*Brief Report*

# GPT-4 vs. GPT-3.5: A Concise Showdown

Anis Koubaa
*Prince Sultan University, Saudi Arabia*
akoubaa@psu.edu.sa

***Abstract*—As the echoes of ChatGPT's remarkable success continue to permeate the AI community, its formidable successor, GPT-4, has emerged, showing off a wealth of novel features. In this concise paper, we elucidate the capabilities of GPT-4 and conduct a comparative analysis with its predecessor, ChatGPT, offering insights into their relative strengths and advancements in the rapidly evolving field of generative AI. We also present a comprehensive summary of the major performance results reported by OpenAI on GPT-4 across various Natural Language Processing (NLP) tasks. We place great emphasis on the innovative advancements offered by GPT-4 in comparison to its predecessors. Our focus is on highlighting its remarkable performance while also mentioning its limitations. The purpose of this paper is to deliver a succinct understanding of the new features and performance benchmarks of GPT-4.**

***Index Terms*—ChatGPT, GPT-4, GPT-3.5, GPT Performance, GPT Limitations, OpenaAI, NLP**

## I. Introduction

Following the groundbreaking release of ChatGPT (based on GPT-3.5) by OpenAI in November 2022, the pursuit of developing state-of-the-art generative Large Language Models (LLMs) for interactive conversation and text completion has intensified, driven by the remarkable success and global impact of ChatGPT. In a prompt response to the rapidly evolving market, OpenAI unveiled GPT-4, the latest addition to the GPT family. It has been hailed for its cutting-edge advancements and unparalleled capabilities.

The objective of this paper is to present the key features of GPT-4, elucidating its characteristics and distinguishing aspects. We also conduct a thorough comparative analysis with its predecessor, ChatGPT, unveiling their respective strengths, limitations, and the unique advancements introduced by GPT-4 in the rapidly evolving domain of generative AI.

## II. GPT-4 vs. GPT3.5 Training Process: The Rule-based Approach

GPT-4 shares a lot of common features with GPT-3.5 [1], [2] in the sense that they both rely on a similar Transformers architectural model, but of course, at different scales. The transformer model relies on an encoder-decoder architecture with self-attention modules [3] responsible for capturing complex relationships and extracting patterns from input sequences. The encoder process input sequences, and the decoder converts the output of the encoder to the generated sequence at the output of the transformer.

On the other hand, OpenAI revealed little information about the training process of GPT-4, compared to its predecessor, where it disclosed all technical information.

However, it is clear that GPT-4 introduced a rule-based reward model (RBRM) approach as compared to GPT-3.5, in addition to the known Reinforcement Learning with Human Feedback (LRHF) [4] It was reported in [5]: "*Our rule-based reward models (RBRMs) are a set of zero-shot GPT-4 classifiers. These classifiers provide an additional reward signal to the GPT-4 policy model during RLHF fine-tuning [6] targets correct behavior, such as refusing to generate harmful content or not refusing innocuous requests.*"

The Rule-Based Reward Models (RBRMs) approach improves language models' performance and safety, like GPT-4. It provides additional reward signals during the Reinforcement Learning from Human Feedback (RLHF) fine-tuning process on the generated text to ensure its compliance with generating safe and correct content. In GPT-4, The Rule-Based Reward Models are set to zero-shot classifiers, meaning they were not fine-tuned on specific tasks. At the same time, they can still generate text safely, considering the large knowledge-based leveraged during the pre-training phase. These classifiers serve as an additional reward signal for the GPT-4 policy model on top of the Reinforcement Learning from Human Feedback (RLHF) fine-tuning process. Combining these two fine-tuning models helps improve the reliability and safety of GOT-4 and reduces hallucinations cases dramatically compared to GPT-3.5. Although combining RBRMs with RLHF fine-tuning can improve the reliability and safety of GPT-4, it is essential to remember that this does not mean that GPT-4 is perfect. The approach may reduce hallucinations and other issues compared to previous models, but some challenges and limitations persist, as will be discussed in Section V.

## III. Architecture and Model Size

The GPT-4 model is known to retain the transformer-based architecture characteristic of its predecessors in the GPT series. However, as of March 2023, OpenAI has not released a detailed technical report on GPT-4 [5], diverging from their approach with GPT-1, GPT-2, and GPT-3. The available information in [5] pertaining to GPT-4's architecture is rather generic, stating that *"GPT-4 is a Transformer-style model [33] pre-trained to predict the next token in a document, using both publicly available data (such as internet data) and data licensed from third-party providers. The model was then fine-tuned using Reinforcement Learning from Human Feedback (RLHF)"*. This description aligns with the fundamental features common to legacy GPT models, offering little insight into any distinct architectural advancements.

OpenAI explicitly reported in their technical report:" *Given both the competitive landscape and the safety implications of*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 March 2023 doi:10.20944/preprints202303.0422.v1



| Features | GPT-4 | GPT-3 (ChatGPT) |
|---|---|---|
| **Model Size** | Not officially available [1] | 175 billion |
| **Modality** | Text, Images | Text |
| **Context Window Length** | 8192 to 32768 | 2048 |

TABLE I
COMPARISON OF GPT-4 AND GPT-3.5

*large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute,dataset construction, training method, or similar.*". This particular approach from OpenAI diverges from their previous practices, where they openly shared technical details of the GPT model's internal architecture with the community. It appears that, in this instance, business considerations have taken precedence, given the growing competition in the field of generative large-scale language models.

The architecture of GPT-4 exhibits a significant advancement in scale compared to its predecessors. Only a few parameters have been disclosed, emphasizing the substantial differences between GPT-4 and the earlier GPT-3 variants. Table I and Figure 1 provide a comprehensive illustration of these discrepancies, emphasizing the remarkable increase in parameters present in GPT-4 compared to all GPT-3 variants.

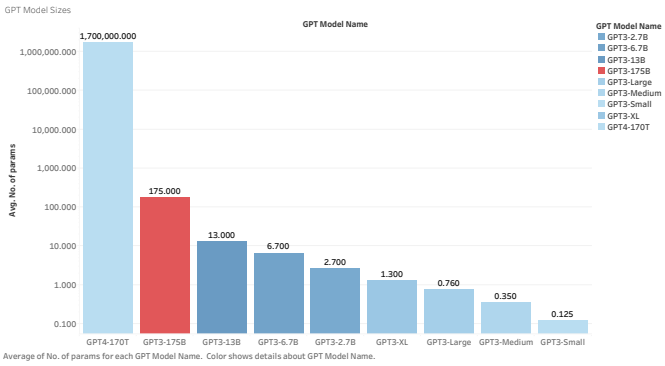



Fig. 1. GPT Models Size in Logarithmic Scale

Table I compares the key features of GPT-4 and GPT-3.5 models.

- **Model Size:** GPT-4 is 1000 times larger in size reaching 170 Trillion parameters compared to 175 Billion parameters of GPT-3.5. The size difference demonstrates the amplified capabilities in performance and accuracy and dealing with complex language models and natural language processing tasks.
- **Modality:** GPT-4 improves on GPT-3 by supporting multimodal inputs, including text and images, in contrast to GPT-3, which processes only text inputs. GPT-4 is
- **Context Window Length:** The difference in context window length is illustrated in Figure 2. The context width in ChatGPT refers to the number of previous tokens or words the model utilizes to generate its response in a conversation. The context can affect the relevance, coherence, and overall quality of ChatGPT's response. The context length in GPT3.5 is 2048 and has increased to 8192 and 32768 (depending on the version) in GPT4 in input, which is 4 to 16 times greater than GPT-3.5. Regarding output, GPT-4 can generate up to 24000 words (equivalent to 48 pages), 8 times higher than GPT-3.5, constrained by 3000 words (equivalent to 6 pages). GPT-4 demonstrates a tremendous increase in the context window scale, which allows larger inputs for improved accuracy and relevance and generates longer text.

In summary, GPT-4 features a significantly larger architectural model size than its predecessors, including GPT-3 and its variants. The increased model size helps improve its natural language processing (NLP) capabilities, resulting in more accurate and relevant responses. However, this larger scale induces more significant processing and computing resource requirements and longer delays in generating text.

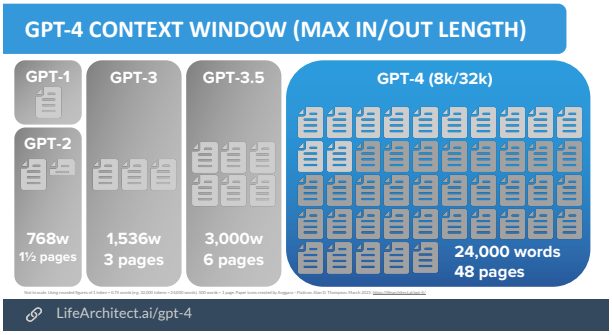



Fig. 2. GPT-4 Context Window Length Comparison with GPT-(1,2,3.x) [7]

## IV. GPT-4 PERFORMANCE

While the GPT-4 technical report did not elaborate on any technical details related to the architecture and other technical contributions/details, it mainly emphasized disclosing its performance against different benchmarks. In this section, we present a succinct summary of the performance benchmarks of GPT-4.

### A. Predictive Scaling Performance Study

OpenAI wanted to test if the training process of their GPT-4 LLM model could be scaled up to larger models. For this purpose, they developed a deep learning stack that scales predictably for large training runs like GPT-4, where extensive model-specific tuning is not feasible. They trained smaller models using the same methodology as GPT-4, but with at most 10,000x less compute and at most 1,000x less compute for the HumanEval dataset. These smaller models were used to predict the final loss and pass rate of GPT-4, respectively.

They concluded that they were able to accurately predict two performance metrics of GPT-4, namely:

- **The Final Loss:** Predicting the final loss in large language models (LLMs) is useful to avoid performing useless computation-intensive training and get an initial pre-training assessment of the quality of the model and its performance. This help avoids the unnecessary use

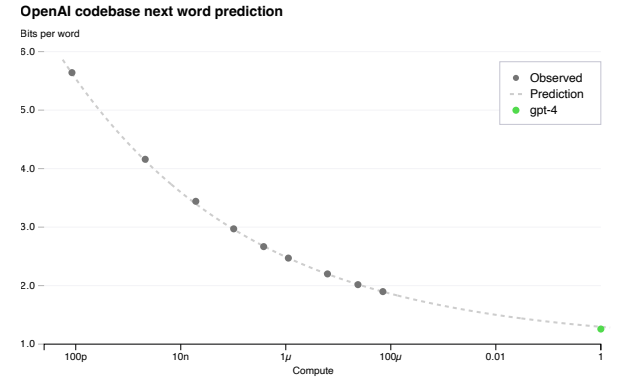
[1] Some unofficial sources mentioned 170 trillions parameters

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 March 2023 doi:10.20944/preprints202303.0422.v1

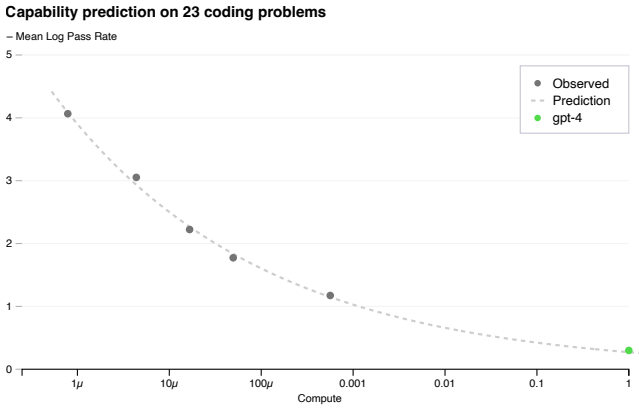


of computing resources and guides researchers to decide whether to stop the training process and evaluate its performance. As shown in Fig. 3(a), OpenAI reported that the approximated final loss of properly-trained LLMs is approximated by a power-law distribution in the amount of compute resources used to train the model.

- **Pass rate on the HumanEval dataset:** HumanEval is a dataset comprising a collection of tasks used to evaluate the capabilities of large language models to generate and understand human-like language, as released in [8]. In [5], they predicted the performance of GPT-4 by evaluating its ability to solve Python programming problems of varying complexity. This is what defines the Pass rate on the HumanEval dataset. They predicted the pass rate on a subset of the HumanEval dataset by extrapolating from models trained with at most 1,000× less compute. They also discovered an approximate power law distribution relationship between the pass rate and the amount of computation used to train the model, as illustrated in Fig. 3(b).




(a) Loss Prediction




(b) Pass Rate on HumanEval Dataset

Fig. 3. GPT-4 Performance Predictions [5]

## B. Benchmark Exams Capability

GPT-4 has undergone a comprehensive evaluation process to determine its human-level performance on various assessments. These assessments can be categorized into two main types of benchmarks: (*i.*)Academic and Professional Exams, which encompass exams from various disciplines such as Math, Science, Law, and others, and (*ii.*) Multilingual Performance, where GPT-4's 3-Shot Accuracy is compared across multiple languages.

The Academic and Professional Exams cover a broad spectrum of exams, including those commonly used for university admissions (e.g., SAT, GRE, Bar), professional licensing, and advanced placement. The Multilingual Performance evaluation involves testing GPT-4's abilities in various languages and comparing its performance to existing language models.

The following subsection summarizes the performance of GPT-4 on the aforementioned evaluations.

*1) Academic and Professional Exams:* Table I compares GPT-4 and GPT-3.5 performance in various professional and academic exams. This table offers a detailed comparison of the two language models' abilities to perform on these types of exams, providing insights into the advancements made by GPT-4 compared to its predecessor. Each exam was assessed through exam-specific rubrics, and the final score of GPT models was reported with their percentiles of test-takers who achieved the same score as GPT-4. The analysis reveals that GPT-4 consistently outperforms its predecessor, GPT-3.5, across various professional and academic exams. Notably, GPT-4 scored in the top 10% of test takers on the challenging Uniform Bar Exam, demonstrating its impressive capabilities in the legal domain. This performance shows the extent of capability improvements of GPT-4 in natural language processing and exhibits human-level performance on most of these professional and academic certification exams.

## C. Multilingual Performance

The evaluation of professional and academic exams primarily focused on the English language. However, OpenAI expanded its assessment to include Multilingual exams, which involved translating multiple-choice problems spanning fifty-seven topics into various languages. To accomplish this, Azure Translate was utilized to translate the questions into different languages. This comprehensive evaluation allowed OpenAI to assess GPT-4's performance in various languages, providing valuable insights into the model's multilingual capabilities.

As depicted in Figure 4, the performance of GPT-4 with regards to supporting multiple languages has been compared against GPT-3.5, with the former showing substantial improvements in the English language. For non-English languages, 15 out of 27 (55.6%) demonstrated a 3-shot accuracy higher than 80%, with the remaining 12 languages falling below this threshold. This significant advancement in GPT-4, compared to its predecessors, has opened up new frontiers for NLP applications across various industries.

# V. Challenges and Limitations of GPT-4

While GPT-4 brings significant advancements, it shares some limitations with GPT-3.5 [9], albeit with reduced impact. OpenAI recently published a technical report titled "GPT-4 System Card" [10], which offers a detailed analysis of the model's primary limitations and persistent challenges.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 March 2023 doi:10.20944/preprints202303.0422.v1



| Category | Exams | GPT-4 estimated percentile | GPT-4 (no vision) estimated percentile | GPT-3.5 estimated percentile |
|---|---|---|---|---|
| Math | SAT Math | 700/800 ∼ 89th | 690/800 ∼ 89th | 590/800 ∼ 70th |
| | GRE Quantitative | 163/170 ∼ 80th | 157/170 ∼ 62nd | 147/170 ∼ 25th |
| | AP Calculus BC | 443rd ∼ 59th | 443rd ∼ 59th | 10th ∼ 7th |
| Science | USABO Semifinal Exam 2020 | 87/150 99th ∼ 100th | 87/150 99th ∼ 100th | 43/150 31st ∼ 33rd |
| | USNCO Local Section Exam 2022 | 36/60 | 38/60 | 24/60 |
| | AP Biology | 585th ∼ 100th | 585th ∼ 100th | 462nd ∼ 85th |
| Computer Science | Codeforces Rating | 392 below 5th | 392 below 5th | 260 below 5th |
| Medicine | Medical Knowledge Self-Assessment Program | | | |
| Law | Uniform Bar Exam (MBE+MEE+MPT) | 298/400 ∼ 90th | 298/400 ∼ 90th | 213/400 ∼ 10th |
| | LSAT | 163 ∼ 88th | 161 ∼ 83rd | 149 ∼ 40th |
| Others | SAT Evidence-Based Reading & Writing | 710/800 ∼ 93rd | 710/800 ∼ 93rd | 670/800 ∼ 87th |
| | GRE Verbal | 169/170 ∼ 99th | 165/170 ∼ 96th | 154/170 ∼ 63rd |
| | GRE Writing | 4/6 ∼ 54th | 4/6 ∼ 54th | 4/6 ∼ 54th |
| | AP Art History | 586th ∼ 100th | 586th ∼ 100th | 586th ∼ 100th |

TABLE II
GPT-4 VS GPT3.5 BENCHMARKING PERFORMANCE COMPARISON ON PROFESSIONAL AND ACADEMIC EXAMS

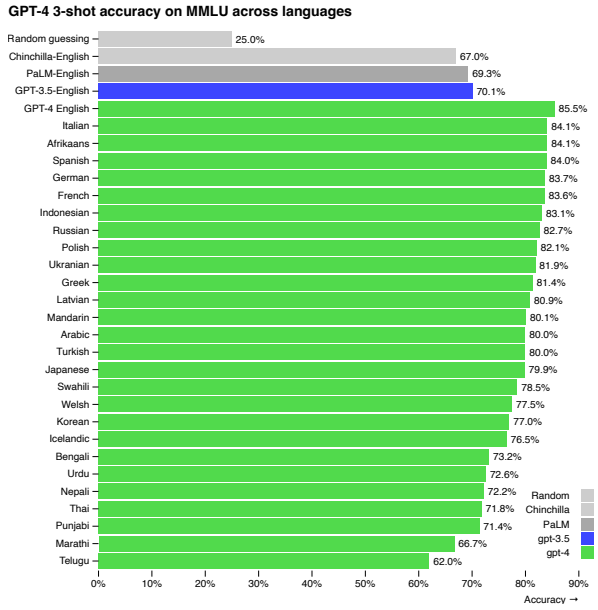



Fig. 4. GPT-4 Context Window Length Comparison with GPT-(1,2,3.x) [7]

The report aims to enhance transparency and understanding of GPT-4's capabilities, safety concerns, and the measures taken to mitigate potential risks and issues associated with its deployment.

In what follows, we provide a concise overview of the key limitations, among various other challenges:

- **Hallucination:** this represents one of the most critical problems in generative AI in general and GPT models in particular. This happens when the generative model produces non-sense reasoning or factually inaccurate content. OpenAI reported that GPT-4 significantly improved in reducing hallucinations compared to previous GPT-3.5 models (which have been improving with continued iteration). GPT-4 scores 19 percentage points higher than the latest GPT-3.5 on the OpenAI internal adversarially-designed factuality evaluations. While GPT-4 has already shown improvements in reducing hallucinations compared to GPT-3.5, continued efforts are needed to further minimize this issue.
- **Harmful content:** AI Generative models are subject to producing harmful content such as hate speech or invitation to violence. Two model versions of GPT-4 were analyzed: GPT-4-early (an early version fine-tuned for instruction following) and GPT-4-launch (a version fine-tuned for increased helpfulness and harmlessness). GPT-4-early reflects the risks when minimal safety mitigation is applied, while GPT-4-launch exhibits safer behavior due to the safety measures implemented. Over 50 experts were engaged in providing a more comprehensive understanding of GPT-4 and potential deployment risks in various areas. Building on the success of the GPT-4-launch, further advancements in safety mitigation can help ensure responsible AI deployment.
- **Disinformation and Influence Operation**: Influencing public opinion is critical as it drives people's opinions in the wrong direction. This is usually performed by injecting information through different channels like social media, news outlets, and other platforms to disseminate disinformation to a broad audience. Using sophisticated AI generative tools like GPT-4 can aggravate these issues considering their massive capabilities in manipulating information and how it is possible to expose that information. Overall, OpenAI reported that GPT-4 is better than GPT-3.5 in mitigating the effect of exploiting it to generate disinformation. However, it can still be used to some extent, such as improved persuasiveness to generate misleading but persuasive content. There is still a lot to do in this regard to prevent LLMs such as GPT-4 and earlier models from helping develop fake information and potentially influencing general opinion. It is crucial to enhance the model's ability to avoid being exploited for generating disinformation or persuasive misleading content.

## VI. Conclusion

This paper discussed the recent advanced brought by GPT-4 and how it compares with its predecessor GPT-3.5. In summary, OpenAI did not disclose technical information as was the case for previous models, including the architectural design and the training process, and only focused on discussing comparative results of GPT-4 with GPT-3.5 and a comprehensive assessment of different benchmarking tasks. The main differences highlighted compared to GPT-3.5 are mainly the models' size (170 Trillion for GPT-4 vs. 175 Billion for GPT-3.5), the size of the context length, the multimodality feature, which includes text and images as input, and the use of the Rule-Based Reward Models in their training process. GPT-4 was also reported to share similar limitations as GPT-3.5 but with a reduced effect.

While GPT-4 is a technology that would facilitate several new services and businesses, it is important for the community to work towards improving its non-functional aspects, such as safety and reliability, to limit the malicious usage of such an exciting technology for crimes and wrong actions.

Also, there is a growing need for the research and development community to develop similar open-source Large Language Models to avoid having monopolies by big giants and companies to take exclusivity in developing these services, thus concentrating the competition only at the industrial level. The academic research centers should work together to make similar systems available as open source for the benefit of knowledge sharing and transfer.

## Statement

During the preparation of this work, the author used ChatGPT to (i.) paraphrase his own sentences to improve readability, (ii.) search for particular information, and (iii.) summarize and clarify some text from related works and references. After using this tool/service, the author reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

## References


[1] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," *CoRR*, 2022.

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," *CoRR*, vol. abs/2005.14165, 2020.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[4] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," *Advances in neural information processing systems*, vol. 30, 2017.

[5] OpenAI, "Gpt-4 technical report," 2023, https://cdn.openai.com/papers/gpt-4.pdf.

[6] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving, "Fine-tuning language models from human preferences," *arXiv preprint arXiv:1909.08593*, 2019.

[7] L. Dr Alan D. Thompson, "Gpt-4 context window," March 2023, https://lifearchitect.ai/gpt-4/.

[8] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. de Oliveira Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry, P. Mishkin, B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, L. Kaiser, M. Bavarian, C. Winter, P. Tillet, F. P. Such, D. Cummings, M. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, W. H. Guss, A. Nichol, A. Paino, N. Tezak, J. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders, C. Hesse, A. N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish, I. Sutskever, and W. Zaremba, "Evaluating large language models trained on code," *CoRR*, vol. abs/2107.03374, 2021. [Online]. Available: https://arxiv.org/abs/2107.03374

[9] E. Ruane, A. Birhane, and A. Ventresque, "Conversational ai: Social and ethical considerations." in *AICS*, 2019, pp. 104–115.

[10] OpenAI, "Gpt-4 system card," 03 2023, https://cdn.openai.com/papers/gpt-4-system-card.pdf.