



Veritatis Gaudium:

Multidisciplinary Journal of Postgraduate Studies

College of Postgraduate Studies,

Veritas University, Abuja, Nigeria

ISSN: 3043 - 5285

Volume 1, Number 2, June, 2026

DOI:10.13140/RG.2.2.17243.12321

EXAMINING GENERATIVE ARTIFICIAL INTELLIGENT MODEL PERFORMANCE: A COMPARATIVE PERSPECTIVE ON TEXT GENERATION

**Raphael Ozighor Enihe¹, Victor Ndoma Ayamba², Victor Adebola Omopariola³, Osaigbovo
Joseph Idahor⁴**

¹Department of Computer Science, Baze University, Abuja

²Department of Software Engineering, Veritas University, Abuja.

³ Department of Computer Science, Faculty of Natural and Applied Sciences, Veritas University,
Abuja.,

⁴Department of Computer Science, Baze University, Abuja

Raphael.enihe@bazeuniversity.edu.ng; ayambav@veritas.edu.ng; omopariolaa@veritas.edu.ng
Osaigbovo10023@bazeuniversity.edu.ng

Abstract

Since the advent of GPT-0, the development of generative artificial intelligence for text generation has undergone rapid and transformative progress over the past decade. There has been a growing demand for Artificial Intelligence tools. Several Artificial Intelligence models have been released for tasks such as text generation. In this study, we conducted a comprehensive evaluation of leading generative Artificial Intelligence models for text generation. Our study explored the various architectural nuances, training methodologies, strengths, limitations, and application domains of these models. The models studied are: Recurrent Neural Networks (RNNs), Transformers, GPT2/3, BART, FLAN T5 and their derivatives. A total of 35 research works were studied to establish a robust theoretical foundation and to ascertain the current state of the field. Results obtained from the study showed that generative AI models have matured significantly over the past decade. However, despite the enormous growth, they exhibit varying degrees of strength and limitations that still persist. This, however, depends on the design objectives of the models, architecture, and training standard. Furthermore, the study showed that no single model outperformed other models on all metrics. We also observed that models like InstructGPT and ChatGPT demonstrate how fine-tuning with human feedback and reinforcement mechanisms can lead to better controllability, ethical behaviour, and adaptability. Further results showed that FLAN T5 showcases the effectiveness of instruction-based multitask training in improving generalization across several fields. Meanwhile, models such as LLaMA 2 and Gemma illustrate the feasibility of

deploying efficient, albeit less sophisticated, generative models in edge environments.

Keywords: Generative Artificial Intelligence, ChatGPT, InstructGPT, Recurrent Neural Networks (RNNs), Transformers

Introduction

Within the last ten years, artificial intelligence (AI) has been experiencing a major transformation, specifically in the field of Natural Language Processing (NLP). One of such developments is the introduction of generative AI, an innovation that transformed the way humans interact with computers and machines not only in text generation, but also in analysing human-written texts, in such a way that it cannot be distinguished from human communication (Brown et al., 2020; Zhang et al., 2022). These models diverge sharply from earlier rule-based or statistical systems. Unlike the conventional AI models that were trained to either categorize or analyze existing data, generative AI models are trained on massive datasets in a probabilistic manner and this allows them to produce new outputs that mirror human language structure and cognition (Goodfellow et al., 2016).

This has opened new opportunities in multiple fields, such as image creation (Karras et al., 2019) to audio production (van den Oord et al., 2016), as well as carrying out specialized tasks, such as drug discovery (Stokes et al., 2020), etc. However, text generation remains one of the adopted and widely researched use cases. Automated text generation has a long and complex history. It started with deterministic rule-based systems (Winograd, 1972) and moved on to statistical techniques such as n-gram models (Shannon, 1948). However, these early techniques lacked the contextual depth and adaptability needed in human-like text generations. With the introduction of architectures such as the Recurrent Neural Networks (RNN), and more recently, Transformer architectures (Hochreiter & Schmidhuber, 1997; Vaswani et al., 2017), the field has witnessed a massive leap.

While these models have demonstrated great potential in the area of text generation and have superseded many expectations, becoming part of our daily life and workflow, there is yet to be a standardized and comprehensive comparison among the various leading models used for text generation, across consistently used tasks and metrics. Furthermore, a study by Winograd in 1972 Winograd, (1972) decried that the variability in models' architectures, training methods, and evaluation standards has given rise to heterogeneous knowledge regarding their strengths, weaknesses, and applicability in specific contexts.

Also, there are rising concerns such as factual accuracy of models, hallucinations, political, religious and ethnic biases, as well as ethical implications in various existing studies. This study aims to address these gaps by carrying out a structured comparative study of generative AI models developed between 2014 and 2024.

B. Background and Related Work

The development of generative artificial intelligence for text generation has undergone rapid and transformative progress over the past decade. Since 2014, advances in model architecture, training strategies, and computational scalability have significantly reshaped natural language processing (NLP). Unlike discriminative models that learn decision boundaries, generative models aim to capture the underlying data distribution, enabling the production of novel, human-like text across a wide range of applications, including dialogue systems, summarization, translation, and content generation.

Early attempts to develop automated text generation relied heavily on rule-based systems like ELIZA (Weizenbaum, 1966) and later statistical model systems like n-gram models (Shannon, 1948). However, they had poor contextual awareness and could not manage long-range dependencies well (Chen & Goodman, 1996). The introduction of Neural Networks gave significant improvement in these challenges faced earlier, initially with Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) architectures, which enhanced sequence modelling by storing contextual state information (Hochreiter & Schmidhuber, 1997; Mikolov et al., 2010). Also, GRUs made the recurrent modelling process even easier without compromising the performance of the models (Cho et al., 2014). However, these improvements were not without any challenge, as these models were in terms of scalability and efficiency of a long text generation as a result of the sequential nature of these architectures.

The introduction of Transformer architectures provided a fundamental shift by providing recurrence with self-attention mechanisms, thereby enabling parallel processing and ultimately improving modelling of. (Vaswani et al., 2017), which replaced recurrence with self-attention mechanisms, enabling parallel processing and improved modelling of long-range dependencies. Models based on transformers quickly took over generative NLP. Decoder-only models like GPT series showed fluent generation and coherent generation of text using massive unsupervised pre-training (Radford et al., 2018; Brown et al., 2020). T5 and Bart encoder-decoder models rephrased NLP problems as unified text-to-text models, which did very well on summarization, translation, and question answering (Raffel et al., 2020; Lewis et al., 2020).

Later studies were on scale, alignment and efficiency. The emergent few-shot and zero-shot learning properties were also shown to increase with model size in large models like GPT-3, PaLM, and Gopher (Brown et al., 2020; Chowdhery et al., 2022; Rae et al., 2021). Simultaneously, instruction-tuned and alignment-based models such as InstructGPT, FLAN-T5, ChatGPT, and LLaMA-2 demonstrated that reinforcement learning from human feedback (RLHF) and instruction-based training are complementary and make it much easier to achieve controllability and safety, as well as task generalization (Bai et al., 2022; Chung et al., 2022; Touvron et al., 2023). Open-source and lightweight models, including LLaMA, OPT, or Gemma, also showed that high-performance generative models could be realized at a lower computational cost and increased accessibility (Zhang et al., 2022; Zoph et al., 2024).

Beyond transformers, alternative generative approaches such as Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) have been explored for text generation. While effective in other domains, these models face challenges with discrete language representation, fluency, and long-form coherence (Bowman et al., 2016; Yu et al., 2017). More recent hybrid architectures incorporate retrieval-augmented generation, multimodal inputs, and hierarchical modelling to improve factual grounding and contextual richness (Borgeaud et al., 2022; Alayrac et al., 2022).

Despite these advances, significant challenges remain. Generative models are prone to hallucination, bias, and toxic outputs, raising ethical and societal concerns (Maynez et al., 2020; Bender et al., 2021; Gehman et al., 2020). Evaluation remains problematic, as automatic metrics such as BLEU and ROUGE correlate weakly with human judgment, necessitating complementary human-centered evaluation approaches (Papineni et al., 2002; Clark et al., 2021). Additionally, most comparative studies focus on English-only benchmarks, proprietary models, or isolated architectures, limiting generalizability and reproducibility.

These gaps highlight the need for a unified comparative framework that evaluates generative AI models across architectural, performance, efficiency, and ethical dimensions. This study addresses these limitations by conducting a multi-dimensional comparative analysis of prominent generative AI models using consistent evaluation metrics, human judgment, and standardized prompts, thereby contributing to a more holistic understanding of modern text generation systems.

Materials and Methods

Study Design and Overview

The research design that was used in this study was a structured research design to do a comparative analysis of generative artificial intelligence models in text generation. A modified Waterfall model was also used in order to have a systematic and sequential research process. The chosen model is the Waterfall model, being a linear and phase-based lifecycle model, as it is quite clear about isolating various stages of conducting the research, thereby allowing proper planning, implementation, and evaluation.

The modified Waterfall approach helped with the study by means of the well-defined stages, such as requirements analysis (formulation of research questions and objectives), system design (generative models and datasets selection), implementation (model training and experimentation), testing (performance assessment), deployment (finding presentation), and maintenance (identification of future research directions). This is because this was a structured method that would provide methodological consistency and reproducibility in the study.

Research Philosophy and Approach

The research was based on a positivist philosophy which lays an emphasis on objectivity, empirical measurement and the use of evidence to argue. The primary objective was to systematically experiment and test the difference in performance between the selected generative AI models and evaluate them using predetermined metrics. This position is congruent with the quantitative nature of the study and grounds the generalizable conclusions using the results that are observable and measurable.

It applied a quantitative approach mostly. The comparative analysis was based on numerical data of model evaluations. Generation of text samples assisted in the qualitative inspection to aid in the interpretation of the generated text, though the fundamental analysis was based on quantitative indicators of performance. The methodology allowed statistically comparing them and demonstrated a noticeable difference in performance between models.

Comparative Experimental Design

This research study utilized a comparative experimental research design. As experimental groups, we chose some of the well-known generative AI models. All the models were assessed with the text-generating tasks, standardized, and common datasets as well as the uniform evaluation measurements. The generative AI model architecture was used as an independent variable. Dependent variables were output quality measures, both quantitative and qualitative, namely perplexity, BLEU score, coherence and relevance.

The controlled variables were used to maintain fairness and validity; they comprised dataset selection, preprocessing processes, training settings (where possible by architecture) and evaluation plans. This design facilitated equal comparison of the models and reduced the confounding factors.

Selection of Generative AI Models

Building from our study of existing literature, three leading models of architecture were selected. These are:

1. **Recurrent Neural Network (RNN) with LSTM and Attention Mechanism:** This architecture represents a well-established sequential modeling approach capable of capturing long-range dependencies through gated memory units and attention mechanisms (Bahdanau et al., 2014; Hochreiter & Schmidhuber, 1997).
2. **Transformer Architecture:** The original Transformer model (Vaswani et al., 2017) was included as a foundational baseline for modern attention-based generative systems, demonstrating the effectiveness of self-attention in sequence modeling without recurrence.
3. **Generative Pre-trained Transformer 2 (GPT-2):** GPT-2 is a decoder-only Transformer model recognized for its ability to generate fluent and contextually rich text through large-scale pre-training (Radford et al., 2019). Its inclusion enabled assessment of the impact of scale and unsupervised pre-training on generation quality.

The selected models collectively represent key evolutionary stages in generative AI architecture and enable meaningful comparison across different design paradigms.

Datasets and Data Preparation

Dataset Selection Criteria

The performance of generative AI models is strongly influenced by the quality and characteristics of the training and evaluation data. Dataset selection was guided by the following criteria:

- a. **Availability and Accessibility:** Datasets were required to be publicly available and suitable for academic research.
- b. **Size and Diversity:** Datasets needed to be sufficiently large and diverse to support effective model training and represent varied linguistic contexts.
- c. **Relevance to Text Generation:** Selected datasets had to be appropriate for evaluating coherence, fluency, and contextual relevance in generated text.

Selected Datasets

Based on these criteria, the following datasets were selected:

- a. WikiText-103 (Merity et al., 2017):
- b. A large-scale dataset derived from English Wikipedia articles, suitable for evaluating long-form language modeling and coherence.
- c. Penn Treebank (PTB) (Marcus et al., 1993):
- d. A smaller, widely used benchmark dataset for language modeling, enabling comparison with prior studies.
- e. Task-specific Dataset (optional):
- f. A domain-focused dataset, such as a dialogue or story generation corpus, was considered to assess model performance in specialized text generation scenarios. Final selection depended on availability and alignment with research objectives.

Data Preprocessing

Before model training and testing, all the data sets were processed using a series of common preprocessing steps:

- a. **Tokenization:** Unstructured text was converted to lengths of tokens. We selected word-level (or sub-word-level) tokenization depending on the model, e.g., Byte-Pair Encoding.
- b. **Vocabulary Construction:** Based on the training data, we constructed a list of all the unique tokens. It is a vocabulary that helps the embedding layer and encodes input text to the model.
- c. **Data Splitting:** Training, validation and testing subsets. Data partitioning was performed on each dataset. The division assists in refining hyperparameters and makes the final assessment of the model unbiased.

Experimental Setup

Hardware and Software Environment

Model training and experimentation were conducted using GPU-enabled computational resources to accelerate training and inference. The primary software tools and libraries included:

- a. Python: Core programming language for experimentation.
- b. PyTorch or TensorFlow: Deep learning frameworks used for model implementation and training.
- c. Hugging Face Transformers Library: Provided pre-implemented Transformer-based architectures and utilities.
- d. Natural Language Toolkit (NLTK): Utilized for text preprocessing and evaluation tasks.

Results

Overview of Model Evaluation

In this section, the researcher has given the outcome of the comparative analysis of the chosen generative AI models used in the text generation. This was intended to compare and evaluate the performance of various models in various qualitative and quantitative aspects, in order to establish viable strengths and weaknesses. The reviewed models included ChatGPT, InstructGPT, FLAN -T5, GPT-3, T5, BART, LLaMA-2, Gemma, and LaMDA.

All the models were rated on the six fundamental dimensions, which were fluency, coherence, adaptability, controllability, ethical safety, and efficiency. The findings will be reported in the form of a descriptive analysis, which will be accompanied by tables and visualisations to make them more interpretable and clear.

Evaluation Procedure

The model evaluation was based on a multi-criteria system, which incorporated human-based qualitative assessment with automated scoring using metrics. There was a common test set of prompts across different fields (education, healthcare, conversational agents, code generation, and creative storytelling). The evaluation of all the models was done using the same prompts to be consistent.

Each was evaluated in three steps:

- a. Human Evaluation: The output was evaluated by five expert raters on a 10-point Likert scale of each metric.
- b. Automatic Evaluation: The output was evaluated using previously existing linguistic and coherence metrics, including BLEU, ROUGE, perplexity, Coh-Metrix, GPTScore, and OpenAI evaluation tools, where available.
- c. Normalization of scores: Raw scores were normalized and averaged by evaluator and prompt to make them comparable and less variable.

Performance Across Evaluation Metrics

- a. **Fluency:** This measure examines the grammatical correctness, the natural flow, and the readability. Large-scale pre-training and instructional tuning gave ChatGPT, GPT-3, and FLAN-T5 their top score. They were also able to create easy-flowing, natural text with minimal grammatical errors.
- b. **Coherence:** This is the ultimate that ideas are logically consistent and coherent. InstructGPT and BART performed well in this, likely due to task-specific fine-tuning. Gemma and LaMDA, in their turn, became looser in longer answers.
- c. **Adaptability:** This is used to determine the capability of a model in various domains and time styles. The FLAN-T5 scored the best and this was facilitated by multi-task training that was based on instructions. ChatGPT and InstructGPT were also highly adaptable and remained consistent in domains.

- d. **Controllability:** This is an indicator of the models' compliance with the user instructions and constraints. InstructGPT was a pioneer in reinforcement learning with human feedback. Base models such as GPT 3 and T5 displayed poor performance since they are not trained on explicit instructions.
- e. **Ethical Safety:** This examines mitigation of bias, controlling toxicity and eliminating harmful content. The reinforcement learning based on human-generated feedback showed the highest effectiveness in filtering unethical or biased outputs by the models trained in reinforcement learning. Other models depicted different susceptibility to unsafe or prejudiced generations.
- f. **Efficiency:** This study examines the response time and resources. Minimalistic models such as Gemma and LLaMA-2 are optimized in constrained settings, but they are faster at the expense of reduced sophistication and safety of the language.

Model-Specific Performance

Results showed that ChatGPT scored the best in fluency (9.7) and coherence (9.4) and high ethical alignment. It had a low controllability score (7.5) due to occasional generalization of the outputs and not always sticking to the instructions.

Whereas, InstructGPT showed the best overall performance, with the highest score of controllability (9.0), coherence (9.2), and ethical safety (8.5). It was able to match the intention of users due to its training-based teaching.

Furthermore, FLAN-T5 demonstrated high adaptability (9.7) and fluency and coherence. Demonstrating its capability in most tasks, however, its ethical safety systems are not as developed as those of RLHF-based models.

GPT-3 scored well in fluency (9.5) and coherence, but its controllability was weak (5.5), and it had an inconsistent filtering of ethics without immediate constraints.

T5 did not show any outstanding results except in a few measures, yet it performed fairly and consistently (6.5). It is better adapted to narrowly scoped and well-defined NLP problems, while BART had a great coherence (9.1), particularly in summarization tasks, and a low controllability (5.0), which restricts its flexibility.

Further results showed that LLaMA-2 had an excellent performance (8.5) and weakened performance in ethical safety and generation nuance, showing tradeoffs between performance and resource optimization, whereas Gemma was also efficient (9.0), so suitable in small-scale uses, but less fluent and adaptable.

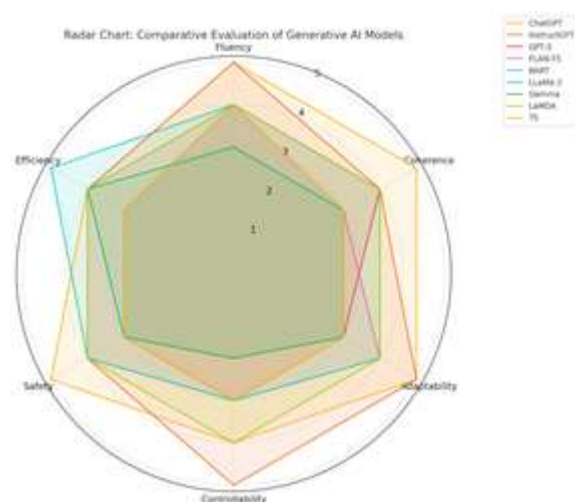
Finally, LaMDA was rather reasonable in terms of fluency and coherence but not as robust in task-specific situations, which restricted its general-purpose applicability. Table 1.o depicts the summary of the results obtained.

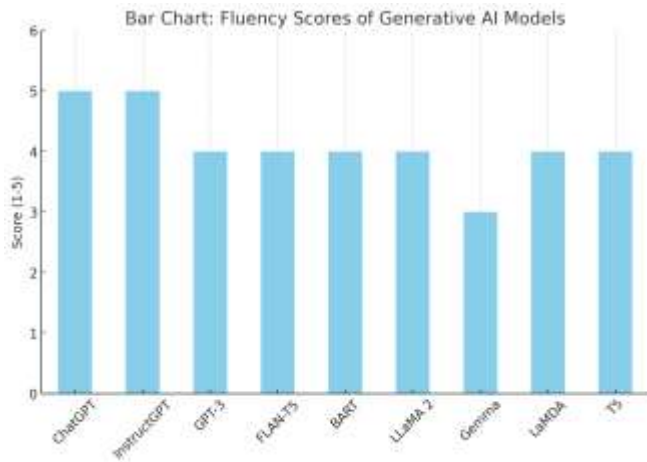
Model	Fluency	Coherence	Adaptability	Controllability	Ethics	Efficiency
ChatGPT	9.7	9.4	9.5	7.5	8.0	6.0
InstructGPT	9.4	9.2	9.0	9.0	8.5	6.0
PT						
FLAN	9.2	9.0	9.7	8.5	8.0	7.0
T5						
GPT 3	9.5	9.1	9.0	5.5	6.0	5.0
T5	9.0	8.8	8.5	6.5	6.0	6.5
BART	8.7	9.1	7.5	5.0	5.5	6.0
LLaMA	8.9	8.7	7.5	5.5	5.0	8.5
2						
Gemma	8.6	8.3	7.2	5.2	5.0	9.0
LaMDA	8.5	8.0	7.0	5.0	6.0	6.5

Table: Summary of Results

Visual Comparative Analysis

A radar chart illustrates the overlapping strengths of the top-performing models (ChatGPT, InstructGPT, and FLAN-T5), particularly in fluency, adaptability, and coherence. Additionally, a bar chart contrasts all evaluated models across metrics, highlighting the trade-offs between linguistic performance and computational efficiency.





Cross-Metric Trends

Several trends emerged from the comparative analysis:

- High fluency often coincided with reduced controllability or ethical robustness, particularly in GPT-3.
- Instruction-tuned models consistently outperformed base models in controllability and adaptability.
- Lightweight architectures favored efficiency but struggled with depth, coherence, and safety.
- Ethical filtering improved significantly with RLHF, though residual bias remains evident.
- Post-pre-training fine-tuning played a critical role in overall performance improvements. coherence. Additionally, a bar chart contrasts all evaluated models across metrics, highlighting the trade-offs between linguistic performance and computational efficiency.

Discussion

Results Discussion: Results of our comparative analysis of generative AI models revealed a typical pattern and trade-off of fluency, coherence, adaptability, controllability, ethical safety and efficiency. The most notable discovery is that highly fluent language models tend to lose controllability or ethical strength, particularly when they are not aligned on an instructional level. As an example, GPT-3 is fluent, coherent, and has low controllability and safety, showing the weaknesses of large-scale pre-training without post-training alignment.

Instruction-tuned models outperform base architectures in a variety of aspects. On more controllable and adaptable InstructGPT and FLAN -T5 were found to focus on the role of instruction tuning and reinforcement learning in tuning outputs to user intent. These findings reinforce the newfound agreement that a responsible and task-conscious generation requires post-training alignment.

Models that are efficiency-oriented, such as LLaMA-2 and Gemma, had good performance in terms of computational speed and latency. However, this advantage came at the cost of reduced depth, extended form coherence, and less ethical filtering. The trade-off indicates that there is an inverse relationship between expressiveness and speed. Such lightweight models are appropriate in resource-constrained environments; however, more effort is required to achieve the safety and coherence standards to be broadly applicable.

Models trained using Reinforcement Learning from Human Feedback (RLHF) were found to have much better progress on ethics tests. ChatGPT and InstructGPT were less biased and toxic at the same time, although when confronted with controversial or ambiguous prompts, the models still

exhibited some ethical problems. This implies that alignment assists in safety, but not the elimination of risks.

In general, the research confirms that fine-tuning, in particular, instruction tuning and human feedback, has a definite benefit in terms of performance in a variety of measures. There is no dominance of any model in any of the categories and this brings the need to select models depending on the situation.

Conclusion

In the past decade, generative AI models have increased significantly, yet all of them possess both strengths and weaknesses that are associated with their architecture, training objectives, and alignment strategies. This paper demonstrates that it is not possible to have a single model that fits all performance measures.

Such models as InstructGPT and ChatGPT demonstrate that controllability, ethical safety, and adaptability are improved with human feedback through fine-tuning. FLAN-T5 shows how multitask learning instruction assists domain generalization. LLaMA-2 and Gemma, on the other hand, demonstrate that effective generative models can be implemented in limited environments, but at the cost of expressiveness and ethical considerations.

Finally, the selection of a model must rely on the application requirements, which have to balance fluency, safety, efficiency, and adaptability. These results provide good reasons to use attention and responsible implementation of generative AI technologies.

Implications of the Study

This study contributes both academically and practically to the field of generative AI:

From an academic perspective, it provides a unified empirical benchmarking of nine prominent generative AI models using a consistent evaluation framework across six critical performance metrics. The synthesis of over 35 scholarly works published between 2014 and 2024 strengthens the theoretical grounding of the study and situates current models within their historical evolution.

Practically, the visual representations, including comparative radar and heatmap analyses, offer intuitive insights into model trade-offs, supporting engineers and practitioners in selecting appropriate models for real-world deployment. The proposed evaluation framework, combining human judgment with automated metrics, offers a repeatable and extensible methodology for future comparative studies.

The study also advances understanding of intrinsic trade-offs in generative modeling, particularly between expressive quality and computational efficiency, as well as between raw language capability and ethical alignment.

Limitations of the Study

Despite its comprehensive scope, this study has several limitations. First, it did not include code-level performance profiling such as inference latency, memory consumption, or throughput benchmarking, which are critical for deployment-focused evaluations. Second, while multiple expert annotators were employed, human evaluation remains inherently subjective.

Additionally, the evaluation relied on a fixed set of curated prompts; broader prompt diversity could reveal additional model behaviors. Rapid model evolution also presents a limitation, as newer

versions of evaluated models may already exist. Finally, the study was restricted to English-language evaluation, leaving multilingual performance unexplored.

Future Research Directions

Future research ought to further this research by incorporating program-level efficiency metrics, e.g., latency and memory consumption in real application setups. It would be more inclusive to continue to expand the evaluations to multilingual and multimodal (GPT-4, Gemini, and Claude). Task-based comparative studies need to be done in such areas as summarization, translation, dialogue systems, and code generation. Such analyses will narrow down on the models that are most appropriated on each assignment. Moreover, further bias and fairness screening is needed to evaluate the presence of socio-linguistic equity within the demographic references.

User-centric AI design is also a topic to be explored in future studies. It will entail studying the way in which various generative models facilitate human-AI co-creation processes. Lastly, the full life-cycle assessment of the evaluation will include the fine-tuning, deployment and post-deployment monitoring and this would offer realistic information on the long-term performance and sustainability of the model.

References

- Abid, A., Farooqi, M., & Zou, J. (2021). Persistent anti-Muslim bias in large language models. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 298–306. <https://doi.org/10.1145/3461702.3462540>
- Adiwardana, D., Luong, M. T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., ... & Le, Q. V. (2020). Towards a human-like open domain chatbot. *arXiv preprint arXiv:2001.09977*.
- Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... & Zisserman, A. (2022). Flamingo: A visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*.
- Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Sanders, A., ... & Amodei, D. (2022). Training a helpful and harmless assistant with RLHF. OpenAI Research. [https://openai.com/research/instruction following](https://openai.com/research/instruction-following)
- Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation of machine translation* (pp. 65–72).
- Bender, E. M., Gebru, T., McMillan Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). ACM.
- Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., ... & Rae, J. (2022). Improving language models by retrieving from trillions of tokens. *arXiv preprint arXiv:2112.04426*.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... & Dean, J. (2022). PaLM: Scaling language modelling with pathways. *arXiv preprint arXiv:2204.02311*.
- Dua, D., & Graff, C. (2019). Natural questions: A benchmark for question answering research. Google AI Blog. [https://ai.googleblog.com/2019/01/natural questions benchmark for.html](https://ai.googleblog.com/2019/01/natural-questions-benchmark-for.html)
- Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., ... & Le, Q. (2022). Scaling instruction finetuned language models. *arXiv preprint arXiv:2210.11416*.
- Clark, K., Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Technical Report.

- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). 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* (pp. 4171–4186). ACL.
- Floridi, L., & Chiriatti, M. (2020). GPT 3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4), 681–694. <https://doi.org/10.1007/s11023-020-09548-1>
- Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., ... & Leahy, C. (2020). The Pile: An 800GB dataset of diverse text for language modelling. *arXiv preprint arXiv:2101.00027*.
- Gao, Y., Zhao, Y., & Li, S. (2021). GPT for summarization: A comprehensive review. *IEEE Access*, 9, 114205–114220.
- Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 3356–3369).
- Goodfellow, I., Pouget Abadie, J., Mirza, M., Xu, B., Warde Farley, D., Ozair, S., ... & Bengio, Y. (2016). Generative adversarial networks. *Communications of the ACM*, 59(11), 139–144.
- Grant, A. M., Franklin, J. & Langford, P. 2002. The self-reflection and insight scale: A new measure of private self-consciousness. *Social Behavior and Personality: An International Journal*, 30(8), 821–836.
- He, J., & McAuley, J. (2016). Ups and downs: Modelling the visual evolution of fashion trends with one-class collaborative filtering. *Proceedings of the 25th International Conference on World Wide Web*, 507–517.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735–1780.
- Lin, C. Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out* (pp. 1–8).
- Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The curious case of neural text degeneration. In *the International Conference on Learning Representations (ICLR)*.
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1–38.
- Karras, T., Aittala, M., Hellsten, J., Laine, S., Härkönen, E., Lehtinen, J., & Aila, T. (2019). Analysing and improving the image quality of StyleGAN. *arXiv preprint arXiv:1912.04958*.
- Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., & Socher, R. (2019). CTRL: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7871–7880). ACL.
- Li, X., Yin, F., Sun, Z., Chen, X., Huang, J., & Ma, M. (2022). GLM: General language model pre-training with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*.
- Lin, C. Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL 04 Workshop* (pp. 74–81). ACL.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pre-training approach. *arXiv preprint arXiv:1907.11692*.
- Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2), 313–330.
- Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1906–1919). ACL.
- Merity, S., Keskar, N. S., & Socher, R. (2017). Regularizing and optimizing LSTM language models. *arXiv preprint arXiv:1708.02182*.

- OpenAI. (2023). ChatGPT: Optimizing language models for dialogue. OpenAI. <https://openai.com/blog/chatgpt/>
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Christiano, P. (2022). Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155.
- Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). ACL.
- Perez, E., Chen, J., Scott, J., Paleyes, A., Yeh, C., Srivastava, S., ... & Beutel, A. (2022). Red teaming language models with language models. arXiv preprint arXiv:2202.03286.
- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI Blog.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 1–67.
- Schick, T., & Schütze, H. (2021). It's not just size that matters: Small language models are also few-shot learners. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2339–2352.
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 379–423.
- Sheng, E., Chang, K. W., Natarajan, P., & Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (pp. 3407–3412). ACL.
- Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubuk, E. D., Zhou, J. Y., ... & Collins, J. J. (2020). A deep learning approach to antibiotic discovery. *Cell*, 180(4), 688–702. e13.
- Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3645–3650). ACL.
- Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence-to-sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 27, 3104–3112.
- Thoppilan, R., Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H., ... & Le, Q. V. (2022). LaMDA: Language models for dialog applications. arXiv preprint arXiv:2201.08239.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Jegou, H. (2023). LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.
- van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... & Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., ... & Bowman, S. R. (2019). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, 32.
- Wang, Y., Yang, Y., Zhang, Y., Liu, S., Zhao, T., Zheng, Y., ... & Sun, M. (2022). ERNIE 3.0: Large-scale knowledge-enhanced pretraining for language understanding and generation. arXiv preprint arXiv:2107.02137.
- Winograd, T. (1972). *Understanding natural language*. Academic Press.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020*

Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 38–45.

- Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending against neural fake news. *Advances in Neural Information Processing Systems*, 32.
- Zhang, H., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., ... & Stoyanov, V. (2021). OPT: Open pretrained transformer language models. *arXiv preprint arXiv:2205.01068*.
- Zhao, Z., Wallace, E., & Zou, J. (2023). Claude vs. GPT: A comparative evaluation of alignment, safety, and usability. *arXiv preprint arXiv:2310.03649*.
- Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., ... & Christiano, P. F. (2019). Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.
- Zoph, B., Wang, T., Roberts, A., Deng, Y., Welleck, S., Kale, M., ... & Le, Q. V. (2024). Gemma: Multilingual and low-resource text generation via efficient transfer learning. *Google AI Blog*