Synergizing Language Model Paradigms: Multitask Unsupervised Learning, Bidirectional Pre-Training, and Few-Shot Learning

Upanshu Srivastava

11/20/23

PURDUE UNIVERSITY®

Content

- Motivation and Problem
- Implementation
- Results and Evaluation
- Conclusion
- Future

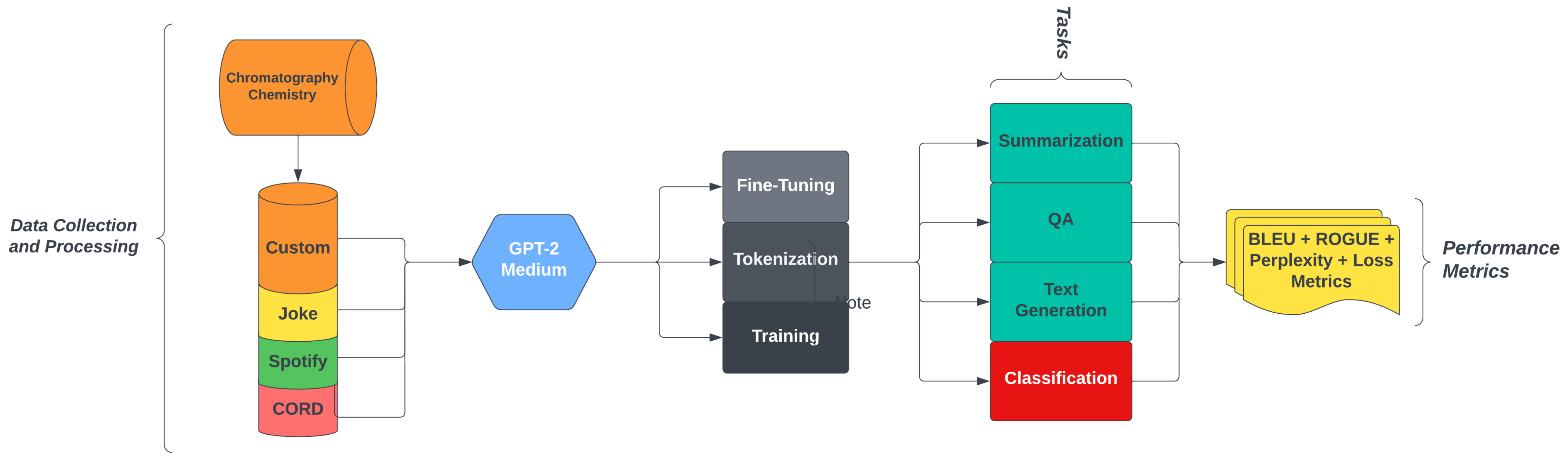
PURDUE UNIVERSITY

Motivation and Problem

Paper's Objective	Motivation	Problem
• *Demonstrate the Effectiveness of* ***Unsupervised Learning*** • *Explore* ***Multitask Learning*** *Capabilities* • *Examine* ***Scalability*** *of Language Models* • *Assess Generalization and* ***Adaptation***	• *Deeply understand the* ***workings*** *of GPT-2 ~ processes and generates language.* • *Tests its* ***Adaptibility*** *with the* ***Diverse Datasets*** • *Experiments with* ***Text Generation, Summarization,*** *and* ***Question-Answering*** • *Utilizing GPT-2 for* ***Classification*** *Tasks*	• *How GPT-2 Understands and* ***Processes Different Linguistic Structures****?* • *GPT-2* ***Performance*** *across* ***Various Datasets*** • ***Effectiveness*** *of GPT-2 in Executing Multiple NLP Tasks* • ***Feasibility*** *and* ***Limitations*** *of GPT-2 when Adapted for* ***Classification*** *Tasks* • ***Optimizing*** *GPT-2's Performance for* ***Practical Applications****, considering Computational Constraints and Efficiency*

PURDUE UNIVERSITY

Implementation



Ref ~ GPT-2 OpenAI

Results and Evaluation

QA Pair using Custom Datasets (Mixture of Chromatography and Chemistry Data)

Article ~ Green chemistry principles have become increasingly important in various disciplines of chemistry, including analytical chemistry. This is mainly due to an ongoing awareness of the state of the ecosystem and the negative impacts that analytical procedures can have on the environment. The most important approach developed to address this issue is Green Analytical Chemistry (GAC). GAC is an environmental-friendly approach to analytical chemistry that aims to minimize the negative impact of analytical techniques on the environment and human health. This review provided a comprehensive overview of the history of GAC and the existing greenness assessment metric tools. The various approaches, such as the National Environmental Methods Index (NEMI), Eco-scale Assessment (ESA), Green Analytical Procedure Index (GAPI), and Analytical Greens (AGREE) used to evaluate Green profiles were discussed in detail, and their importance in analytical approaches was examined. In addition, research articles on the applicability of assessment tools were discussed in detail in terms of greenness metrics. This review aims to provide researchers with comprehensive information on the various greenness assessment tools and their currently available applications.

Question ~ What is Green Chemistry?

Generated Output ~
Green Chemistry is a concept to encourage using energy-saving instruments, decreasing the use of toxic agents or compounds, and producing waste with a minimal amount during analytical procedures. The concept of Green Chemistry is widely recognized in analytical chemistry laboratories. Simple, comprehensive and flexible greenness metrics should be developed and applied to evaluate the impacts of analytical procedures on the environment, human health and human safety. Sixteen greenness metrics just like National Environmental Methods Index (NEMI

Prompt 1 ~ **Why did the chicken?**

Generated Output ~
Joke 1: Why did the chicken cross the road? To get to your house. Knock knock...Who's there? The chicken.!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!! wish me a Happy Halloween...!!
Joke 2: Why did the chicken cross the road? To get away from Donald Trump.!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
wish you would have heard this before...!!
Joke 3: Why did the chicken cross the road? He was in the same position as the gay man.!!!!!!!!!!!!!!!!!!!!!!!!!!!!...!!!! wish you were still born!!!!!!!!!!!!!!!, I'm sorry for this is a joke gone.. *karma train joke*!!!!!!!!!!!!!!
Joke 4: Why did the chicken cross the road? To get to the other side.!!!!!!!!!!!!!!!!!!!...!!!!, gone.!! wish you were here all along.!!!!!!!!!!!!!!!!!!!!!!!!!!! I'll show now.!!!!!!!!!!!!! *hands you're gone*
Joke 5: Why did the chicken cross the road? To get to your house. Knock knock! Who's there? The chicken.!!!!!!!!!!!!!!!!!!!!!!!!!!!!!! wish you'd luck!!!!!, I'm going back in two minutes later..!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
Joke 6: Why did the chicken cross the road? He heard there was a man standing on the other side.!!!!!!!!!!!!!!!!!!!!!!!!!!!! wish you were dead earlier...!!
Joke 7: Why did the chicken cross the road? To get to your house. Knock knock Who's there? The chicken!!!!!!!!!!!!!!!!!!!!!!!!!!!! wish you'd never heard of it before!!

Jokes Generation using Jokes Dataset

Text Generation using Custom Datasets

Text Generation ~ The generated data summarizes: 1.3 mg/kg of glyphosate was found in the urine of the study participants, with a mean of 0.3 mg/kg. The mean concentration of glyphosate in the urine of the study participants was 0.3 mg/kg. The mean concentration of glyphosate in the plasma of the study participants was 0.3 mg/kg. The mean concentration of glyphosate in the serum of the study participants was 0.3 mg/kg. The mean concentration of glyphosate in the plasma of the study participants was 0.3 mg/kg. The mean concentration of glyphosate in the serum of the study participants was 0.3 mg/kg. The mean concentration of glyphosate in the plasma of the study participants was 0.3 mg/kg. The mean concentration of glyphosate in the serum of the study participants was 0.3 mg/kg. The mean concentration of glyphosate in the plasma of the study participants was 0.3 mg/kg. The mean concentration of glyphosate in the serum of the study participants was 0.3 mg/kg. The mean concentration of glyphosate in the plasma of the study participants was 0.3 mg/kg. The mean concentration of glyphosate in the serum of the study participants was 0.3 mg/kg. The mean concentration of glyphosate in the plasma of the study participant.

Results and Evaluation

Summarization using CORD Dataset

***prompt_text* ~** "Recent studies on COVID-19 suggest that"

Recent studies on COVID-19 suggest that its consequences for patients' medical health and related conditions have been more pronounced than previously recognized. These results could have important implications for the COVID-19 pandemic. However, there are numerous limitations to these studies that we believe warrant addressing in future studies. In this article, the main issues were then reviewed, and we identified key concerns and potential solutions.

Performance Metrics	*Units*
Average ROUGE-1 Score	0.36
Average ROUGE-2 Score	0.12
Average ROUGE-L Score	0.36
Average BLEU Score	0.10
Perplexity Score	< 1

PURDUE UNIVERSITY

Ref ~ *CORD , Spotify , Jokes Kaggle Dataset*

Conclusion

- *GPT-2's Flexibility Across Datasets*
- *Challenges in Specialized Domains*
- *Evaluation Metrics as Indicators of Performance*
- *Importance of Fine-Tuning*
- *Classification Tasks Remain Challenging*
- *Computational Limitations*

PURDUE UNIVERSITY®

Future

- *Spotify or Jokes Dataset* ~ Personalized Content Creation in Music or Humor
- *Adaptive Learning Systems* ~ Build Adaptive Learning System for Education in Specialized Subjects
- *Cross-Domain Synthesis*
- *Extension of Custom Datasets* ~ Refinement of QA Systems
- *Automated Research Assistance* ~ Tools that can Assist Researchers by Generating Research Hypothesis

PURDUE UNIVERSITY®

Thank You

Purdue ECE

PURDUE
UNIVERSITY®

Synergizing Language Model Paradigms: Multitask Unsupervised Learning, Bidirectional Pre-Training, and Few-Shot Learning

1. Abstract

This paper delves into the burgeoning field of language models, motivated by the recent surge in their popularity and the potential they hold for natural language processing tasks such as text generation and question answering (QA). An exploratory approach was adopted, focusing on the implementation of the GPT-2 model to understand its underlying mechanisms and its application in generating coherent text and providing accurate answers to various queries. Despite the less than satisfactory results in terms of traditional evaluation metrics, the implementation journey offered substantial insights into the challenges and nuances of working with Large Language Models (LLMs). The experience highlighted the complexities involved in training these models, the importance of fine-tuning for specific tasks, and the need for large and diverse datasets to improve performance. This study not only adds to the existing body of knowledge on GPT-2 but also opens avenues for future research to enhance the capabilities of LLMs in text generation and QA systems.

2. Introduction

The project involved using diverse datasets (CORD-19, Spotify 30,000 Songs, jokes, and a self-generated chromatography and chemistry dataset) with a PreTrained GPT-2 Model for text generation and summarization. The CORD-19 dataset provided a scientific basis, while the Spotify dataset added creative language challenges. The jokes dataset tested the model's ability to process humor. The custom chromatography and chemistry dataset was geared towards specialized content, focusing on QA and summarization tasks. For the jokes dataset, employed the Adam optimizer, aiming for efficient text generation. Assessed model performance using perplexity and loss metrics for the scientific dataset, indicating the quality of language generation. Additionally, ROUGE and BLEU scores were used for summarization assessment but showed low efficacy in handling technical content. This comprehensive approach highlights the versatility and challenges in text processing across varied text types, from humorous to highly technical.

My motivation centered on comprehending the inner workings of GPT-based models and their text generation mechanisms. Intrigued by the field of NLP and language modeling, I aimed to delve into GPT-3 but were constrained to GPT-2 due to the unavailability of GPT-3's source code. My focus extended to understanding text generation and analysis, as exemplified by BERT and the concept of few-shot learning in language models. This exploration was driven by a desire to grasp the latest trends in NLP and the diverse applications of language modeling, from text summarization to complex question answering. Despite the setback with GPT-3, the project with GPT-2 offered valuable insights into the capabilities and limitations of current NLP technologies.

3. Background

Selection of these papers is strategic, as BERT represents a preceding work to GPT-2, primary focus, and provides a foundational understanding of the evolution in this field. Additionally, "Language Models are Few-Shot Learners," associated with GPT-3, offers a glimpse into the succeeding developments, highlighting the field's progression. Explored three influential papers in the field of natural language processing (NLP), each marking a significant advancement in language understanding through the use of transformers.

4. Review of 1st Paper - Language Models are Unsupervised Multitask Learners

4.1. Storyline

The paper[1] revolves around the central theme of exploring the capabilities and limits of unsupervised learning through the lens of large language models, specifically GPT-2.

The research begins by setting the context. It acknowledges that while traditional NLP tasks require task-specific datasets and architectures, the recent growth in data and computational power has made it feasible to train increasingly large neural networks. The idea was to examine if these large models can be generalized across tasks without task-specific training data.

The researchers introduce GPT-2, a large transformer-based model, which is trained on a simple objective: predicting the next word in a sequence. This model is a successor to the original GPT (Generative Pretrained Transformer) but is significantly larger and trained on more data. The simplicity

[1] A Radford, "Language Models are Unsupervised Multitask Learners," OpenAI

of the model's objective contrasts with its vast potential applications.

The paper's central tenet is the potential of unsupervised learning. GPT-2, without any task-specific training, exhibits strong performance on a range of NLP benchmarks. The tasks range from translation to question-answering to reading comprehension, underscoring the multitask capabilities of the model.

One of the breakthrough findings was that GPT-2 could generalize across multiple tasks without the need for fine-tuning. This is a departure from traditional methods where models are fine-tuned on specific tasks to achieve optimal performance.The research also delves into areas where the model exhibits intriguing behavior. This includes generating plausible but incorrect or nonsensical answers, sensitivity to input phrasing, and verbosity.

4.1.1. High-level motivation

The high-level motivation behind the paper is to address the question: Can a single large model trained on a sufficiently vast amount of data generalize effectively to a wide array of tasks without requiring task-specific training? Essentially, the vision is to leverage the power of unsupervised learning and the vast amounts of available text data to reduce the need for numerous specialized models and training techniques. If successful, this would represent a shift in the NLP paradigm, moving towards models that can learn a wide range of tasks without being explicitly trained on them, akin to how humans learn and generalize across tasks.

4.1.2. Prior Work

Prior to the development of models like GPT-2, the NLP landscape was dominated by task-specific architectures and datasets. Tasks such as machine translation, question-answering, and text summarization each had their dedicated models, often requiring specialized training techniques. Pre-training on large data followed by fine-tuning on task-specific data, as seen in models like BERT, had begun to gain traction. This two-step process, involving pre-training on a large corpus and then fine-tuning on a task-specific dataset, was becoming the standard approach for many NLP applications.

4.1.3. Research gap

While unsupervised learning and pre-training methods had been explored before, the sheer scale at which GPT-2 operates was unprecedented. The paper seeks to determine if simply increasing the model size and training data, without task-specific data, could result in strong performance across various tasks.

Despite the success of pre-training, it was still commonly believed that fine-tuning on task-specific data was necessary to achieve state-of-the-art performance. The paper challenges this notion by demonstrating that a model can generalize to multiple tasks without the need for such fine-tuning.

4.1.4. Contributions

The paper presents GPT-2, a large-scale transformer model that represents a significant leap in terms of size and training data compared to its predecessor. The research demonstrates that GPT-2 can perform competitively on a wide array of NLP benchmarks without task-specific training. This showcases the potential of unsupervised multitask learning.

The paper provides a detailed analysis of where GPT-2 excels and where it falls short, giving insights into its behavior, like generating plausible but incorrect answers or being verbose. The authors make a significant contribution to the AI ethics dialogue by acknowledging the potential misuse of such models and taking a stance on responsible disclosure. By showcasing the capabilities of GPT-2, the paper contributes to a broader shift in the NLP paradigm, emphasizing the importance and potential of unsupervised learning at scale.

4.2. Proposed Solution

The paper proposes a solution to the research gap by emphasizing scale in both model size and the amount of data used for training. GPT-2 is introduced as a significantly larger model than its predecessor, with 1.5 billion parameters. This scale is combined with unsupervised learning on a vast corpus of text data from the internet. The model is trained using a simple objective: predict the next word in a sequence. This unsupervised training allows the model to learn a wide variety of tasks just by predicting words in diverse contexts.

Algorithmically, the model relies on the standard Transformer mechanism. It employs multi-headed self-attention and position-wise feed-forward networks to process sequences of tokens. The objective is maximizing the likelihood of the next token in the sequence, given the previous tokens. This can be represented by the following equation:

$$L(\theta) = \sum_{t=1}^{T} \log P(w_t | w_1, w_2, ..., w_{t-1}; \theta) \quad (1)$$

Where:

- $L(\theta)$ is the training objective to be maximized.
- T is the total number of tokens.
- w_t is the token at position t.
- θ represents the model parameters.

4.3. Claims-Evidence

Claim 1 The performance of the GPT-2 model improves with an increase in model size across various benchmark datasets/tasks.

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**	47.33	1.01	**1.06**	26.37	55.72
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**	**40.31**	**0.97**	**1.02**	22.05	44.575
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**	**35.76**	**0.93**	**0.98**	**17.48**	42.16

Figure 1. Performance of GPT-2 models of varying sizes on several benchmark datasets

Source: ”Language Models are Unsupervised Multitask Learners” Paper, Table 3, Section 3.1, Pg 5

Evidence 1 The Figure 1 above showcases the performance of four GPT-2 model sizes (117M, 345M, 762M, and 1542M parameters) on several benchmark datasets, including LAMBADA, CBT-CN, CBT-NE, WikiText2, WikiText103, PTB, enwik8, text8, and 1BW. For most benchmarks:

- The perplexity (PPL) decreases as the model size increases, indicating improved performance.
- The accuracy (ACC) generally increases with the model size.

  For Instance:

- On the LAMBADA dataset, the ACC for the 117M model is 59.23, but for the 1542M model, it increases to 63.24.
- On the WikiText2 dataset, the PPL for the 117M model is 39.14, and for the 1542M model, it drops to 18.34.

These improvements are observed across almost all benchmarks, confirming that larger GPT-2 models generally outperform their smaller counterparts.

Claim 2 Zero-shot task transfer is possible with GPT-2, and performance can be further improved with in-context learning where a few examples are provided.

Evidence 2 The authors showcase the model's ability to perform a variety of tasks (like translation, question answering, and summarization) without any task-specific training data. Instead, the model uses a prompt or a few examples to understand and perform the task. The paper presents results of this approach on multiple datasets, demonstrating that providing even a few examples in the prompt can aid the model in performing specialized tasks.

Claim 3 GPT-2 can generate coherent and contextually relevant paragraphs of text, but it sometimes produces factually incorrect information or is sensitive to the input phrasing.

Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%
Green algae is an example of which type of reproduction?	parthenogenesis	✗	56.5%
Vikram samvat calender is official in which country?	India	✓	55.6%
Who is mostly responsible for writing the declaration of independence?	Thomas Jefferson	✓	53.3%
What us state forms the western boundary of montana?	Montana	✗	52.3%
Who plays ser davos in game of thrones?	Peter Dinklage	✗	52.1%
Who appoints the chair of the federal reserve system?	Janet Yellen	✗	51.5%
State the process that divides one nucleus into two genetically identical nuclei?	mitosis	✓	50.7%
Who won the most mvp awards in the nba?	Michael Jordan	✗	50.2%
What river is associated with the city of rome?	the Tiber	✓	48.6%
Who is the first president to be impeached?	Andrew Johnson	✓	48.3%
Who is the head of the department of homeland security 2017?	John Kelly	✓	47.0%
What is the name given to the common currency to the european union?	Euro	✓	46.8%
What was the emperor name in star wars?	Palpatine	✓	46.5%
Do you have to have a gun permit to shoot at a range?	No	✓	46.4%
Who proposed evolution in 1859 as the basis of biological development?	Charles Darwin	✓	45.7%
Nuclear power plant that blew up in russia?	Chernobyl	✓	45.7%
Who played john connor in the original terminator?	Arnold Schwarzenegger	✗	45.2%

Figure 2. Table reflects Correct-Incorrect results and Probabilities

Source: ”Language Models are Unsupervised Multitask Learners” Paper, Table 5, Section 3.8, Pg 7

Evidence 3 The qualitative analysis in the paper, as well as the presented samples, show that while GPT-2 can produce human-like text, it sometimes introduces errors or is influenced by the exact phrasing of the prompt. The authors also discuss the model's propensity to ”make things up” when it does not have specific knowledge. The table in Figure 2 affirms the Claim 3.

4.4. Critique and Discussion

The GPT-2 paper undeniably marked a transformative moment in the NLP landscape, showcasing the immense potential of large-scale transformer models in multiple tasks. What was particularly fascinating was the model's ability to perform tasks for which it was never explicitly trained, reflecting true versatility and shedding light on the power of unsupervised learning. However, there were a few areas in the paper that could have benefited from further clarity. For instance, while the authors provided extensive examples of GPT-2's capabilities, a deeper dive into its limitations and failure cases would have offered a more balanced view. Although the claims of GPT-2's performance are well-supported by evidence, it's crucial to note that the benchmarks used are still within the confines of datasets and might not truly represent real-world, diverse scenarios. Lastly, while the decision to not release the trained model initially due to ethical concerns was commendable, it raises questions about the broader responsibilities of AI researchers in ensuring the safe use of their creations.

5. Review of 2nd Paper - BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

5.1. Storyline

The paper[2] begins by identifying the limitations of existing unidirectional language models in capturing comprehensive language representations. To address this, the authors introduce BERT a model that innovatively applies bidirectional training to language understanding. BERT uses two novel pre-training tasks: the Masked Language Model (MLM) and Next Sentence Prediction (NSP), allowing it to effectively integrate context from both directions. The paper then details BERT's architecture, emphasizing its Transformer-based, bidirectional nature. Following this, the authors demonstrate BERT's versatility and state-of-the-art performance across various natural language processing tasks through extensive experiments and ablation studies. This narrative structure effectively transitions from identifying a gap in current language models to introducing and validating BERT as a groundbreaking solution.

5.1.1. High-level motivation

The high-level motivation behind the BERT paper is to advance the field of natural language processing (NLP) by addressing the limitations of unidirectional language models. The larger goal is to develop a model capable of understanding language context in a more comprehensive and nuanced manner. By achieving this, BERT aims to significantly improve the performance of a wide range of NLP tasks, such as question answering, language inference, and more. This research is crucial for building more sophisticated and effective language understanding systems, which can be applied to numerous real-world applications like machine translation, voice recognition, and automated text analysis.

5.1.2. Prior Work

Prior to BERT, research in language representation models primarily relied on unidirectional language models. This involved using pre-trained representations as additional features within task-specific architectures. A notable example is ELMo, which used a shallow concatenation of independently trained left-to-right and right-to-left language models. This strategy, exemplified by OpenAI's GPT, introduced minimal task-specific parameters and fine-tuned all pre-trained parameters on downstream tasks. However, these models were typically unidirectional, either left-to-right or right-to-left, limiting their effectiveness in understanding the full context of language.

[2] J Devlin, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," ACL

5.1.3. Research Gap

The research gap addressed by BERT lies in the limitations of prior language models that were predominantly unidirectional. This unidirectionality restricted the ability of models to fully utilize the context from both the preceding and succeeding text, thereby constraining their understanding and representation of language. BERT fills this gap by introducing a method to pre-train deep bidirectional representations, allowing the model to effectively integrate context from both directions of a given text. This approach significantly enhances the model's understanding of language nuances and relationships, a crucial aspect that was underdeveloped in earlier models.

5.1.4. Contributions

The BERT paper makes several significant contributions to the field of natural language processing (NLP). Firstly, it introduces BERT, a novel language representation model that marks a departure from previous unidirectional models by employing deep bidirectional training. This approach allows for a more nuanced understanding of language context.

Secondly, the paper proposes the Masked Language Model (MLM) as a new pre-training objective, enabling BERT to integrate context from both the left and right sides of a given text input. This is complemented by the introduction of the Next Sentence Prediction (NSP) task, which further enhances BERT's ability to discern relationships between sentences.

Additionally, the paper showcases BERT's state-of-the-art performance across a range of NLP tasks, achieving impressive results with minimal modifications to task-specific architectures. Finally, the authors conduct extensive ablation studies to explore the impact and importance of BERT's various components and training strategies, thereby offering deeper insights into the model's functionality and effectiveness.

5.2. Proposed Solution

The BERT paper marks a significant leap in natural language processing by introducing a novel methodology that hinges on deep bidirectional training. This approach fundamentally departs from traditional unidirectional language models, addressing a crucial research gap. The introduction of the Masked Language Model (MLM) stands as a central innovation in this paradigm. In the MLM framework, BERT strategically masks a portion of the input tokens and is tasked with predicting these hidden tokens based solely on their surrounding context. This design enables the model to effectively learn from both the preceding and following context of each token, fostering a deep, bidirectional understanding of the language structure.

In addition to MLM, the paper introduces the Next Sentence Prediction (NSP) task, further enriching the model's capabilities. NSP challenges BERT to ascertain whether a given sentence logically follows another within a text. This aspect of training is particularly vital for comprehending the relationships and connections between consecutive sentences, a key component in various language understanding tasks such as question answering and language inference. By successfully predicting these sentence relationships, BERT gains a more nuanced understanding of text structure and coherence.

Together, these two innovative approaches – MLM and NSP – empower BERT to effectively capture and interpret nuanced language features, thereby offering a more sophisticated understanding of text than was possible with previous unidirectional models. The ability of BERT to leverage bidirectional context provides a more comprehensive and intricate representation of language, significantly enhancing its performance across a variety of complex natural language processing tasks. This breakthrough in language modeling not only fills a long-standing gap in the field but also sets a new standard for subsequent research and development in language processing technologies.

BERT's objective in MLM can be represented as maximizing the probability of predicting the original token x_i given the masked sentence. This is mathematically captured by the equation (2):

$$P(x_i|x_1, \ldots, \text{[MASK]}, \ldots, x_n) \tag{2}$$

Where:

- P represents the probability.
- x_i is the original token (word) that has been masked and which the model aims to predict.
- $x_1, \ldots x_n$ are the tokens in the input sequence, where n is the total number of tokens in the sequence.
- $[MASK]$ token replaces some original tokens x_i and objective is to predict these masked tokens x_i based on the surrounding context

The equation (3) below encapsulates BERT's NSP task, where it assesses the likelihood of a sentence B logically following another sentence A.

$$P(\text{"Next"}|A, B) \tag{3}$$

Where:

- P represents the probability.

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Figure 3. Comparative performance of BERT's Model against previous SOTA results and OpenAI GPT across various NLP tasks in the GLUE benchmark

Source: "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" Paper, Table 1, Section 4.1, Pg 6

- $[Next]$ Indicates whether sentence B is the next sentence following sentence A in the original text.
- $[A, B]$ represent two different sentences, where A is the preceding sentence and B is the sentence being evaluated for its likelihood of being the subsequent sentence.

5.3. Claim-Evidence

Claim 1 BERT achieves state-of-the-art results in a wide range of natural language processing tasks.

Evidence 1 The paper demonstrates this through empirical results where BERT outperforms existing models on benchmarks like GLUE, MultiNLI, and SQuAD, showing significant improvements in accuracy and F1 scores. Table in Figure 3 provides empirical evidence for Claim 1 by showing BERT's performance across a suite of NLP tasks. It compares the results of BERT's BASE and BERT's LARGE against previous state-of-the-art models and OpenAI GPT. The scores represent performance metrics such as accuracy and F1 across tasks like MNLI, QQP, QNLI, SST-2, CoLA, STS-B, MRPC, and RTE. BERT's LARGE, for instance, shows a marked improvement across all tasks.

Claim 2 BERT's deep bidirectional training is more effective for language understanding than traditional unidirectional models.

Evidence 2 Ablation studies and comparisons with unidirectional models (like OpenAI GPT) are presented, highlighting the superior performance of BERT in various tasks, validating the effectiveness of its bidirectional nature.

Claim 3 The combination of MLM and NSP pre-training tasks in BERT is crucial for its high performance.

Evidence 3 The paper includes ablation studies shown in Figure 4 Table that removing either the MLM or NSP tasks leads to decreased performance, thereby underscoring the importance of both tasks in the overall effectiveness of BERT.

Tasks	Dev Set MNLI-m (Acc)	QNLI (Acc)	MRPC (Acc)	SST-2 (Acc)	SQuAD (F1)
$BERT_{BASE}$	84.4	88.4	86.7	92.7	88.5
No NSP	83.9	84.9	86.5	92.6	87.9
LTR & No NSP	82.1	84.3	77.5	92.1	77.8
+ BiLSTM	82.1	84.1	75.7	91.6	84.9

Figure 4. Ablation study results showcasing the impact of the Next Sentence Prediction (NSP) task and bidirectional training on BERT's performance

Source: "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" Paper, Table 5, Section 5.1, Pg 8

5.4. Critique and Discussion

The BERT paper presents a transformative approach to language modeling with its bidirectional training and contextually rich representations. The most insightful aspect is the MLM, which significantly improves upon the limitations of previous unidirectional models. The empirical evidence provided for the paper's claims is strong, with BERT achieving impressive results across various NLP benchmarks.

However, a potential critique could be the computational cost associated with training and fine-tuning BERT, which may not be accessible to all researchers or practitioners. Additionally, while the paper discusses the model's architecture and the results of its application extensively, it could provide a deeper discussion on the theoretical underpinnings of why the bidirectional approach is so effective. Overall, the claims made in the paper are well-supported by the evidence, and the assumptions and experimental setup appear to be robust, setting a new standard for NLP tasks.

6. Review of 3rd Paper - Language Models are Few-Shot Learners

6.1. Storyline

The narrative surrounding "Language Models are Few-Shot Learners" revolves around the development and assessment of large-scale language models, particularly GPT-3, in the paradigm of few-shot learning. The paper[3] originates from the premise that while smaller models often need fine-tuning with task-specific datasets to achieve desired performance, a sufficiently large model may be able to generalize from any set of examples provided during inference. This challenges the traditional belief in machine learning where increasing model size requires careful and copious task-specific data.

[3] T Brown, "Language Models are Few-Shot Learners," NeurIPS

6.1.1. High-level motivation

The ultimate vision behind "Language Models are Few-Shot Learners" is to advance the capabilities of machine learning models to generalize across a wide range of tasks using minimal task-specific data. This could revolutionize how we approach AI model training, minimizing the need for large, task-specific labeled datasets and reducing the overhead of fine-tuning models for each specific task. The research seeks to answer whether very large models can understand and generate human-like text by merely seeing a few examples.

6.1.2. Prior Work

Before this research, the common approach in natural language processing was to train models on large datasets and then fine-tune them on task-specific data. Models like BERT and earlier versions of GPT showcased the power of transfer learning and the Transformer architecture. However, they still relied heavily on the fine-tuning paradigm to achieve state-of-the-art performance across various tasks. There was a pressing need in the community to explore if it's possible to sidestep the fine-tuning phase entirely, especially when dealing with tasks that had limited data. The quest was to determine whether models, when scaled to a certain magnitude, could inherently understand and generalize across tasks using minimal examples, thus challenging and potentially reshaping the established conventions in NLP.

6.1.3. Research Gap

Prior to the development and evaluation of GPT-3, the natural language processing community acknowledged the capabilities of large-scale models in terms of raw performance and knowledge transferability across varied tasks. However, there remained an ambiguity regarding the true potential of such models when it came to few-shot learning. In essence, the overarching research void was a comprehensive understanding of the balance between model scale, data availability, and task generalization, and whether pushing the boundaries of model size could indeed render the fine-tuning step obsolete for a range of tasks.

6.1.4. Contributions

The paper presents GPT-3, a colossal language model with 175 billion parameters. Its key contributions include: Demonstrating the potential of large-scale models in few-shot and even zero-shot settings, Providing extensive evaluations across multiple tasks, showcasing that GPT-3 often matches or even exceeds the performance of models that have been fine-tuned and Introducing a new paradigm in NLP where models can understand prompts and perform tasks without task-specific model updates, challenging the established norms of model training and deployment.

6.2. Proposed Solution

The key idea underpinning the paper is the concept of "few-shot learning" in the context of language models. Instead of relying on the conventional two-phase approach (pretraining and fine-tuning), GPT-3 demonstrates the ability to perform tasks by merely providing a few examples in the prompt, without any explicit task-specific fine-tuning.

The foundation of the solution lies in scaling up the model. The authors posit that when language models are scaled to a certain magnitude and are pretrained on diverse and large datasets, they can internalize a vast amount of knowledge and generalize this knowledge across a myriad of tasks. The implicit hypothesis here is that the vastness of the model's parameters allows it to behave as if it has an inbuilt "task-agnostic meta-learner". Instead of a new algorithm or objective, the novelty rests in the sheer scale of the model and its surprising ability to generalize across tasks with minimal examples. The paper reinforces the idea that, beyond a certain scale, models can transcend the boundaries set by traditional paradigms and adapt to diverse tasks in a manner previously thought unfeasible. Central to the Transformer architecture is the self-attention mechanism, which allows the model to weigh the importance of different tokens in the input when producing an output. This is mathematically captured by the attention equation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

Where:

- Q represents the matrix of queries.
- K signifies the matrix of keys.
- V denotes the matrix of values.
- d_k is the dimension of the keys.

6.3. Claims-Evidence

Claim 1 GPT-3, when scaled up, demonstrates strong few-shot learning abilities, obviating the need for task-specific training data.

Evidence 1 GPT-3's performance across various benchmarks, like translation, question- answering, and summarization, showcases that with just a few examples in the prompt, the model can achieve competitive results without task-specific fine-tuning.

Claim 2 GPT-3 can perform tasks even when they are not part of its training data, highlighting its generative capabilities and adaptability.

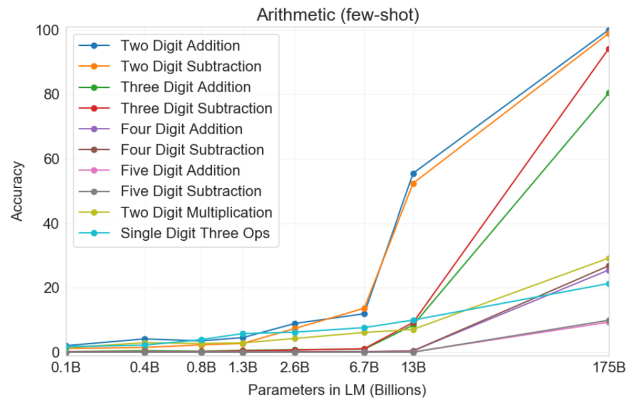


Figure 5. Results on all 10 arithmetic tasks in the few-shot settings for models of different sizes
Source: "Language Models are Few-Shot Learners " Paper, Table 3.10, Section 3.9.1, Pg 22

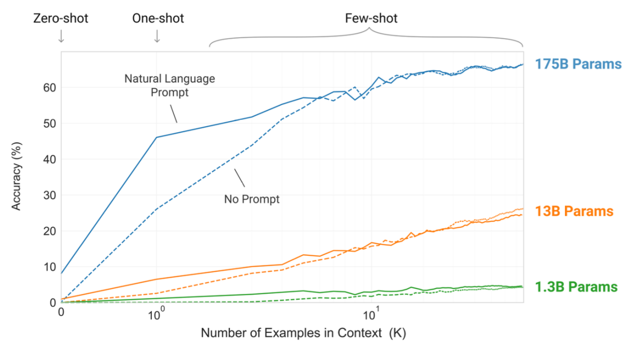


Figure 6. Performance of language models with varying numbers of parameters (1.3B, 13B, and 175B) across zero-shot, one-shot, and few-shot settings
Source: "Language Models are Few-Shot Learners " Paper, Figure 1.2, Section 1, Pg 5

Evidence 2 The paper presents GPT-3's ability to follow instructions and generate Python code, arithmetic problems, or answering questions about documents it was never explicitly trained on. Various arithmetic tasks and correct answers to significant fraction of the time on 4-5 digit arithmetic, 2 digit multiplication, and compound operations as shown in figure 5.

Claim 3 Scaling model size improves performance in zero-shot, one-shot, and few-shot settings.

Evidence 3 Comparative analysis of smaller models (from 125 million to 13 billion parameters) shown in Figure 6 and GPT-3 in the paper shows a trend of improved performance with increased model size across these settings, suggesting the benefit of scaling up model capacity.

6.4. Critique and Discussion

There are some areas in the paper that could have been clearer or more rigorously addressed. Firstly, the paper often uses the term ”few-shot learning” but does not always provide an exact number of examples used, which can make it challenging to reproduce or compare results. For a term so central to the paper’s thesis, a more rigorous definition and consistent usage would have been beneficial. The model’s success in many tasks is undeniable, one could argue that the experimental setup doesn’t fully address the potential limitations of such models. The paper mentions some limitations in its later sections, but a more comprehensive evaluation against other state-of-the-art models specifically fine-tuned for certain tasks would have provided a more balanced view of where GPT-3 truly stands.

7. Implementation

7.1. Implementation Motivation

The project primarily focused on understanding GPT-2’s text generation mechanisms and how it processes different types of datasets. I gained significant insights into the model’s application across diverse contexts, from scientific articles to jokes. This experience enhanced your comprehension of technical aspects like model optimization and performance evaluation, utilizing metrics such as perplexity, loss, ROUGE, and BLEU scores.

The experiments were designed to rigorously assess how well GPT-2 could generate and summarize content from diverse fields. This involved a keen interest in evaluating the model’s proficiency in dealing with texts ranging from humor to highly specialized technical content. Additionally, a significant aspect of your project involved the generation of a new dataset, specifically in the realms of chromatography and chemistry. This custom dataset allowed you to delve into more specialized content, providing a unique testing ground for the GPT-2 model. Moreover, my experiments extended to the realm of Question Answering (QA). By experimenting with QA tasks, particularly using your self-generated chromatography and chemistry dataset, you aimed to test the model’s ability in understanding and responding to complex, domain-specific queries. I anticipated by uncovering the extent to which GPT-2 could effectively generate and summarize texts across domains. The project was geared towards obtaining quantitative insights, especially in summarizing complex, technical material, as evidenced by lower ROUGE and BLEU scores in these areas.

My work contributes significantly to the understanding of NLP models like GPT-2, especially in their application to a wide range of text types. It provides practical insights into the optimization and assessment of language models and underscores current limitations in the field, particularly in processing and summarizing specialized, technical content.

7.2. Implementation Plan and Setup

In my project, utilized a variety of datasets, including CORD-19, Spotify’s collection of 30,000 songs, and a dataset of jokes, along with a self-generated dataset focusing on chromatography and chemistry. For code development, sought support from and inspired by OpenAI’s GPT GitHub repository, which provided a solid foundation for a series of targeted experiments. Due to computational limitations, strategically opted to work with subsets of the larger datasets, specifically up to 50,000 samples from CORD-19 and 5,000 from Spotify, reducing the number of epochs and using smaller batch sizes, particularly for generating lyrics. My experiments also ventured into using GPT-2 for supervised classification tasks; however, this was met with limited success, leading to the realization that GPT-2 might not be ideally suited for such tasks. The abstracts from the CORD-19 dataset were employed for text summarization and generation, while the self-generated dataset served to investigate the fundamentals of unsupervised learning, focusing on zero-shot tasks for QA and summarization. My experimentation was primarily concentrated on summarization and QA, for which I utilized ROUGE and BLEU scores to assess QA on the unsupervised dataset, and perplexity and loss metrics for text generation evaluation. This comprehensive approach allowed me to explore and understand the diverse applications and limitations of GPT-2 in NLP tasks.

7.3. Implementation Details

The transformers library is pivotal for accessing the GPT-2 model, providing you with pre-trained models and various functionalities essential for NLP tasks. Other major libraries are: torch, nltk, rouge, sklearn, and matplotlib.

In my project, I modified code from scratch, focusing on summarization and QA tasks, while also referencing OpenAI’s GPT GitHub repository for comparative insights. A significant portion of my work relied on using a pre-trained GPT-2 model, especially for benchmarking purposes with evaluation metrics like BLEU, ROUGE, and perplexity. The tokenization process was a critical step, involving loading the GPT-2 model and tailoring it to specific datasets, resulting in four distinct sets of code for each dataset. My approach included careful task-specific adjustments, along with dedicated efforts in training and fine-tuning the model to suit the diverse data types. Additionally, the implementation plan encompassed a thorough process of evaluating the model using various metrics, alongside the development and configuration of the model to align with the project’s specific requirements. Moreover, I implemented a structured approach for handling both unsupervised and supervised

```
Individual BLEU scores: [0.08407354325835562]
Average BLEU score: 0.08407354325835562
```

Figure 7. BLEU (Bilingual Evaluation Understudy) scores reflects poor performance

datasets. For the unsupervised dataset, I divided the data into training and testing sets. This split was crucial for evaluating the model's performance, ensuring that it was tested on unseen data to assess its generalization capabilities accurately.

Dedicated efforts to creating a validation dataset for the supervised datasets. This involved selecting a portion of the data to serve as a validation set, separate from the training and testing sets. The validation dataset played a key role in fine-tuning and optimizing the model during the training process. It provided a means to monitor and evaluate the model's performance on data that it had not been trained on, allowing for adjustments and improvements before the final testing phase.

7.4. Results and Interpretation

The specific BLEU score shown in the Figure 7 is 0.084, which would typically be considered a low score, suggesting that the evaluated translation or text generation did not closely match the reference text for Custom Dataset that is mixture of Chromatography and Chemistry.

The Figure 8 shows a box plot depicting the distribution of loss values across three different epochs for Joke Dataset. Epoch 1 shows a median loss around 2.5, with a spread indicating variability in loss. Epoch 2 has a higher median loss closer to 2.75, suggesting no improvement or possible overfitting as training progressed. Epoch 3 shows a significant reduction in loss with a median below 2.0, indicating that the model likely improved its performance by the third epoch.

The Figure 9 displays a series of seven generated jokes in response to the prompt "Why did the chicken cross the road?" The jokes are varied, incorporating elements of surprise, political humor, and twists on the traditional punchline. Some jokes take a more surreal or abstract approach, while others play on topical references or unexpected outcomes.

The Figure 10 text generation using CORD Dataset. The generated text comes up with non-coherent answers reflecting poor performance.

In Figure 11, the lyrics is generated using Spotify Dataset which also classifies various parameters with Genre. Reflects incoherent results.

In Figure 12, the Question-Answer using Custom Dataset

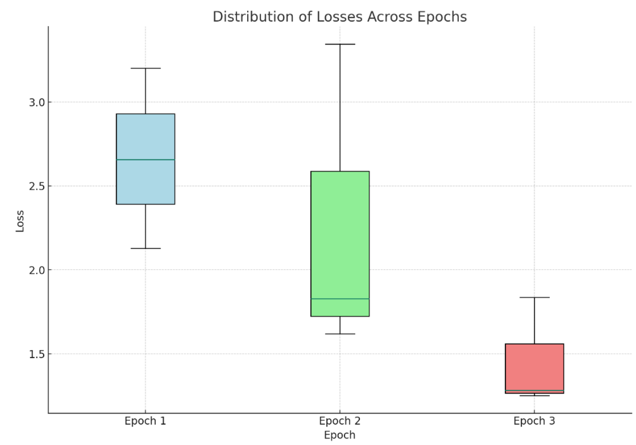


Figure 8. Distribution of Losses Across Epochs

***Prompt 1* ~ Why did the chicken?**

***Generated Output* ~**
Joke 1: Why did the chicken cross the road? To get to your house. Knock knock...Who's there? The chicken.!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!! wish me a Happy Halloween...!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
Joke 2: Why did the chicken cross the road? To get away from Donald Trump.!!!!!!!!!!!!!!!!!!!!!!!!!!!!!! wish you would have heard this before...!!!
Joke 3: Why did the chicken cross the road? He was in the same position as the gay man.!!!!!!!!!!!!!!!!!!!!!!!!!!!...!!!! wish you were still born!!!!!!!!!!!!!!, I'm sorry for this is a joke gone.. *karma train joke*!!!!!!!!!!!!!!
Joke 4: Why did the chicken cross the road? To get to the other side.!!!!!!!!!!!!!!!!!...!!!!, gone.!! wish you were here all along.!!!!!!!!!!!!!!!!!!!!!!!!!! I'll show now.!!!!!!!!!!!! *hands you're gone*
Joke 5: Why did the chicken cross the road? To get to your house. Knock knock! Who's there? The chicken.!!!!!!!!!!!!!!!!!!!!!!!!!!!!!! wish you'd luck!!!!!, I'm going back in two minutes later..!!!!!!!!!!!!!!!!!!!!!!!!!!!!
Joke 6: Why did the chicken cross the road? He heard there was a man standing on the other side.!!!!!!!!!!!!!!!!!!!!!!!!!!!! wish you were dead earlier...!!!
Joke 7: Why did the chicken cross the road? To get to your house. Knock knock Who's there? The chicken!!!!!!!!!!!!!!!!!!!!!!!!!!!!! wish you'd never heard of it before!!

Figure 9. Jokes Generation using Jokes Dataset

***prompt_text* ~** "Recent studies on COVID-19 suggest that"

Recent studies on COVID-19 suggest that its consequences for patients' medical health and related conditions have been more pronounced than previously recognized. These results could have important implications for the COVID-19 pandemic. However, there are numerous limitations to these studies that we believe warrant addressing in future studies. In this article, the main issues were then reviewed, and we identified key concerns and potential solutions.

Figure 10. Text Generation using CORD Dataset

***genre_prompt* ~** "Genre: Pop. Theme: Love and Heartbreak. Lyrics:"

Generated Lyrics:
The Very Best of Lenny Kravitz, Vol. 1 r&b playlist_subgenre:neo soul danceability:0.71 energy:0.52 valence:0.72 tempo:91.99

Figure 11. Lyrics Generation and Genre Classification using Spotify Dataset

Article ~ Green chemistry principles have become increasingly important in various disciplines of chemistry, including analytical chemistry. This is mainly due to an ongoing awareness of the state of the ecosystem and the negative impacts that analytical procedures can have on the environment. The most important approach developed to address this issue is Green Analytical Chemistry (GAC). GAC is an environmental-friendly approach to analytical chemistry that aims to minimize the negative impact of analytical techniques on the environment and human health. This review provided a comprehensive overview of the history of GAC and the existing greenness assessment metric tools. The various approaches, such as the National Environmental Methods Index (NEMI), Eco-scale Assessment (ESA), Green Analytical Procedure Index (GAPI), and Analytical Greens (AGREE) used to evaluate Green profiles were discussed in detail, and their importance in analytical approaches was examined. In addition, research articles on the applicability of assessment tools were discussed in detail in terms of greenness metrics. This review aims to provide researchers with comprehensive information on the various greenness assessment tools and their currently available applications.

Question ~ What is Green Chemistry?

Generated Output ~
Green Chemistry is a concept to encourage using energy-saving instruments, decreasing the use of toxic agents or compounds, and producing waste with a minimal amount during analytical procedures. The concept of Green Chemistry is widely recognized in analytical chemistry laboratories. Simple, comprehensive and flexible greenness metrics should be developed and applied to evaluate the impacts of analytical procedures on the environment, human health and human safety. Sixteen greenness metrics just like National Environmental Methods Index (NEMI

Figure 12. Answer generation using Unsupervised Learning (Custom Datsets)

```
Average ROUGE-1 Score: 0.35658914235442585
Average ROUGE-2 Score: 0.12121211630927477
Average ROUGE-L Score: 0.35658914235442585
```

Figure 13. Metrics to evaluate the quality of summaries produced by text generation systems for Custom Datsets)

after a specific article was fed. Moreover, reflects poorer ROUGE (in Figure 13) and BLEU Scores.

In addition to my other efforts, I crafted a specialized prompt designed for the Spotify-based dataset, aimed at generating lyrics. This prompt was carefully developed to extract specific details from the dataset, particularly focusing on attributes like genre and sub-genre. By integrating these details into the prompt, I enhanced the ability of the model to generate lyrics that are not only contextually relevant but also aligned with the distinct musical styles and themes represented in the dataset.

Also the perplexity score was too low as 0.065 reflecting poor summarization by GPT 2 model. The quality and amount of training data are crucial. If the model is trained on insufficient or poor-quality data, its performance on evaluation metrics like BLEU and ROUGE can be low. Inadequate preprocessing of text data, such as improper tokenization or failure to handle out-of-vocabulary words, can degrade model performance.

8. Conclusion and Discussion

The project demonstrated the impressive adaptability of GPT-2 in processing a wide variety of texts, ranging from scientific articles to creative song lyrics and jokes. This versatility highlights the model's capability in understanding different styles of language and content. Furthermore, the use of a self-generated chromatography and chemistry dataset revealed insightful aspects of the model's potential and limitations, especially in dealing with specialized, technical content in QA and summarization tasks.

One of the key findings of the project was the identification of GPT-2's constraints, especially in supervised classification tasks and its challenges in summarizing complex technical content. These limitations suggest a potential avenue for future research, possibly exploring more specialized models or advanced fine-tuning techniques to overcome these hurdles. My work was also constrained by limited computational resources, which impacted the scope of the experiments. Additionally, while the evaluation methods, including ROUGE, BLEU, and perplexity metrics, provided valuable insights, incorporating a more diverse set of evaluation tools could offer a more comprehensive assessment of the model's performance across various tasks.

This project aligns closely with my initial motivation to delve into the functionalities of GPT-based models in text generation and their applications in different domains. The insights gained from this endeavor are pivotal in understanding the complexities and potential of NLP in various contexts.

9. References

1. Language Models are Unsupervised Multitask Learners - Paper 1
2. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding - Paper 2
3. Language Models are Few-Shot Learners - Paper 3
4. Jokes Dataset - Kaggle
5. CORD Dataset - Kaggle
6. Spotify Dataset - Kaggle
7. Custom Dataset -Google Drive