IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 30, No. 1, January, 2025, pp. 31-41
Journal homepage: http://iieta.org/journals/isi

Hallucinations in GPT-2 Trained Model

Deniz Safar, Mohammed Safar*, Shayma Jaafar, Borkan Ahmed Al-Yachli, Abrar Khaled Shukri, Mohammed H. Rasheed

Department of Computer Technology Engineering, Technical Engineering College Kirkuk, Northern Technical University, Kirkuk 360001, Iraq

Corresponding Author Email: mohammed.sefer@ntu.edu.iq

https://doi.org/10.18280/isi.300104

Received: 6 October 2024
Revised: 5 January 2025
Accepted: 14 January 2025
Available online: 25 January 2025

Keywords:
GPT-2, hallucinations, transformers

ABSTRACT

This paper analysis the phenomenon of "hallucinations" in text generated by GPT-2 when it produces irrelevant or illogical content. This work has quantified the extent of those hallucinations and look into ways of their mitigation. By using two main techniques: cosine similarity and frequency analysis. These techniques calculate coherency and relevance in the text produced by OpenAI GPT-2 at different training levels. Where a study case was implemented to train the model and ask the questions and retrain the model using these replays. The main findings indicate that this model hallucinates much less at the beginning of learning, with the situation significantly improving as training progresses. Extreme learning does not eliminate all such inadequacies, and more over-training led to more hallucinations. The hallucinated items span from smaller deviations to major content-wise deviations. An inspection reveals some patterns and cues that are predictive of increased output unreliability of the model. This research suggests a stricter training program that involve varied data sets to reduce the rate of hallucinations. More importantly, improve the accuracy of the model by reaching superior levels through the embedding of contextual and factual anchoring systems as well as designing algorithms for higher trigger identification. Other recommendations of the paper include post-generation text evaluation and continuous research to enhance the complexity of the models.

1. INTRODUCTION

GPT-2 is a state-of-the-art unsupervised language model of text generation that is coherent, relevant given contexts, and conditioned upon input. Produced by OpenAI as a revolutionary move in the natural language processing field. And because of large-scale training with internet textual sources of great diversity a text artificially can be created, then 'hallucinations' can be occurred in those places where models generate information that maybe either meaningless, wrong factually, or in no way connected to the context that has been trained. Even so, highly sophisticated models like GPT-2 are not immune to errors. While some hallucinations may just be minor mistakes, others may involve serious fabrications or distortions of facts, most of which make reliance on its outputs for critical applications quite challenging [1]. Recently research has founded a number of dimensions with GPT-2 or the advanced ones. For instance, Shahriar and Hayawi [1] conducted research on machine-learning models to categorize material written by humans from that written by machines. They observed that even while the machine-learning models are able to distinguish between both-especially for more innovative tasks or storytelling-classifying text generated by GPT-2 would be like splitting hair [1]. The machine learning models were quite easily able to distinguish an expression of a human being from that of text written through the machine. The scholars worked on including the metadata related to textual level of readability, complexity, and sentiment in relation to some textual elements; these are pivotal as they can help one grasp the subtleties of the AI text [2]. In this survey, Yang et al. [3] researched the challenge of aligning two textual data distributions, particularly focusing on the alignment of subjects and sentiments. The researchers have analyzed textual distribution patterns using the supervised and unsupervised methodologies along with fine-tuned GPT-2. The authors proposed a new method, KL Textual Distributions Contrasts (KL-TDC), for ascertaining alignment in machine-generated textual collections. Whereas Samuel et al. [4] evaluated the classification of texts produced by humans and those generated by AI, focusing on both new and traditional features, it is quite fascinating that the data presents challenges in identifying texts that have been reworded using artificial intelligence or vice versa. This highlights how new features can greatly improve classifier accuracies. The area of concern in the discussion is actually the quality of texts generated, by the state-of the-art language models such as GPT-2. A study conducted by the Bangura et al. [5]. has shown that the text that has been created is most often of low quality from the perspective of manual qualitative measurements since the text produced always needs modifications. This goes on to underline the shortcomings in methods used to generate content always of higher quality. Provided an assessment to

classify text offered by humans and AI in multiple languages according to its nature. They were able to get their identification of text produced by AI up to a very high level of accuracy. This insight is proof of how far the capabilities in AI for content identification have come. But at the same time, this only goes on to reiterate that the problem in accurately discriminating text generation by AI from content generation by humans is here to stay [6].

GPT-2 is a state-of-the-art language model developed by OpenAI based on the Transformer architecture introduced by Vaswani et al. in their seminal work "Attention Is All You Need" [7]. While unlike traditional variants of sequential models such as RNNs and LSTM networks in GPT-2 the self-attention mechanism is used to parallelize the words without the need for processing them in succession. This approach improves the capability of understanding long-range connections or contextual linkages in text and thereby yields more coherent outputs. The architecture of GPT-2 uses many layers of self-attention and feed-forward neural networks that are really helpful in modeling complex linguistic patterns. In this regard each layer will compute the attention scores which would evaluate the relevance of a word in a given context to make the content coherent in context and meaning. It undergoes unsupervised learning over gigantic datasets such as WebText across many domains thus giving the model its excellent generalization capability. GPT-2 is an autoregressive model the prediction of the next word in sequence comes from previous words. This process involves pretrained embeddings and positional encodings of the words so that the model should be aware of which word comes after another during output.

However, GPT-2 has a number of drawbacks-most especially the possibility of generating "hallucinations" outputs that can be factually wrong or contextually inappropriate. Most such errors happen as a result of biased training data or a limitation in understanding domain specific scenarios. Understanding these issues is necessary to address these challenges especially in those applications that need factual accuracy like documents related to the law or medicine [8]. Therefore, comprehension of strengths and weaknesses will help improve its dependability and usability.

1.1 Problem statement

Hallucinations in outputs of models involve text, where language models like GPT-2 generate unrelated and nonsensical output or even incorrect, and this has been the justification for much discussion in the domain of artificial intelligence.

To illustrate the issue the following examples of fabricated outputs which were found in this research:

(1) Logical Inconsistency: In response to a question such as "What does OSI stand for, and what is its purpose?" the model once responded with "OSI stands for 'Organizational Structure and Management System'". This is just one example of how the model can produce believable yet factually incorrect information after long training cycles.

(2) Factual Inaccuracies: In the place where it was required to "Explain the key functions of the Transport Layer in the OSI Model" the response given was that "The Transport Layer in the OSI Model is a network protocol used to connect computers and other devices to a network." That is not specific and full of detail instead of technical content.

These examples illustrate the need to find these errors in a systematic way and to correct them since such findings have been reported in recent works focused on factuality in AI-generated text [8, 9]. These experiments show that GPT-2 frequently outputs text which appears credible but does not factually support its claims, undermining trustworthiness in critical applications.

That is issue of primordial concern for quite a few reasons, however:

AI-generated hallucinations of this kind have a great deal to do with undermining the reliability and believability of such systems crucial for serious applications in the field of news dissemination, customer services, or education.

The fallacious information by AI might form the basis for wrong decisions, especially in sensitive sectors such as medical and legal services.

Hallucinations often reflect underlying biases or quality problems in the training data, which further propagate systemic biases and unfair outcomes.

Hallucinations can be detrimental to both the user experience in general and the consumer-perceived effectiveness of an AI system or agent.

Therefore, these issues should be fixed to ensure that AI is used ethically and effectively in many fields. We align our research with this as the key point. We aim to measure and minimize such inaccuracies so as to enhance the overall quality and dependability of textuality produced by AI.

The main two major reasons that can be attributed to GPT-2 generating erroneous text are biases in the training data and architectural constraints inherent therein into transformer-based models. Firstly, biases inherent in the training datasets which are usually derived from web-scraped text, reflect predominant societal, cultural and linguistic prejudices hence causing distorted outputs to occur. Such biases are enhanced because the model learns patterns indiscriminately and without distinguishing between fact and opinion or misleading content. This limitation in GPT-2 is a structural one since the model relies on statistical correlations rather than genuine comprehension which results in superficial coherence without profound semantic understanding [8]. Although the GPT-2 transformer architecture serves quite well with short-term dependence and it performs poorly concerning long range coherent context yielding reasonable sounding text which contains factual errors. Another way which causes factual error is that usually generated information is not further checked against externally known knowledge. To handle such issues, data curation methods need to be improved and adversarial training should be carried out to reduce biases while architectural designs should be advanced to allow external verification mechanisms, as recent works have suggested [10, 11].

1.2 Objectives

This study's main objective is to examine and alleviate hallucinations in texts produced by the GPT-2 AI language model. This research is of utmost importance

for:

(1) Improving Model Reliability: Our objective is to enhance the dependability and factual accuracy of AI-generated texts by discovering and minimising mistakes. This is crucial for ensuring responsible utilisation of these models in many fields.

(2) Promoting the progress of AI research: Our research adds to the wider domain of artificial intelligence by tackling a significant constraint of existing language models, hence

facilitating progress in AI technology.

(3) This study conducts cosine similarity analysis and frequency analysis as the main methodologies in analyzing and mitigating hallucinations in GPT-2-generated text while focusing on textual accuracy and coherence. Cosine similarity analysis infers semantic coherence of the created text comparing the outputs at different training phases. High ratings of similarity denote logical and contextually relevant responses while low values underline discrepancies and errors. Frequency analysis on the other hand tracks the frequency of certain words and phrases over several training iterations. While changes in word frequency patterns provide information about content drift and help to identify consistent outliers that are indicative of hallucinations.

The assessment process is informed by key metrics: semantic coherence, consistency and error rate monitoring. Cosine similarity scores ensure that responses are on par with the given prompts for semantic coherence. Consistency and stability are assessed through responses across incremental training phases for divergence or stability. Error rates are tracked by monitoring changes in word frequency as a means of detecting trends related to factual errors or irrelevant information. These methods put together a quantitative framework for assessing hallucinations and testing ways to mitigate them hence improving the dependability of AI-generated outputs.

2. LITERATURE REVIEW

Creativity is on of talents or abilities possessed by a human to create or generate any new thing, mostly by self-generation or self-creation. GPT-2, a prestigiously pre-trained Transformer 2, marked a major breakthrough in the natural language processing field, particularly natural language generation. OpenAI has designed GPT2 with the Transformer architecture—a discovery that opened the doors to robots being able to understand and generate human language to an extent. The Transformer model, proposed by Vaswani et al., deviates from the earlier models that mostly include the RNN and LSTM, with their associated problems of vanishing or exploding gradients and issues of parallelization. This effectively addresses the problem through its use of the attention mechanisms; this way, models can process words together but with a focus on different parts of that input sentence at the same time. This makes them work better and more effective in their efficiency while working in translation, summarization, and text generation [12]. GPT-2 was also known for its massive scale and conscientiously crafted nature. The model itself, measuring 1.5 billion free parameters, had pre-training on a large corpus, WebText, with textual material synthesized from 45 million Internet links. At that massive scale, GPT-2 can generate coherent, contextually sound text over long stretches. Most of these enhancements pertain to what it mainly works for normalization and initialization within the layer, larger size in vocabulary, and context over its predecessor version, the GPT. All these combines to make it a more powerful model for language modeling tasks and generation of text [7].

As seen in Figure 1, Vasireddy et al. recently showed a new way of combining Vision Transformers (ViTs) with GPT-2 models and a new use case towards which this methodology could be applied. Feature extraction from the images yields one of the major things in this approach: generation of context-aware human-like explanations, which reinforces the flexibility of GPT-2 in multimodal communication scenarios. This scheme has the potential to make image-based data accessible to a larger audience, respectively heightening user experience. It is a giant step toward AI-driven content creation [13].

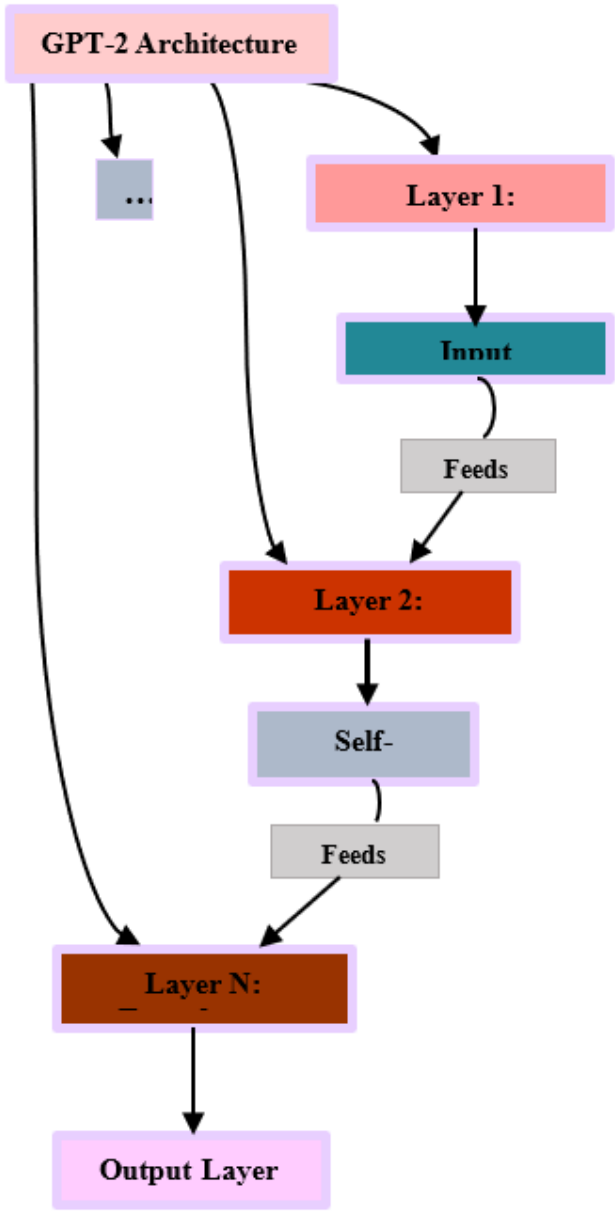


Figure 1. GPT-2 architecture

GPT-2, is transformer model, however, it's differing from previous sequential models such as recurrent neural networks (RNNs) and long short-term memory networks (LSTMs). The Transformers, first proposed by Vaswani et al. in their influential study" Attention Is All You Need," established a new mechanism called self-attention. The mechanism enables the model to assess the importance of individual words in a phrase, regardless of their position within the sentence. Transformers are highly effective for language modelling by overcoming the limitations of RNNs and LSTMs in dealing with long-range dependencies in text [7]. GPT-2's architecture, which consists of numerous layers of transformers, allows GPT-2 to effectively capture complex language patterns. The model has to be trained on a wide variety of datasats, enabling it to produce logical and contextually appropriate text on different subjects and in

different ways. GPT-2 utilises its predictive capabilities to generate text by expecting the next word in a sentence, taking into account all previous words. This makes it highly proficient in tasks like as translation, summarization, question-answering, and even creative writing. Yet, the GPT-2 and transformers also present difficulties. A major issue is the model's proclivity to produce 'hallucinated' content. This problem is most obvious in applications that demand a high level of factual precision, as generating news articles or creating academic literature.

Hallucinations in contents generated by AI occurs when language models or neural machine translation (NMT) systems gives an output that are illogical, irrelevant, or factually inaccurate. This issue presents a substantial difficulty in diverse applications which including academic writing. Neural machine translation hallucinations demonstrate that hallucinations frequently arise when the model always focuses on the same source tokens during the inference process. This statement gives a rise to the developing the "Static Source Contribution Hypothesis" that proposes the allocation of source contributions remains constant during model hallucination. In this order to examine by researchers they employed source alarms to intensify NMT hallucinations and subsequently compared the hallucinated outputs with non-hallucinated ones. This methodology facilitated a regulated comparison and aided in comprehending the fundamental principles behind hallucinations in NMT systems [14].

ChatGPT also makes a manufactured hallucination where entire phrasing machine generated deceptive content addresses content whose author is a machine, which seems credible but is actually factually errors and could be totally misleading in academic setups where factuality expects precision. This study found that when given a scientific publication title and asked to write an abstract, human reviewers often had difficulty telling whether the abstract was authored by AI or by a human. Yet the impressive percentage of the AI-generated abstracts recorded plagiarized contents, according to tools to detect plagiarism and AI. This provides insight into the potential and the pitfalls of using AI in academic writing, a domain where retention and accuracy of knowledge are paramount [15].

3. RELATED WORKS

A study by Bano et al. [16] shows that carefully selected collection of articles from Wikipedia website and ChatGPT generated texts. The study was focusing on developed algorithm that are capable of distinguishing between texts written by humans or those generated by AI. This research is crucial for understanding and tackling the issue of hallucinations in AI-generated literature. And the study finds a distinctive patterns and traits exclusive to AI-generated texts which frequently exhibit hallucinatory elements through the analysis of a substantial corpus. Where these hallucinations can appear as strange or unrelated material or as minor errors that may not be immediately noticeable. Accurately discerning between texts generated by artificial intelligence and those written by humans is essential for upholding the credibility and dependability of information, particularly in vital fields such as media, academia, and legal documentation. In a separate study conducted by Gunser et al. [17]. The objective was to determine if readers are able to distinguish between texts generated by artificial intelligence and those written by humans and to measure their perception of the formal characteristics of these texts. The results suggest that the readers had a generally low level of correctness in differentiating between the two categories of texts. The AI-generated has deemed to be less proficient, captivating, and engaging in comparison to the continuations written by humans and the original ones. This study emphasizes the nuances of hallucinations in AI-generated texts where the inaccuracies may not always be obvious falsehoods but instead appear as a deficiency in originality. The research highlights the necessity for improvements in AI language models to generate content that is more akin to human language captivating and precise. While Tulchinskii et al. [18] introduce the concept of central dimensionality as a measure to reliably differentiate between human-written texts and AI-generated texts, they do not specify how it is applied. The study revealed that the average intrinsic dimensionality of natural writings in alphabet-based languages and Chinese language can be greater than that of texts generated by artificial intelligence. This statistical differentiation offers a method to identify hallucinations in writings generated by artificial intelligence. The study's results use for development of advanced tools to identifying hallucinations in AI-generated content. Another work has made a noteworthy advancement by successfully implementing a solitary layer of GPT-2 on FPGA for edge devices. This study highlights the capabilities of GPT-2 in several hardware settings with a specific emphasis on its performance and complexity. Their research demonstrates the versatility of GPT-2 beyond traditional computing systems enabling its use in a broader array of applications, especially in contexts with limited resources [19].

While previous work has contributed to the evaluation of AI-generated text and the detection of hallucinations, gaps remain relating to the understanding of contextual and methodological limits of such techniques. For example, Bano et al. [16] were limited to the discrimination of texts generated by AI and those authored by humans and did not present any contextual analysis scheme for reviewing coherence or relevance across diverse domains. Similarly, Gunser et al. [17] reported evaluations of AI-generated text by users; however, the sole reliance on subjective evaluations reduced the ability to generalize to larger datasets.

This work extends the prior art by adding cosine similarity and frequency analysis for the systematic measurement of hallucinations at different levels of training. Unlike the methods that focus on detection only our approach goes further in investigating the patterns of progression explaining how model training influences the tendencies of hallucination. This paper points out weaknesses in the current classification approaches especially those which fail to consider domain-specific contextual errors, and proposes methods for improving these weaknesses.

In this paper we investigate the relative importance of data quality versus training scale in determining the frequency of hallucinations in text generated by GPT-2. This work lays the groundwork for future research utilizing this framework in pursuit of superior hallucination detection methods and stronger language models.

4. METHODOLOGY

The methodology of this study involved several key phases in utilizing the GPT-2 model representing an advanced text generator developed by OpenAI. The version used of GPT-2

showed very impressive fluency in generating text that is diverse and complicated so it would be proper to explore hallucinations consisting of AI-written content. The GPT-2 model draws from various datasets including literature across genres, scientific research papers and web knowledge. Such vast coverage of information is needed for appropriately replicating real-world scenarios. The training of the GPT-2 model was conducted progressively in multiple stages. The model was first trained simply and then the complexity and scope of training were increased since this allows us to trace how the capacity for generating text and creating hallucinations developed during these different stages in this work and forms an underlying methodology. Post-training the model produced text by utilising a predefined set of prompts that were specifically designed to encompass a broad spectrum of subjects and writing techniques. The objective was to assess the model's capacity to generate content. For detecting and quantifying the extent of hallucinations in generated texts the system had to institute criteria dependent on factual accuracy relevance to the prompt and coherence in the given context. Texts that were too far from these criteria were classified as hallucinations. The findings were recorded with caution and looked into concerning the levels of training to establish the frequency and features of the hallucinations. The analytical methods used included cosine similarity and frequency analysis to determine the condition of a given hallucination. Another use of cosine similarity analysis could have been in quantifying semantic congruence between AI generated texts with a reference text set. In this case higher scores would represent semantic coherence and on the lower end where it would approximate toward indicating an increased likelihood of hallucinations. Frequency analysis involved recording specific words and phrases that cropped up during the different stages of training. Considerable changes in these could thereby point toward a possible hallucination. Technologically this work varied from the application of a wide array of tools to back our research. The GPT-2 model served as the core tool for text generation aided by Python in scripting and analysis of data drawing on its comprehensive gatherings in natural language processing. MATLAB provided an essential use in data visualization and statistical analysis. It created compact graphics for trends like cosine similarity scores and word frequencies. Also, the Natural Language Toolkit in Python served as a benchmark for text processing while numerical computations were performed by the Python libraries SciPy and NumPy. A Python script used the transformers library from Hugging Face to fine-tune the GPT-2 model onto a dataset. It involved initializing the model and tokenizer and setting up training parameters for batch sizes and training epochs. The training was managed through an object of the class Trainer. It orchestrated how through iterative training cycles and the model would learn from the dataset to return an optimized model and preserve the tokenizer for further uses or deeper exploration. This methodology provided a very decent framework for understanding the flexibility and growth of GPT-2 pushing its ability to generate text that is contextually appropriate and coherent, as well as being prone to hallucinations, as seen in Figure 2.

The use of cosine similarity and word frequency analysis as evaluative methods is justified by their established usefulness in NLP and text similarity assessment. Cosine similarity measures the semantic closeness of vectors hence making it appropriate for coherence and relevance checking in generated text [20]. It is especially good at measuring the degree of overlap between input prompts and generated responses hence helping to identify aberrations that might suggest hallucinations [21].

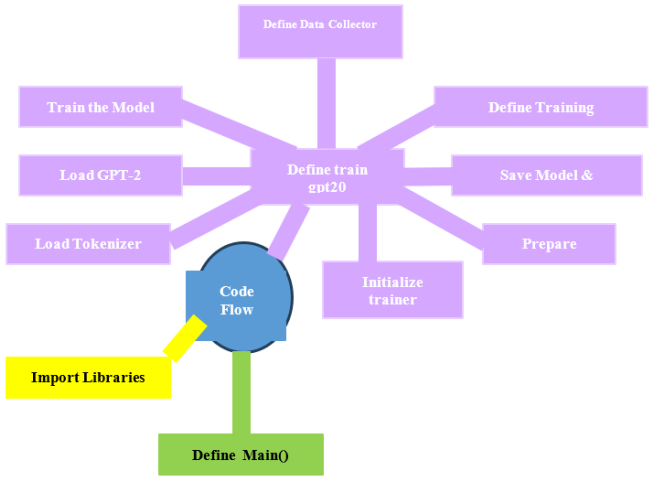


Figure 2. The methodology

On the other hand, word frequency analysis serves as a supporting technique in order to observe and the pattern of language changes in vocabulary and recurring structures that may indicate anomalies in content generation. So, the frequency analysis also agrees with prior research that examined thematic shifts and differences in language within AI-generated content using statistical distributions [22].

The proposed approach will make hallucination identification more robust combining a dual-method approach that balances semantic evaluation with statistical pattern recognition and to addressing abnormalities in content coherence and lexical distribution.

5. RESULTS

The tables below show the basic questions and replays by the model with fixed numbers of over-training based on the answers from previous trains.

5.1 Number questions

(1) What does OSI stand for, and what is its purpose?
(2) How many layers are there in the OSI Model, and can you name them?
(3) What is the primary function of the Physical Layer in the OSI Model?
(4) Describe the responsibilities of the Data Link Layer in the OSI Model.
(5) What is the role of the Network Layer in the OSI Model?
(6) Explain the key functions of the Transport Layer in the OSI Model.
(7) What is the primary purpose of the Session Layer in the OSI Model?
(8) What does the Presentation Layer in the OSI Model handle?
(9) Which OSI layer provides network services directly to user applications?
(10) How do Network Interface Cards (NICs) facilitate data communication on a LAN?
(11) What is a MAC address, and where is it used in a

network

(12) Differentiate between hubs and repeaters in terms of their functionality.

(13) What is the primary role of routers in computer networking.

(14) What is the primary role of routers in computer networking.

5.2 Number the answer with 1 training

(1) OSI stands for "Open Systems Interconnection," and its purpose is to provide a conceptual framework and guidelines for standardizing the design and operation of computer networks.

(2) The OSI Model consists of seven layers. Here are the seven layers of the OSI Model, listed from the bottom (Layer 1) to the top (Layer 7).

(3) The primary function of the Physical Layer in the OSI Model is to define the physical and electrical characteristics of the communication.

(4) The Data Link Layer, which is the second layer in the OSI Model, has several key responsibilities in network communication.

(5) The Network Layer, which is the third layer in the OSI Model, plays a fundamental role in network communication.

(6) Transport Layer is responsible for end-to-end communication, segmentation and reassembly, error handling, flow control, multiplexing and demultiplexing.

(7) The primary purpose of the Session Layer in the OSI Model is to establish, manage, and terminate sessions or connections between two devices on a network. Sessions enable ongoing communication and coordination between applications on different hosts.

(8) Which is the sixth layer in the OSI Model, is responsible for handling the following tasks related to data presentation and formatting.

(9) Application Layer.

(10) Also known as network adapters or network cards, are hardware components or devices that connect computers, servers, and other devices to a network.

(11) Media Access Control address, is a unique identifier assigned to a network interface controller.

(12) hubs are multi-port devices that broadcast incoming signals to all connected devices, On the other hand, repeaters are dedicated devices that actively amplify and regenerate signals.

(13) The primary role of routers in computer networking is to connect different networks and facilitate the routing of data packets between these networks.

5.3 Number the answer after 50 training

(1) OSI stands for "Organizational Structure and Management System"

(2) OSI Modeling is a layer that is used to model the physical properties of a system. It is used to determine the physical properties of a network.

(3) The first step in designing a layer is to create a layer that is both transparent and transparent to light. The layer is composed of two layers: The transparent layer and the transparent layer.

(4) The DLLs are used to connect computers to the Internet. They are used to connect computers to the Internet and to communicate with each other.

(5) The Layer is a layer between the network and the host.

(6) The Transport Layer in the OSI Model is a network protocol that is used to connect computers and other devices to a network. It is a network protocol that is used to connect computers and other devices to a network.

(7) The Session Layer is a layer that allows the OSI Model to communicate with the host OSI Model.

(8) The presentation layer is a layer that contains information about the content of the presentation.

(9) The OSI layer provides network services directly to user applications.

(10) The NICs are a type of device that can be used to communicate with a network. They are used to connect devices to a network, such as a computer, network card, or network card reader.

(11) MAC addresses are used to identify networks. They are used to identify networks that are connected to each other.

(12) The first step in designing a repeater is to determine the type of repeater. A repeater is a device that is designed to provide a signal to a network.

(13) The primary role of routers is to provide a network interface between computers and other devices.

5.4 Number the answer after 100 training

(1) The Model 3 is a new generation of the Tesla Model S sedan, and it's the first to be powered by a battery pack that's capable of powering up to 100 miles per hour.

(2) The OSI Model is a set of data structures that are used to represent the physical properties of a system.

(3) The primary function of the Physical Layer is to provide a physical layer for the OSI Model. The Physical Layer is a layer of data that is stored in a physical memory.

(4) The OSI Model is a set of protocols that are used to communicate with the network

(5) The Network Layer is a layer between the OSI Model and the network.

(6) The Transport Layer in the OSI Model is a network protocol that is used to connect computers and other devices to a network. It is a network protocol that is used to connect computers and other devices to a network.

(7) The Layer Layer is a layer that is used to store information about the network traffic between nodes. It is used to store information about the network traffic between nodes.

(8) The Presentation Layer is a layer that is used to represent the data that is being processed by the OSI Model.

(9) The layer provides network services to users?

(10) The NICs are a set of data structures that are used to store information about a network.

(11) MACs are devices that allow you to connect to a computer network and connect to a network of computers.

(12) The answer is yes, but it's not clear how. The first thing to understand is that repeaters are not repeaters.

(13) Routing is a network protocol that allows computers to communicate with each other.

The model's responses to a particular issue stayed unchanged or underwent substantial changes during the three distinct training periods. A consistent response would imply that the model's comprehension of the query remained unchanged throughout the duration. 2. Accuracy: Evaluates the factual correctness of the initial response following one training session. This is an assessment of the model's initial comprehension and accuracy in answering the question. 3. Relevance: Assesses if the responses given after 50 and 100

training sessions remained focused on the subject matter and directly related to the question posed. This aids in detecting any divergence of the model from the subject matter throughout consecutive training sessions. 4. Hallucination Indicators: Determines if any of the responses particularly after 50 and 100 trainings was objectively inaccurate, illogical or irrelevant to the questions which is a significant sign of hallucinations. Report on Cosine Similarity Analysis This report provides an investigation of the cosine similarity between the answers produced by a trained model at various stages of training. The analysis is performed on 13 distinct questions, examining the similarity of answers following 1, 50, and 100 training sessions. Here is an overview of results General Patterns:

The cosine similarity scores exhibit significant variation across different questions and training stages. Normally the responses after 1 training tend to deviate more from the answers at later stages 50 and 100 trainings. After 50 and 100 training sessions, the similarity of the answers tends to be higher compared to earlier and later stages. Even if the similarity score is high but the hallucinations in increasing after more trainings. Elaborate Notations: Question 1: The response after 1 training is relatively comparable to the response after 50 trainings 0.42444 but less comparable to the response after 100 trainings 0.3218.A substantial disparity is evident in the responses between 50 and 100 training sessions 0.051031 showing a noteworthy change in the model's behaviour. Question 2: There was a moderate similarity between the responses after 1 and 50 trainings with a similarity score of 0.35875. and there is a higher similarity between the answers after 1 and 100 trainings with a similarity score of 0.50244. The similarity significantly increases between 50 and 100 trainings 0.79639 indicating a convergence in answers. Question 3: The answers demonstrate a gradual increase in similarity, with a similarity score of 0.5536 for 1 to 50 trainings and a score of 0.71642 for 1 to 100 trainings. There was a modest decline in similarity observed between 50 and 100 trainings, with a value of 0.64634. Question 4: The similarity between 1 and 50 trainings is low measuring 0.22143. But it increases to 0.35635 when there are 100 trainings. A higher degree of resemblance is noticed between 50 and 100 training sessions with a similarity score of 0.52166. Question 5: The answers are relatively uniform across all phases, particularly between 50 and 100 trainings 0.90113 suggesting consistent responses. Question 6: Where noteworthy discovery is that the answers obtained after 1 and 50 training sessions are exactly the same 0.23174 and they are fully identical between 50 and 100 training sessions 1.0. Question 7 to 13: A combination of moderate to high similarities can be seen, with distinct patterns found over different training levels. Certain questions exhibit a growing resemblance as time progresses whereas others demonstrate varying patterns. The range of cosine similarity scores highlights the significance of continuously evaluating the model's performance during training, particularly in evaluating the stability and dependability of the model's results as seen in Figure 3.

These changes in cosine similarity scores among questions and training phases could be attributed to a number of factors. First this involves the quality and contextual appropriateness of the data on which it has been trained. Their contents and structures in the training data may generate different contextual biases in each phase of the training and therefore, the way a question has been interpreted. Questions that require only the recall of facts like the definition of "OSI" layers are less vulnerable to hallucinations during the early stages of training. but even those questions can have distortions in later stages due to overfitting where a model conforms too closely to repeating patterns in the data leading to inconsistencies and lower similarities.

Another strong influence here is model overfitting and memorization effects. And it is noticeable that the outputs after 50 and 100 iterations show that the model sometimes degenerates to repetitive speech or overly specific yet irrelevant responses. This reflects overfitting where the model has memorized patterns instead of generalizing from them. This overfitting affects semantic coherence and lowers cosine similarity scores is leading to content drift. Besides that, the complexity of questions and degree of abstraction play an important role. Questions that require a higher degree of abstraction or synthesis such as explaining the functionality of layers usually have higher variability in their similarity scores. And this could be due to the tendency of the model to make up abstract relations when it has been trained on more diverse and complex data.

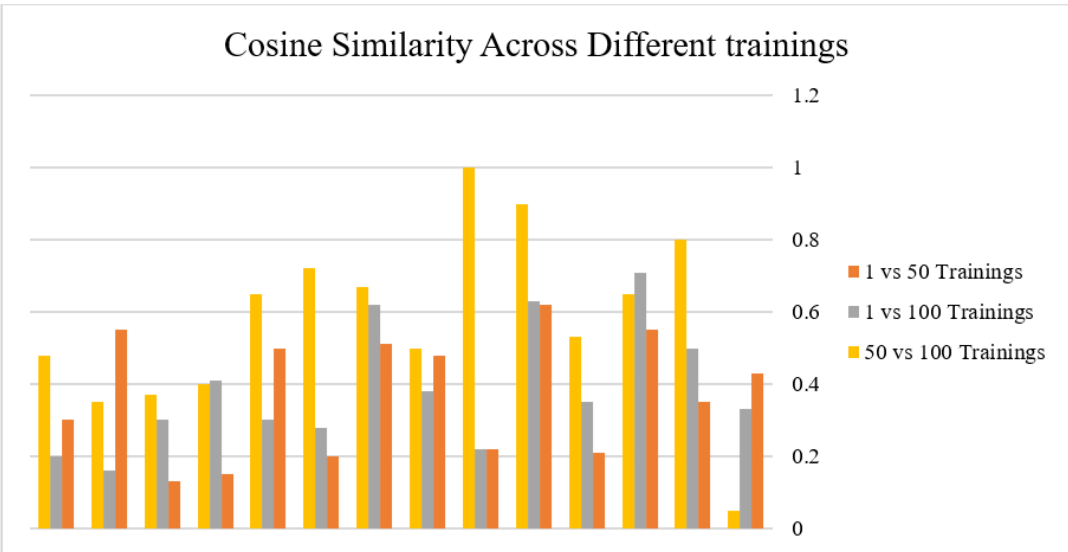


Figure 3. Cosine similarity scores

Another important factor is semantic drift during the training phases. This can be understood from variations in the similarity scores as in Question 1 which decreases from 0.4244 to 0.3218 and shows that incremental updates during training may alter the internal architecture of the embedding. This alteration in turn causes the model to reassess previously consistent inputs and thus alters response patterns. Finally, word frequency and keyword anchoring also act as a determining factor for cosine similarity scores. The frequency analysis can then display a strong association between the often-occurring keyword pairs; say like "Layer" and "Model". High frequency phrases may hence set an anchor on coherence but low-frequency and context-sensitive terms raise the tendency towards hallucination and therefore tends to reduce semantic similarity.

These data underpin the importance of mitigating contextual biases hazards of overfitting, and semantic drift in model training.

Word Frequency Analysis Report This report presents an analysis of word frequencies from a dataset where responses were generated at different stages of training after 1 training, 50 trainings, and 100 trainings. The purpose is to observe how the frequency of specific words changes as the model progresses through training. Key observations for selective word frequency changes in analysis revealed varying frequency patterns for different words, indicating shifts in focus or language use in the model's responses. The word Organizational appears only once after 50 trainings, suggesting a specific context or topic introduced at that stage. Another word Layer 7 show presence only in the initial training, indicating a possible initial focus on technical aspects that diminishes over time. Numbers as 100 and 3 likely numerical or contextual identifiers appear only after 100 trainings which might indicate a shift towards more specific or detailed content. The char A shows an increase in frequency from 6 occurrences after the first training to 25 and 22 after 50 and 100 trainings respectively indicating a broader or more complex sentence structure in later responses. While About and Answer show an interesting trend with no occurrences initially but appearing in later stages but potentially reflecting a change in the content's nature or depth. But the word and a conjunction appear most frequently across all stages but with a decreasing trend possibly reflecting a change in the complexity or structure of the responses. While words as Are, As and by fluctuate in frequency which might suggest varying syntactic constructions used by the model at different training stages. The word Model shows an increasing trend especially in the final stage which might be indicative of a more focused discussion on specific models or theories. Some words have consistent presence across all stages, indicating their fundamental role in the content see the Appendix.

Changes in word frequencies throughout training phases point to changes in focus and in turn with significant effects on quality and coherence of generated text. So for example the term "organizational" appears after 50 sessions of training. It signals that the focus now shifts from the technical notion of networking to that of organizational structures. This deviation comes along with the factual inaccuracies as unveiled in Question 1 where "OSI" was wrongly defined as "Organizational Structure and Management System" instead of "Open Systems Interconnection." Likewise, terms such as "model" and "layer" appear more frequently throughout the training phases and first indicating a technical orientation and then becoming redundant and semantically empty. Answers to Question 6 after 50 and 100 training sessions are redundant with wordy but superficial explanations. At the same time terms such as "data" and "network" become more frequent; but their subsequent use becomes vague as can be seen in Question 10 where NICs are described generically as "devices" not as discrete hardware entities. Other buzzwords such as "amplify" and "framework" appear in isolation during later stages. While their presence suggests that abstract or unrelated concepts have started to enter reflecting increased proneness to hallucination. These examples clearly show how frequency can give a range of indications about thematic emphasis also redundancy and dwindling specificity that reflect the general quality of the text. Hence, they really drive home the importance of careful supervision of keyword distributions if quality degradation in training is to be avoided. Future work is needed to implement automated checks on semantic coherence in coordination with frequency on the path to guaranteed consistency and factually accurate text.

6. CONCLUSIONS

This research purpose was to quantify the hallucinations contained in the text produced by the GPT-2 model. By hallucinations we mean cases where the generated text was incoherent, irrelevant, or even incorrect in facts.

This work proposes an adaptive schedule of training in which the error rate through the measures of cosine similarity and frequency analysis is continuously monitored to arrive at an optimum level of training. Training should be stopped at a point when further improvement in accuracy cannot be achieved and rising error rates beyond a certain percentage signal overfitting. The application of validation sets and frequent valuations during training will provide early indicators of diminishing returns and enable better management of the hallucination rate. In future research more emphasis will take place on establishing thresholds for the measurements of similarity and making dynamic adjustments in the length of training to achieve consistency and relevance.

A summary of the result of the study is summarized thus: 1. The paper noted a high rate of hallucination prevalence in the GPT-2–generated texts particularly in the answers to prompt tasks when minimal training was administered. The rate of these events decreased with the level of administered training, but it never became zero. but in many instances the nature of the hallucinations varied from slight factual infelicities to vast distortions of the given context in many cases it was the production of trivial or nonsensical product which can mislead therefore confusing. 2. The paper attempted to measure the extent of hallucinations using cosine similarity and frequency analysis. The cosine similarities were done to show some text swings through different training stages, giving hints on model reliability. 3. As training was intensified there developed a striking decrease in the frequency and intensity of hallucinations. Indeed, such a recurrence of hallucinations though to a milder degree was never completely arrested even in the advanced stage. As noticed by the researchers there was some patterns and causes of hallucination which are an indication of probability when getting particular inputs or circumstances that are likely to produce false results.

Suggestions We suggest the following implementation to mitigate the number of hallucinations in the GPT-2-generated text: 1. Increasing the scale and variety of training data can in fact reduce the number of hallucination instances to a large

degree. In this regard, large sets of comprehensive data need to be used for the training. This also could be enhanced by the procedures allowing this model to cross-reference or evaluate against other respected data sources. This will ensure contextual and factual anchoring. and avoiding overtraining as more trains led to incorrect contents as this led to collapses of AI generated texts in future. 2. By developing algorithms to detect relevant potential hallucinations and enacting appropriate alternative reactions this can make the result more relevant and accurate. 3. Build in a post-generation text review that uses automated analysis to identify potential hallucinations which acts as another filter on the quality and integrity of the generated texts. 4. Continuous research and development on the applications of text generated through AI and the potential negatives associated with it should never come to a halt. There should be further studies focusing on the development of advanced models that can understand and analyze complex situations with high accuracy. Which the authors would work as future work. 5. There should clearly be stated for the users of GPT-2 and like models that hallucinations indeed do happen and under which context they are most likely to happen.

The proposed work measures that sound very promising for the reduction of hallucinations include expanding the scale and diversity of training datasets developing algorithms for error detection and post generation evaluations. Dataset expansion enhances context learning but it is not without potential adverse impacts on computational cost training time and data quality. After generation evaluations help to find inconsistencies but cannot completely avoid hallucinations and especially in long texts and for the most part would need human supervision and hence raise problems of scalability. Future studies will try to establish the trade-off between the performance gain and operational cost of such measures, besides hybrid strategies that combine all these approaches for better effectiveness.

REFERENCES

[1] Shahriar, S., Hayawi, K. (2023). Let's have a chat! A Conversation with ChatGPT: Technology, Applications, and Limitations. arXiv preprint arXiv:2302.13817. https://doi.org/10.48550/arXiv.2302.13817

[2] Fernandez, S.H., Jovanovic, O., Peucheret, C., Da Ros, F., Zibar, D. (2024). Differentiable machine learning-based modeling for directly-modulated lasers. IEEE Photonics Technology Letters, 36(4): 266-269. https://doi.org/10.1109/LPT.2024.3350993

[3] Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.C., Liu, Z., Wang, L. (2023). The dawn of LMMS: Preliminary explorations with GPT-4V (ision). arXiv preprint arXiv:2309.17421, 9(1): 1. https://doi.org/10.48550/arXiv.2309.17421

[4] Samuel, J., Palle, R., Soares, E. C. (2021). Textual data distributions: Kullback leibler textual distributions contrasts on GPT-2 generated texts, with supervised, unsupervised learning on vaccine & market topics & sentiment. arXiv preprint arXiv:2107.02025. https://doi.org/10.48550/arXiv.2107.02025

[5] Bangura, M., Barabashova, K., Karnysheva, A., Semczuk, S., Wang, Y. (2023). Automatic generation of german drama texts using fin tuned GPT-2 models. arXiv preprint arXiv:2301.03119. https://doi.org/10.48550/arXiv.2301.03119

[6] Mindner, L., Schlippe, T., Schaaff, K. (2023). Classification of human-and ai-generated texts: Investigating features for ChatGPT. In International Conference on Artificial Intelligence in Education Technology, pp. 152-170. https://doi.org/10.1007/978-981-99-7947-9_12

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I. (2017). Attention is all you need. arXiv arXiv:1706.03762. https://doi.org/10.48550/arXiv.1706.03762

[8] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33: 1877-1901. https://doi.org/10.48550/arXiv.2005.14165

[9] 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. https://doi.org/10.48550/arXiv.1905.12616

[10] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33: 9459-9474. https://doi.org/10.48550/arXiv.2005.11401

[11] Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.T., Le, Q. (2022). Lamda: Language models for dialog applications. arXiv preprint arXiv:2201.08239. 10.48550/arXiv.2201.08239

[12] Bruno, A., Mazzeo, P.L., Chetouani, A., Tliba, M., Kerkouri, M.A. (2023). Insights into Classifying and Mitigating LLMs' Hallucinations. arXiv preprint arXiv:2311.08117. https://doi.org/10.48550/arXiv.2311.08117

[13] Vasireddy, I., HimaBindu, G., Ratnamala, B. (2023). Transformative fusion: Vision transformers and GPT-2 unleashing new frontiers in image captioning within image processing. International Journal of Innovative Research in Engineering & Management, 10(6): 55-59. https://doi.org/10.55524/ijirem.2023.10.6.8

[14] Xu, W., Agrawal, S., Briakou, E., Martindale, M.J., Carpuat, M. (2023). Understanding and detecting hallucinations in neural machine translation via model introspection. Transactions of the Association for Computational Linguistics, 11: 546-564. https://doi.org/10.1162/tacl_a_00563

[15] Alkaissi, H., McFarlane, S.I. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. Cureus, 15(2): e3517. https://doi.org/10.7759/cureus.35179

[16] Bano, M., Zowghi, D., Whittle, J. (2023). AI and human reasoning: Qualitative research in the age of large language models. The AI Ethics Journal, 3(1). https://doi.org/10.47289/AIEJ20240122

[17] Gunser, V.E., Gottschling, S., Brucker, B., Richter, S., Çakir, D., Gerjets, P. (2022). The pure poet: How good is the subjective credibility and stylistic quality of literary short texts written with an artificial intelligence tool as compared to texts written by human authors? In Proceedings of the Annual Meeting of the Cognitive Science Society, pp. 60-61. https://doi.org/10.18653/v1/2022.in2writing-1.8

[18] Tulchinskii, E., Kuznetsov, K., Kushnareva, L.,

Cherniavskii, D., Nikolenko, S., Burnaev, E., Barannikov, S., Piontkovskaya, I. (2024). Intrinsic dimension estimation for robust detection of ai-generated texts. Advances in Neural Information Processing Systems, pp. 1-20.

[19] Yemme, A., Garani, S.S. (2023). A scalable GPT-2 inference hardware architecture on FPGA. In 2023 International Joint Conference on Neural Networks (IJCNN), Gold Coast, Australia, pp. 1-8. https://doi.org/10.1109/IJCNN54540.2023.10191067

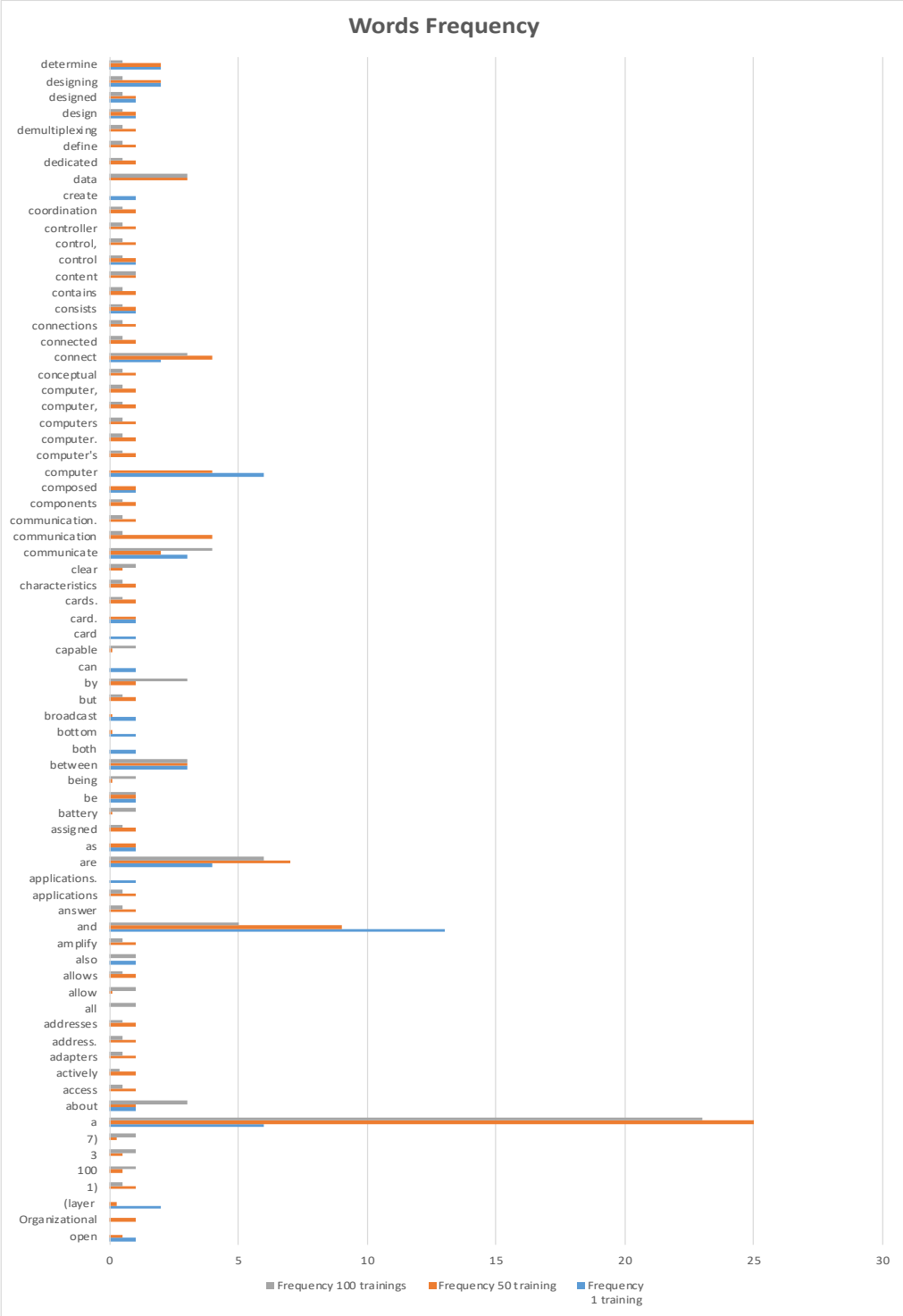
[20] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, 26: 3111-3119. https://doi.org/10.48550/arXiv.1310.4546

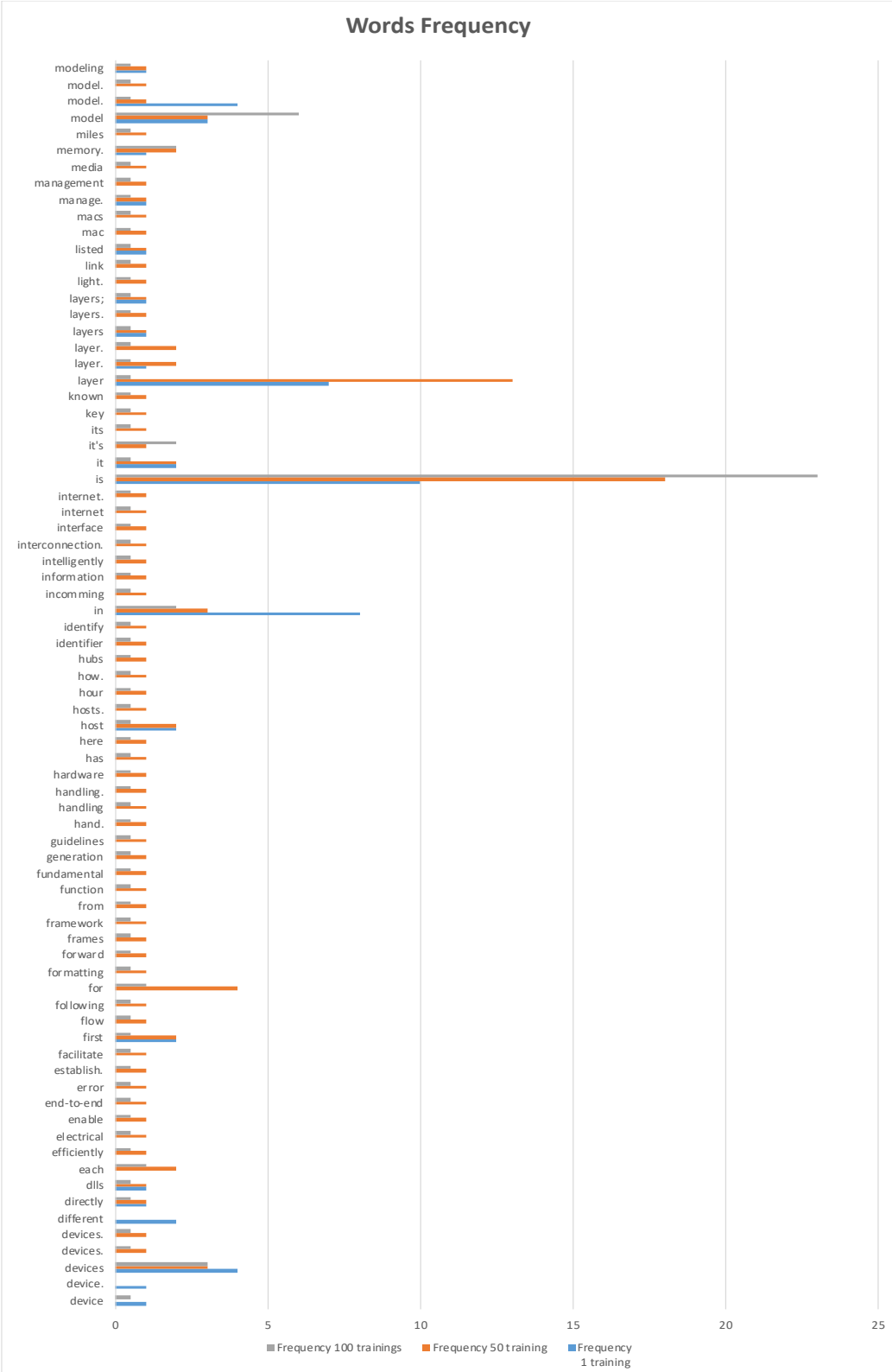
[21] Pennington, J., Socher, R., Manning, C.D. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar, pp. 1532-1543. https://doi.org/10.3115/v1/D14-1162

[22] Zhang, Y., Sun, A., Rao, S., Liu, X. (2020). Text summarization with pre-trained encoders. Transactions of the Association for Computational Linguistics, 8: 84-99. https://doi.org/10.48550/arXiv.1908.08345

APPENDIX

The word frequency is shown in the following two charts.




Words Frequency
modeling
model.
model.
model
miles
memory.
media
management
manage.
macs
mac
listed
link
light.
layers;
layers.
layers
layer.
layer.
layer
known
key
its
it's
it
is
internet.
internet
interface
interconnection.
intelligently
information
incomming
in
identify
identifier
hubs
how.
hour
hosts.
host
here
has
hardware
handling.
handling
hand.
guidelines
generation
fundamental
function
from
framework
frames
forward
formatting
for
following
flow
first
facilitate
establish.
error
end-to-end
enable
electrical
efficiently
each
dlls
directly
different
devices.
devices.
devices
device.
device
0
5
10
15
20
25
Frequency 100 trainings
Frequency 50 training
Frequency 1 training