is one to one if and only if the kernel is {0}. Statement 2   Q is an ideal in R.</p> <ul style="list-style-type: none"><li>• <b>Choice A:</b> True, True</li><li>• <b>Choice B:</b> False, False</li><li>• <b>Choice C:</b> True, False</li><li>• <b>Choice D:</b> False, True</li></ul>	<p><b>Domain:</b></p> <p><b>Abstract Algebra</b></p> <p><b>GT:</b> D</p> <p><b>O1:</b> C</p>	<p><b>O1 Reasoning:</b></p> <p>The correct choice is <b>Choice C: True, False</b>.</p> <p><b>Explanation:</b></p> <p><b>Statement 1:</b> "A ring homomorphism is one to one if and only if the kernel is {0}."</p> <ul style="list-style-type: none"><li>• <b>True.</b> In ring theory, a ring homomorphism <math>\phi : R \rightarrow S</math> is <b>injective (one-to-one)</b> precisely when its <b>kernel</b> <math>\ker(\phi) = \{r \in R \mid \phi(r) = 0_S\}</math> contains only the <b>zero element</b> of <math>R</math>. If <math>\ker(\phi) = \{0\}</math>, then <math>\phi</math> has no non-trivial elements mapping to zero, ensuring injectivity. Conversely, if <math>\phi</math> is injective, the only element that can map to zero is the zero element itself.</li></ul> <p><b>Statement 2:</b> "Q is an ideal in R."</p> <ul style="list-style-type: none"><li>• <b>False.</b> Without specific context, it's generally <b>not true</b> that <math>\mathbb{Q}</math> (the set of rational numbers) is an ideal in <math>R</math> (which could be any ring). For example:<ul style="list-style-type: none"><li>• If <math>R = \mathbb{Z}</math> (the ring of integers), <math>\mathbb{Q}</math> is <b>not</b> a subset of <math>\mathbb{Z}</math>, let alone an ideal.</li><li>• If <math>R = \mathbb{Q}</math> itself, the only ideals are <math>\{0\}</math> and <math>\mathbb{Q}</math>, since <math>\mathbb{Q}</math> is a field.</li></ul></li></ul> <p>In most standard contexts, <math>\mathbb{Q}</math> does not form a non-trivial ideal in typical rings <math>R</math>.</p>

**Question:** The Milky Way is part of a giant supercluster with a diameter of 160 Mpc. What is the name of this supercluster?

- **Choice A:** Virgo
- **Choice B:** Laniakea
- **Choice C:** Sculptor
- **Choice D:** Boötes

**Domain:**  
**Astronomy**

**GT:** B

**O1:** B

**Laniakea Supercluster** is the massive structure that encompasses the Milky Way galaxy. Here's a breakdown of the options:

- **Choice A: Virgo**
  - **Virgo Supercluster** is an earlier classification that includes the Local Group (which contains the Milky Way, Andromeda, and other nearby galaxies). However, Virgo is a part of the larger Laniakea Supercluster. Therefore, while Virgo is related, it does not represent the entire supercluster containing the Milky Way.
- **Choice B: Laniakea**
  - **Laniakea Supercluster** is the correct answer. Announced in 2014 by a team led by astronomer R. Brent Tully, Laniakea spans approximately **160 megaparsecs (Mpc)** in diameter. It includes the Virgo Supercluster and other galaxy clusters, forming a vast network that defines our position in the universe. "Laniakea" means "immense heaven" in Hawaiian, reflecting its vastness.
- **Choice C: Sculptor**
  - **Sculptor Supercluster** is a different structure and does not include the Milky Way. It is named after the Sculptor constellation but is not related to the supercluster containing our galaxy.
- **Choice D: Boötes**
  - **Boötes Supercluster** is another distinct supercluster named after the Boötes constellation. It is separate from the Laniakea Supercluster and does not include the Milky Way.

**Question:** A stakeholder orientation includes all of the following activities except:

- **Choice A:** Generating Data About Stakeholder Groups
- **Choice B:** Assessing The Firm's Effects On Stakeholder Groups
- **Choice C:** Distributing Stakeholder Information Throughout The Firm
- **Choice D:** Minimizing The Influence Of Stakeholder Information On The Firm

**Domain:**  
**Business**  
**Ethics**

**GT:** D

**O1:** D

**Stakeholder Orientation** is a business approach that emphasizes the importance of understanding and addressing the needs, interests, and influences of all stakeholders involved with or affected by the firm. This includes not only shareholders but also employees, customers, suppliers, communities, and other relevant groups.

Let's analyze each option:

- **Choice A: Generating Data About Stakeholder Groups**
  - **Included in Stakeholder Orientation.** Collecting data about stakeholders is fundamental to understanding their needs and expectations. This information helps the firm tailor its strategies and operations to better serve and engage with its stakeholders.
- **Choice B: Assessing The Firm's Effects On Stakeholder Groups**
  - **Included in Stakeholder Orientation.** Evaluating how the firm's actions impact its stakeholders is crucial for responsible management and ensuring that the firm maintains positive relationships and minimizes negative impacts.
- **Choice C: Distributing Stakeholder Information Throughout The Firm**
  - **Included in Stakeholder Orientation.** Sharing information about stakeholders across the organization ensures that all departments and employees are aware of stakeholder expectations and can align their actions accordingly. This promotes a unified approach to stakeholder engagement.
- **Choice D: Minimizing The Influence Of Stakeholder Information On The Firm**
  - **Not Included in Stakeholder Orientation.** On the contrary, stakeholder orientation advocates for integrating stakeholder information into the firm's decision-making processes. Minimizing the influence of such information would be contrary to the principles of stakeholder orientation, which seeks to prioritize and incorporate stakeholder insights to enhance the firm's sustainability and success.



**Question:** Which of the following is true in diplopia?

- **Choice A:** Diplopia can never occur if one eye is covered
- **Choice B:** The outer image is always the false image
- **Choice C:** A fourth nerve palsy occurs when the patient looks upwards
- **Choice D:** A sixth nerve palsy causes a divergent squint

**Domain:**  
**Clinical**  
**Knowledge**

**GT:** B

**O1:** None

None of the provided choices are entirely true regarding diplopia. Here's a detailed analysis of each option:

**Choice A: Diplopia can never occur if one eye is covered**

- **False.** While **binocular diplopia** (double vision caused by misalignment of the eyes) typically resolves when one eye is covered, **monocular diplopia** (double vision that persists in one eye regardless of eye alignment) can still occur even if one eye is covered. Monocular diplopia is less common and usually results from issues within the eye itself, such as cataracts, corneal irregularities, or retinal problems.

**Choice B: The outer image is always the false image**

- **False.** In diplopia, there is no consistent rule that designates the outer (or inner) image as the "false" one. The nature of the double images depends on the underlying cause of the diplopia, such as nerve palsies, muscle imbalances, or refractive errors. Both images can be real or distorted, and their positions can vary based on the specific condition.

**Choice C: A fourth nerve palsy occurs when the patient looks upwards**

- **False.** The **fourth cranial nerve (trochlear nerve)** controls the **superior oblique muscle**, which primarily facilitates downward and inward eye movements. A fourth nerve palsy typically results in difficulties with looking downward, especially when the eye is adducted (turned toward the nose), and can cause vertical diplopia. It is **not** specifically associated with looking upwards.

**Choice D: A sixth nerve palsy causes a divergent squint**

- **False.** The **sixth cranial nerve (abducent nerve)** controls the **lateral rectus muscle**, which abducts the eye (moves it outward). A sixth nerve palsy leads to an inability to abduct the affected eye, resulting in **esotropia** (inward turning of the eye), not a **divergent (exotropia)** squint. Therefore, this statement is incorrect.

## 5. Conclusions

In this work, we conducted a comprehensive evaluation of the OpenAI o1 (o1-preview) model across diverse domains, highlighting its strengths in STEM, clinical knowledge, and high school-level subjects while identifying areas for improvement in professional contexts such as law and accounting. Our analysis underscores the model's ability to generalize across multiple domains, demonstrating its potential for both academic and professional applications. However, the observed variability in performance suggests that further refinement is needed to enhance its capabilities in tasks requiring contextual depth and domain-specific expertise. Future research should focus on extending this evaluation framework to include additional complex and real-world datasets, particularly in professional and interdisciplinary domains. Exploring methods to fine-tune the model for domain-specific applications and incorporating human feedback for iterative improvements could further enhance its alignment with task-specific requirements.

## Author statements

### *Ethical Approval*

Ethical approval was not required because no personal data was used. Any analysis presented were aggregated.

### *Competing Interests*

None declared.

## References

1. E. Kasneci *et al.*, "ChatGPT for good? On opportunities and challenges of large language models for education," *Learn. Individ. Differ.*, vol. 103, p. 102274, Apr. 2023, doi: 10.1016/j.lindif.2023.102274.
2. S. Shahriar, N. Al Roken, and I. Zuolkernan, "Classification of Arabic Poetry Emotions Using Deep Learning," *Computers*, vol. 12, no. 5, Art. no. 5, May 2023, doi: 10.3390/computers12050089.
3. S. Shahriar *et al.*, "Putting GPT-4o to the Sword: A Comprehensive Evaluation of Language, Vision, Speech, and Multimodal Proficiency," *Appl. Sci.*, vol. 14, no. 17, Art. no. 17, Jan. 2024, doi: 10.3390/app14177782.
4. K. Hayawi, S. Shahriar, H. Alashwal, and M. A. Serhani, "Generative AI and large language models: A new frontier in reverse vaccinology," *Inform. Med. Unlocked*, vol. 48, p. 101533, Jan. 2024, doi: 10.1016/j.imu.2024.101533.
5. L. Yan *et al.*, "Practical and ethical challenges of large language models in education: A systematic scoping review," *Br. J. Educ. Technol.*, vol. 55, no. 1, pp. 90–112, 2024, doi: 10.1111/bjet.13370.
6. A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nat. Med.*, vol. 29, no. 8, pp. 1930–1940, Aug. 2023, doi: 10.1038/s41591-023-02448-8.
7. S. Shahriar and N. Al Roken, "How can generative adversarial networks impact computer generated art? Insights from poetry to melody conversion," *Int. J. Inf. Manag. Data Insights*, vol. 2, no. 1, p. 100066, Apr. 2022, doi: 10.1016/j.jjimei.2022.100066.
8. L. Zhao *et al.*, "When brain-inspired AI meets AGI," *Meta-Radiol.*, vol. 1, no. 1, p. 100005, Jun. 2023, doi: 10.1016/j.metrad.2023.100005.
9. H. Zhao, A. Chen, X. Sun, H. Cheng, and J. Li, "All in One and One for All: A Simple yet Effective Method towards Cross-domain Graph Pretraining," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, in KDD '24. New York, NY, USA: Association for Computing Machinery, Aug. 2024, pp. 4443–4454. doi: 10.1145/3637528.3671913.
10. "Introducing OpenAI o1." Accessed: Dec. 06, 2024. [Online]. Available: <https://openai.com/o1/>

11. T. Zhong *et al.*, "Evaluation of OpenAI o1: Opportunities and Challenges of AGI," Sep. 27, 2024, *arXiv*: arXiv:2409.18486. doi: 10.48550/arXiv.2409.18486.
12. W. Fan *et al.*, "A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, in KDD '24. New York, NY, USA: Association for Computing Machinery, Aug. 2024, pp. 6491–6501. doi: 10.1145/3637528.3671470.
13. P. Lewis *et al.*, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2020, pp. 9459–9474. Accessed: Dec. 06, 2024. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>
14. S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, and S. Nanayakkara, "Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering," *Trans. Assoc. Comput. Linguist.*, vol. 11, pp. 1–17, Jan. 2023, doi: 10.1162/tacl\_a\_00530.
15. S. Wu *et al.*, "Retrieval-Augmented Generation for Natural Language Processing: A Survey," *CoRR*, Jan. 2024, Accessed: Dec. 06, 2024. [Online]. Available: <https://openreview.net/forum?id=bjrU9flhD3>
16. T. Carta, C. Romac, T. Wolf, S. Lamprier, O. Sigaud, and P.-Y. Oudeyer, "Grounding Large Language Models in Interactive Environments with Online Reinforcement Learning," in *Proceedings of the 40th International Conference on Machine Learning*, PMLR, Jul. 2023, pp. 3676–3713. Accessed: Dec. 06, 2024. [Online]. Available: <https://proceedings.mlr.press/v202/carta23a.html>
17. S. Shahriar and K. Hayawi, "Let's Have a Chat! A Conversation with ChatGPT: Technology, Applications, and Limitations," *Artif. Intell. Appl.*, vol. 2, no. 1, Art. no. 1, 2024, doi: 10.47852/bonviewAIA3202939.
18. Y. Bai *et al.*, "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback," Apr. 12, 2022, *arXiv*: arXiv:2204.05862. doi: 10.48550/arXiv.2204.05862.
19. H. Lee *et al.*, "RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback," Oct. 2023, Accessed: Dec. 06, 2024. [Online]. Available: <https://openreview.net/forum?id=AAxIs3D2ZZ>
20. D. Hendrycks *et al.*, "Measuring Massive Multitask Language Understanding," presented at the International Conference on Learning Representations, Oct. 2020. Accessed: Dec. 05, 2024. [Online]. Available: <https://openreview.net/forum?id=d7KBjml3GmQ>
21. F. Beeson, *leobeelson/llm\_benchmarks*. (Dec. 05, 2024). Accessed: Dec. 05, 2024. [Online]. Available: [https://github.com/leobeelson/llm\\_benchmarks](https://github.com/leobeelson/llm_benchmarks)

**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.