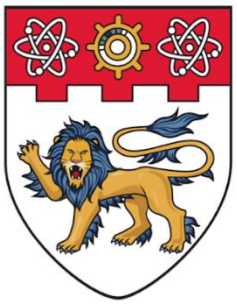


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

SC4079 Final Year Project

Catalyzing Content Generation and Automation: Empowering Diverse Written Content Creation through Generative AI

Submitted by: Lam Hao Fah

Supervised by: Dr Owen Noel Newton Fernando

Examined by: Ast/P Chan Wei Ting, Samantha

School of Computing and Data Science

Nanyang Technological University, Singapore

Submitted in Partial Fulfilment of the Requirements for the Degree of Bachelor of Computing

Table of Contents

1. Introduction

1.1 Background

The emergence of Large Language Models (LLMs) has fundamentally transformed the landscape of automated content generation, establishing a new paradigm in computational linguistics and natural language processing. These sophisticated neural architectures have demonstrated unprecedented capabilities in emulating diverse writing styles, tonal variations, and structural paradigms across multiple domains of written communication. Leveraging transformer-based architectures with attention mechanisms, these models have transcended previous limitations in automated text generation by capturing complex semantic relationships and contextual nuances inherent in human language.

The evolution of LLMs represents a significant advancement beyond traditional rule-based and statistical methods of text generation. Through extensive pre-training on diverse corpora comprising billions of text samples, these models have developed robust representations of linguistic patterns, enabling them to generate coherent, contextually appropriate, and stylistically diverse content.

Contemporary LLMs operate on principles of transfer learning and fine-tuning, allowing them to adapt pre-existing linguistic knowledge to specialized domains and specific stylistic requirements. This adaptability has democratized access to sophisticated content generation tools, empowering organizations and individuals to produce written materials at unprecedented scale and efficiency while maintaining quality standards previously achievable only through human authorship.

1.2 Motivation

The exponential growth in digital content consumption has precipitated an equally substantial demand for high-quality, contextually relevant, and engaging written materials across diverse platforms and channels. Traditional content creation methodologies, predominantly reliant on human authors, face inherent limitations in scalability, consistency, and production efficiency. These constraints have become particularly pronounced in contemporary digital ecosystems where content personalization, multilingual capabilities, and rapid deployment significantly influence engagement metrics and conversion rates.

The emergence of Generative AI presents a compelling solution to these challenges by offering a framework for content automation that preserves - and potentially enhances - creative elements in written communication. By implementing LLM-based content generation systems, organizations can substantially reduce production timelines while maintaining consistent quality standards across diverse content categories. Furthermore, these systems offer the potential for dynamic content adaptation based on audience characteristics, contextual factors, and performance analytics - capabilities that exceed the practical limitations of conventional content creation workflows.

This research is motivated by the need to systematically evaluate the efficacy of different LLM architectures in generating diverse written content and to establish empirical benchmarks for their performance across various dimensions of content quality. By developing and applying creativity-based evaluation metrics, this study aims to provide actionable insights into the comparative advantages of different models and to identify optimization strategies for enhancing their content generation capabilities. Moreover, this investigation seeks to contribute to the broader discourse on human-AI collaboration in creative domains by delineating the complementary strengths of automated systems and human authors.

1.3 Objectives & Scope

This project pursues the following primary objectives:

1. Comparative Model Analysis: Systematically evaluate the relative performance of three distinct language model architectures (GPT-2, GPT-Neo, and OPT-125m) in generating creative, coherent, and contextually appropriate written content across diverse domains.
2. Creativity Metrics Development: Establish and validate a comprehensive evaluation framework specifically designed to quantify the creative dimensions of machine-generated text, encompassing fluency, flexibility, originality, and elaboration.
3. Parameter Efficiency Investigation: Assess the viability of lightweight language models in creative content generation tasks, with particular emphasis on their performance relative to more computationally intensive architectures.
4. Practical Applications Exploration: Develop a functional prototype demonstrating real-world implementation pathways for language model-assisted content creation across multiple domains.

The scope of this research is defined by the following parameters:

1. Model Selection: The study focuses exclusively on decoder-only transformer architectures, with a focus of smaller parameter models (GPT-2, GPT-Neo, and OPT-125m).
2. Content Domains: The evaluation spans five distinct thematic domains: technological forecasting, scientific reporting, counterfactual reasoning, creative narrative, and aesthetic analysis.
3. Evaluation Dimensions: Assessment is limited to four primary dimensions of creativity: fluency (linguistic coherence), flexibility (contextual adaptability), originality (novelty of content), and elaboration (depth and detail).
4. Implementation Context: The practical application prototype is developed within a web-based environment, focusing on interactive content generation with real-time quality assessment.

1.4 Project Resources

To facilitate effective model training and evaluation, the following computational resources and technical frameworks were used.

1.4.1 Computational Infrastructure

The specifications of the systems used in the computing environment were as follows.

- Kaggle's NVIDIA Tesla P100 GPU: Equipped with 16GB VRAM, optimized for high-throughput parallel processing and CUDA-accelerated deep learning workloads.
- Personal NVDIA 4060 laptop GPU: Featuring 8GB VRAM, providing a balance of power efficiency and AI compute capability for local experimentation and fine-tuning.

1.4.2 Model Development Frameworks

The following frameworks were used to optimize model training, evaluation and deployment.

- PyTorch Ecosystem: Served as the primary framework, offering dynamic computational graphs, efficient memory management, and seamless integration with CUDA for hardware acceleration.
- Hugging Face Transformers: Enabled seamless access to pre-trained language models, facilitating transfer learning, model fine-tuning, and advanced NLP capabilities.
- Hugging Face Datasets: Provided an optimized dataset handling pipeline, ensuring scalable ingestion, preprocessing, and on-the-fly transformations with minimal memory overhead.
- Accelerate Framework: Optimized multi-GPU training, implemented mixed-precision computation for improved efficiency, and streamlined large-scale distributed model training.

2. Literature Review

This review examines the evolution of language models for content generation, creativity assessment frameworks in AI-generated text, and model efficiency considerations relevant to this study's focus on parameter-efficient language models for creative content generation.

2.1 Evolution of Language Models for Content Generation

2.1.1 Pre-Transformer Approaches

Early text generation systems relied primarily on rule-based and statistical methods with limited generative capabilities. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) architectures represented significant advances by capturing sequential dependencies in text. Karpathy [30] demonstrated character-level language modeling using RNNs, establishing a foundation for neural text generation. However, these approaches struggled with long-range dependencies and coherence in extended text generation tasks.

2.1.2 Transfromer Architecture and GPT Models

The introduction of the Transformer architecture by Vaswani et al. [26] revolutionized NLP by implementing a self-attention mechanism capable of capturing contextual relationships regardless of sequential distance. This breakthrough enabled efficient parallel processing and the development of more capable language models like GPT (Generative Pre-trained Transformer).

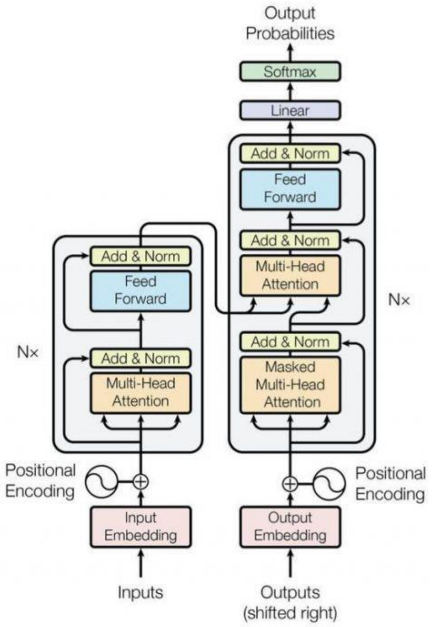


Figure 1: The Transformer - model architecture.

Figure 1: Transformer Model Architecture [26]

Radford et al. [27] introduced the original GPT model, demonstrating the efficacy of unsupervised pre-training followed by task-specific fine-tuning. The subsequent GPT-2 model represented a substantial scaling of this approach with 1.5 billion parameters, exhibiting remarkable zero-shot task performance across diverse generation tasks. As noted by Radford et al., GPT-2 significantly advanced automated content generation by facilitating contextually coherent and stylistically versatile outputs.

2.1.3 Parameter-Efficient Models

While scaling parameters has proven effective for enhancing capabilities, recent research has focused on developing more efficient architectures. EleutherAI's GPT-Neo adapted the GPT architecture with local attention patterns and sparse factorization techniques to optimize computational efficiency. Trained on "The Pile" - an extensive, diverse dataset spanning multiple domains - GPT-Neo's smaller parameter variant (125 million parameters) demonstrated computational efficiency without substantial performance degradation [10].

Similarly, Meta AI's OPT models demonstrated comparable performance while prioritizing efficiency and transparency. The OPT-125m model "employed learned positional embeddings

and optimization techniques tailored to minimize computational load," showing that parameter-efficient models could deliver competitive results in content generation tasks [12]. These developments highlight the potential for smaller, more accessible models to achieve competitive performance in specific tasks, thereby "democratizing access to powerful generative AI technologies even within resource-constrained environments" [12].

2.2 Creativity Assessment in AI-Generated Text

2.2.1 Multidimensional Creativity Models

Zhao et al. [13] emphasize the importance of comprehensive creativity evaluation, recommending multi-dimensional metrics encompassing these four key dimensions:

- **Fluency**: Fluency has traditionally been evaluated using metrics like sentence length variability and lexical diversity. Kann, Rothe, and Filippova [14] demonstrated correlations between sentence structure variability and human quality judgments, while Feng et al. [15] suggest that moderate sentence length correlates strongly with perceived quality. McCarthy and Jarvis [16] established that lexical diversity metrics effectively predict human judgments of writing quality and textual fluency.
- **Flexibility**: Addressing a model's adaptability across diverse contexts, flexibility can be measured by topical diversity and semantic range. Tausczik and Pennebaker's [18] research established that higher variety of linguistic features indicates increased topical breadth. Reimers and Gurevych [19] showed that assessing semantic dispersion through sentence embeddings effectively quantifies a model's conceptual breadth, providing insight into its adaptability across varying discourse domains. Barzilay and Lapata [20] further introduced entity-based approaches to evaluating the smoothness of topic transitions.
- **Originality**: A core dimension of creativity, originality evaluates the uniqueness of generated content. Gatt and Krahmer [21] proposed that n-gram diversity metrics effectively capture linguistic novelty. Ethayarajh [22] examined contextual elasticity in word embeddings as a potential measure of innovative expression. Furthermore, dynamic reference comparisons with existing corpora, such as Wikipedia articles, offer robust means of quantifying textual innovation relative to established knowledge.

- **Elaboration**: Elaboration assesses depth and detailed development within generated texts. McNamara et al.'s Coh-Metrix framework [23] provides automated metrics for assessing text elaboration through syntactic and cohesion features. Biber and Gray's [24] analysis of phrasal elaboration underscores the importance of measuring syntactic complexity and descriptive richness. Together, these approaches enable comprehensive evaluation of how extensively generated text develops and explains concepts.

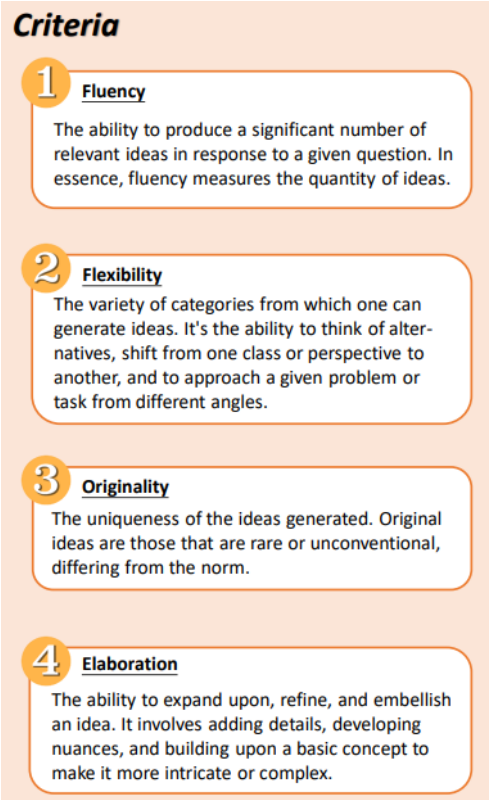


Figure 2: Creativity Criteria [13]

2.2.2 Automated Evaluation Systems

Recent advances in automated evaluation have moved beyond traditional metrics like BLEU and ROUGE. The TTCS (Task-specific, Thorough, Comparable, and Scalable) framework proposed by Zhao et al. [13] established a comprehensive approach to creativity assessment in language models.

2.3 Parameter Efficiency

Given the practical constraints on computational resources, parameter efficiency remains critical for widespread adoption of generative AI. Methods like Curriculum Learning, introduced by Bengio et al. and elaborated by Nekamiche [25], dynamically optimize training complexity, enhancing models' learning efficiency. Similarly, Dynamic Evaluation [28] and Low-Rank Adaptation (LoRA), described by Hu et al. [29], present targeted techniques for real-time adaptation and efficient fine-tuning, respectively, significantly improving model performance without extensive retraining.

2.4 Research Gap and Current Study

Despite significant advances in language model development and evaluation, several important research gaps remain in the current literature: (1) while extensive research exists on large-scale language models (>10B parameters) and their creative capabilities, there remains a significant gap in understanding the creative potential of parameter-efficient models specifically for diverse content generation tasks; (2) though creativity assessment frameworks have been developed for computational systems broadly, few studies have applied comprehensive multidimensional evaluation specifically to parameter-efficient language models, particularly comparing models with different architectural approaches but similar parameter counts; and (3) limited research exists connecting theoretical evaluations of language model creativity to practical deployment considerations in resource-constrained environments, creating a gap between academic findings and industry implementation.

This study addresses these gaps by systematically evaluating three distinct language models (GPT-2, GPT-Neo, and OPT-125m) across multiple dimensions of creativity, focusing on their abilities to generate diverse written content types. The research integrates established creativity dimensions (fluency, flexibility, originality, and elaboration) into an automated evaluation framework tailored to assess the creative quality of generated content.

By focusing on smaller models with practical deployment potential, this study aims to bridge theoretical research on language model capabilities with practical applications in content generation, providing insights into efficient model selection and optimization for diverse writing tasks. Furthermore, the implementation of advanced fine-tuning techniques and the

development of a functional prototype demonstrate pathways for real-world application of these models in content creation workflows.

3. Methodology

3.1 Dataset

This project utilized a meticulously curated corpus of text samples characterized by diversity in genre, style, structure, and communicative intent. This dataset was designed to represent various forms of modern written content while maintaining a high level of quality.

3.1.1 Corpus Composition

- News Articles (CNN/DailyMail) [1]:
  - A comprehensive collection of professional journalism consisting of articles from CNN and DailyMail publications.
  - Supports both extractive and abstractive summarization.
  - Selected for their well-structured narrative, diversity of topics, and professional editorial standards.

- Wiki-based Content (WikiText-103) [2]:
  - High-quality, well-formatted articles extracted from Wikipedia.
  - Selected for their encyclopaedic style, informational density, and diverse subject matter.

- Product/Business Reviews (Yelp Reviews) [3]:
  - User-generated reviews of businesses and products from the Yelp platform.
  - Selected to represent consumer-oriented content with descriptive language and evaluative components.

- Academic Content (GLUE QNLI) [4]:
  - Question-Natural Language Inference dataset derived from Stanford Question Answering Dataset.
  - Contains sentence pairs reflecting academic and educational writing styles.
  - Selected to incorporate formal academic writing conventions and domain-specific terminology.

- Narrative Literature (BookCorpus) [5]:

- Collection of novels spanning diverse genres self-published by authors on smashwords.com.
- Contains approximately 7,185 books representing creative, narrative-driven writing.
- Selected to incorporate creative writing elements and long-form narrative structures into the corpus.

3.1.2 Data Preparation

The collected datasets underwent a systematic cleaning and curation process to ensure consistency, quality, and appropriate length distribution:

- Text Normalization and Cleaning:

- Applied Unicode normalization (NFKC) to standardize character representations.
- Implemented comprehensive regex-based cleaning operations including:
  - Non-ASCII character handling: Removing non-ASCII characters
  - HTML tag removal: Stripping any embedded HTML markup
  - Whitespace normalization: Replacing multiple spaces and newlines with single spaces
  - Special characters handling: Catch and replace unexpected symbol sequences

- Length-Based Filtering:

- Implemented dual-threshold filtering to remove outlier documents:
  - Minimum length threshold of 50 characters to eliminate fragments and incomplete entries
  - Maximum length threshold of 5,000 characters to prevent overly long documents from dominating the corpus
- This balanced approach ensured content remained substantive while maintaining manageable sequence lengths for model training.

- Dataset Integration and Consolidation:

- Combined datasets and retained only essential columns ('source' and 'clean_text').
- Ensured consistent formatting to facilitate seamless tokenization.
- Generated dataset statistics to document the final corpus composition.

This methodical approach resulted in a clean, balanced corpus maintaining the distinctive characteristics of each content category while ensuring consistent quality standards across the dataset. The final corpus served as the foundation for model fine-tuning, with source tracking enabling detailed analysis of performance across different content types.
This extensive dataset provided a robust foundation for training models capable of generating diverse content types while maintaining coherence, relevance, and stylistic appropriateness across varying communicative contexts.

3.2 Models

This project implemented and evaluated three distinct language model architectures, with a deliberate focus on parameter-efficient models for content generation. Apart from GPT-2, which serves as a baseline for comparative analysis, the other models were chosen to demonstrate the effectiveness of lightweight architectures in content generation tasks. This is to accommodate limited GPU availability, memory constraints, and practical feasibility in real-world deployment scenarios.

3.2.1 GPT-2

Representing a pioneering architecture in generative language modeling, GPT-2 provided a valuable comparative baseline:

- **Fundamental Design**: Incorporates a vanilla transformer decoder architecture [6] with 1.5 billion parameters in the implementation used.
- **Attention Mechanism**: Utilizes conventional multi-head attention [7].
- **Tokenization Approach**: Employs byte-pair encoding (BPE) with a vocabulary of 50,257 tokens [8], facilitating efficient representation of diverse linguistic elements.

GPT-2 was integrated into this project to provide historical context for model evolution and to establish performance benchmarks.

3.2.2 EleutherAI/GPT-Neo

The GPT-Neo model developed by EleutherAI offers an alternative approach to content generation with specific architectural characteristics:

- **Training Corpus**: Originally trained on The Pile, a diverse 800GB dataset encompassing scientific papers, code repositories, web content, and books [9].
- **Architectural Innovations**: Implements local attention mechanisms and sparse factorized attention patterns [10] to optimize computational resources.
- **Parameter Scale**: Compact architecture with 125 million parameters, offering computational efficiency while maintaining generation capabilities [11].

This model was selected to evaluate the efficacy of more compact architectural approaches in content generation tasks, particularly in scenarios where computational resources or deployment constraints might favour smaller models.

3.2.3 OPT-125m

The OPT-125m model represents Facebook AI Research's approach to efficient large language modeling:

- **Foundation Architecture**: Based on a decoder-only transformer framework with 125 million parameters [12]
- **Optimization Approach**: Uses learned positional embeddings and standard transformer architecture optimized for efficiency
- **Implementation Advantages**: Offers a lightweight yet capable model suitable for content generation tasks with reduced computational requirements

The implementation utilized OPT-125m's capabilities as a smaller yet effective model for content generation across multiple genres while maintaining reasonable computational efficiency.

3.2.4 Model Training Configuration

Each model underwent specialized fine-tuning procedures optimized for content generation tasks, with consistent training infrastructure implemented across all three models:

- **Training Protocol:**

  - Implemented supervised fine-tuning using autoregressive causal language modeling (next-token prediction)

- Utilized Hugging Face's Trainer API with DataCollatorForLanguageModeling configured for non-masked language modeling
- Applied consistent tokenization with max sequence length of 512 tokens
- Implemented efficient processing with multi-worker data loading and parallel preprocessing

- **Optimization Strategy:**

- Applied consistent AdamW optimizer settings across all models
- Enabled mixed precision training for all models to accelerate computation
- Used identical gradient accumulation steps across all architectures

- **Hyperparameter Consistency:**

- Learning rate: 3e-5 applied uniformly across all models
- Epochs: All models trained for exactly 3 epochs
- Batch size: Consistent per_device_train_batch_size of 8 for all models
- Evaluation strategy: Epoch-based evaluation applied uniformly

- **Evaluation Approach:**

- Implemented consistent eval_dataset selection (first 1000 examples) for all models
- Applied identical validation metrics and evaluation frequency
- Maintained uniform early stopping criteria based on validation loss

3.3 Creativity Evaluation Framework

This project implemented an advanced evaluation methodology that combines established creativity assessment approaches with the Task-specific, Thorough, Comparable, and Scalable (TTCS) framework [13] for comprehensive LLM evaluation. This integrated approach enabled nuanced assessment of creative content generation capabilities.

3.3.1 Fluency Evaluation

Fluency evaluation quantified linguistic quality and textual flow through a computationally efficient composite methodology.

- **Sentence Structure Analysis**: Research by Kann, Rothe and Fippova [14] has found that sentence length variability correlates significantly with expert ratings of text quality, while Feng et al. [15] established that maintaining moderate sentence lengths typically optimizes readability.

```
sentences = sent_tokenize(text)
all_words = word_tokenize(text.lower())
words = [w for w in all_words if w.isalnum()]

if len(sentences) > 0:
    avg_sent_length = len(words) / len(sentences)
else:
    avg_sent_length = 0
```

Figure 3: Code for Sentence Length Analysis

As seen in Figure 3, sentence tokenization was implemented using NLTK's sent_tokenize function to calculate average sentence length, with optimal scores assigned to texts with moderate sentence lengths.

- **Lexical Diversity Measurement**: Lexical diversity refers to the measure of unique words to total words in a text. According to McCarthy and Jarvis [16], lexical diversity metrics strongly predict human judgments of writing quality and is effective in judging fluency of a text.

```
if len(words) > 0:
    lexical_diversity = len(set(words)) / len(words)
else:
    lexical_diversity = 0
```

Figure 4: Code for Lexical Diversity Measurement

As seen in Figure 4, volcabulary richness was quantified by calculating the type-token ratio, providing an indicator of linguistic variety and expressive range.

- **Grammatical Structure Evaluation**: Linguistic research by Ragheb and Dickinson [17] discusses the importance of basic syntactic structures in grammaticality assessment. Errors in synthetic structures such as simple subject-verb-object patterns can be some of the clearest indicators that a text is grammatically unsound.

```
doc = nlp(text)
grammatical_score = 0

for sent in doc.sents:
    has_subj = any(token.dep_ in ('nsubj', 'nsubjpass') for token in sent)
    has_verb = any(token.pos_ == 'VERB' for token in sent)
    if has_subj and has_verb:
        grammatical_score += 1

if len(list(doc.sents)) > 0:
    grammatical_score /= len(list(doc.sents))
```

Figure 5: Code for Synthetic Structure Evaluation

As seen in Figure 5, spaCy's dependency parsing was utilized to identify subject-verb relationships within sentences, assessing the presence of fundamental grammatical structures essential for coherent expression.

3.3.2 Flexibility Evaluation

The flexibility dimension assessed each model's adaptability across diverse content requirements through computational analysis of semantic and topical variations.

- **Topic Diversity Measurement**: According to Tausczik and Pennebaker, a higher variety of linguistic features such as nouns indicates increased topical breadth [18]. In the context of Large Language Models (LLMs), this notion of topical breadth can be viewed as one aspect of the model's flexibility: the more diverse the topics it can handle coherently, the more adaptable it appears in generating or processing discourse across different domains.

```
doc = nlp(text)
key_nouns = [token.lemma_ for token in doc if token.pos_ == 'NOUN' and token.text.lower() not in stop_words]
topic_diversity = len(set(key_nouns)) / len(key_nouns) if key_nouns else 0
```

Figure 6: Code for Topical Diversity Measurement

As seen in Figure 6, the range of topics addressed was quantified by analyzing the ratio of unique noun lemmas to total nouns, after filtering stop words.

- **Semantic Range Analysis**: Research by Reimers and Gurevych has found that by examining the distribution or dispersion of sentence embeddings, we can reveal how

widely a text spans conceptually [19]. In this project, semantic range analysis can indicate how effectively an LLM navigates diverse topics or conceptual domains.

```
sentences = list(doc.sents)
if len(sentences) >= 2:
    sent_embeddings = np.array([sent.vector for sent in sentences])
    similarities = cosine_similarity(sent_embeddings)
    semantic_range = 1 - (np.sum(similarities) - len(sentences)) / (len(sentences) * (len(sentences) - 1))
else:
    semantic_range = 0
```

Figure 7: Code for Semantic Range Analysis

As seen in Figure 7, sentence embedding comparisons were implemented using spaCy's vector representations to measure semantic distance between sentences. Cosine similarity calculations between sentence pairs were averaged and inverted to produce a semantic range score.

- **Concept Transition Tracking**: Barzilay and Lapata introduced an entity-based approach to local coherence, showing how tracking the presence or absence of key entities across adjacent sentences helps evaluate the smoothness of topic transitions [20].

```
concept_transitions = 0
prev_key_entities = set()
for sent in sentences:
    sent_entities = set([token.lemma_ for token in sent
                         if token.pos_ in ('NOUN', 'PROPN') and token.text.lower() not in stop_words])
    if prev_key_entities and (len(sent_entities.intersection(prev_key_entities)) / max(1, len(prev_key_entities)) < 0.3):
        concept_transitions += 1
    prev_key_entities = sent_entities
concept_transitions = concept_transitions / (len(sentences) - 1) if len(sentences) > 1 else 0
```

Figure 8: Code for Concept Transition Tracking

As seen in Figure 8, concept transitions are measured by comparing sets of core nouns between sentences. Adept handling of these transitions signals greater flexibility in generating diverse yet coherent text. If an LLM can pivot to new topics without losing coherence, it demonstrates conceptual agility.

3.3.3 Originality Evaluation

Originality metrics focused on quantifying the uniqueness and innovative qualities of generated content using n-gram analysis and vector similarity measures.

- **Lexical Novelty Measurement**: Gatt and Krahmer asserts that diversity metrics (measuring the number of unique 1-grams, 2-grams, etc.) can reflect linguistic

novelty or creativity [21]. Evidently, n-gram diversity helps quantify how varied a model's output is, partially reflecting the model's ability to produce "original" text.

```
doc = nlp(text)
tokens = [token.text.lower() for token in doc if token.is_alpha]

trigrams = [' '.join(tokens[i:i+3]) for i in range(len(tokens) - 2)]
if not trigrams:
    lexical_novelty = 0
else:
    unique_trigrams = set(trigrams)
    lexical_novelty = len(unique_trigrams) / len(trigrams)
```

Figure 9: Code for Lexical Diversity Measurement

As seen in Figure 9, lexical novelty was calculated by comparing the ratio of unique trigrams to total trigrams within the generated content.

- **Phrase Novelty Analysis**: As seen in Figure 10, we compare the proportion of text trigrams not found in a reference corpora, providing a direct measure of linguistic innovation relative to existing content. Figure 11 shows that the reference materials were dynamically sourced from Wikipedia. For each evaluation, the system retrieved the top three Wikipedia articles related to the content topic, creating a domain-relevant reference corpus against which originality could be measured.

```
reference_trigrams = []
for ref_text in reference_texts:
    ref_doc = nlp(ref_text)
    ref_tokens = [token.text.lower() for token in ref_doc if token.is_alpha]
    reference_trigrams.extend([' '.join(ref_tokens[i:i+3]) for i in range(len(ref_tokens) - 2)])

phrase_novelty = (sum(1 for tg in trigrams if tg not in reference_trigrams) / len(trigrams)) if trigrams else 0
```

Figure 10: Code for Phrase Novelty Analysis

```
def get_reference_texts(query):
    search_results = wikipedia.search(query)
    reference_texts = []
    for title in search_results[:3]:
        try:
            page = wikipedia.page(title)
            reference_texts.append(page.content)
            print(f"Retrieved content for page: {title}")
        except Exception as e:
            print(f"Could not retrieve page for {title}: {e}")
    return reference_texts
```

Figure 11: Code for Sourcing Reference Corpora

- **Reference Similarity Calculation**: Research by Ethayarajh focuses on how word embeddings change across different contexts [22]. The geometry of these embeddings shows how "contextually elastic" a model is. For instance, a model that dramatically reconfigures word embeddings in response to nuanced context might be more adept at generating content that deviates from stock phrases or well-trodden linguistic patterns, potentially reflecting an increased scope for "original" expression.

```
doc_vector = doc.vector
reference_similarities = []
for ref_text in reference_texts:
    ref_doc = nlp(ref_text)
    similarity = cosine_similarity(
        doc_vector.reshape(1, -1),
        ref_doc.vector.reshape(1, -1)
    )[0][0]
    reference_similarities.append(similarity)
reference_similarity = max(reference_similarities) if reference_similarities else 0
```

Figure 12: Code for Reference Similarity Calculation

As seen in Figure 12, document vector embeddings and cosine similarity were employed to measure semantic distance between generated text and the aforementioned reference corpora. We can use this to estimate the degree of novelty or distance from existing content.

3.3.4 Elaboration Evaluation

The elaboration dimension evaluated depth and development of ideas within generated content through computational analysis of syntactic structures and explanation indicators.

- **Detail Density Analysis**: The Coh-Metrix research by McNamara, Graesser, McCarthy and Cai offers various text-analysis indices – including counts and ratios of parts of speech – that can be used to gauge a text's level of detail and descriptiveness [23].

```
doc = nlp(text)
tokens = list(doc)

detail_tokens = [token for token in doc if token.pos_ in ('ADJ', 'ADV') or token.dep_ == 'prep']
detail_density = len(detail_tokens) / len(tokens) if tokens else 0
```

Figure 13: Code for Detail Density Analysis

As seen in Figure 13, counts of adjectives, adverbs, and prepositions were correlated with perceived text richness. By quantifying the relative frequency of these elements, we obtain an empirical gauge of how elaborated or vividly detailed a given text may be, facilitating both qualitative assessments (e.g., how engaging or descriptive the prose is) and more systematic evaluations (e.g., comparing different versions of a text or outputs from various language models).

- **Descriptive Richness Calculation:** Biber and Gray argue that academic prose is often characterized by phrasal elaboration rather than long, clausal structures [24]. One way to measure this phrasal complexity is by looking at how densely nouns are modified, for instance, through prepositional phrases, relative clauses, or adjective use.

```
nouns = [token for token in doc if token.pos_ in ('NOUN', 'PROPN')]
adjectives = [token for token in doc if token.pos_ == 'ADJ']
if nouns:
    avg_adj_per_noun = len(adjectives) / len(nouns)
else:
    avg_adj_per_noun = 0

scaled_richness = min(avg_adj_per_noun / 0.5, 1)
```

Figure 14: Code for Descriptive Richness Calculation

As seen in Figure 14, the average number of adjectives per noun was normalized and measured. Higher adjective-to-noun ratios generally signal more extensive modification and elaboration.

- **Explanation Depth Measurement:** As aforementioned, the Coh-Metrix framework [23] also describes how causal, logical, and intentional connectives (e.g., because, therefore, thus) contribute to text cohesion and can serve as indicators of explanatory depth.

```
explanation_keywords = {'because', 'since', 'therefore', 'thus', 'consequently', 'due', 'hence'}
keyword_count = sum(1 for token in doc if token.text.lower() in explanation_keywords)

advcl_count = sum(1 for token in doc if token.dep_ == 'advcl')
total_explanation = keyword_count + advcl_count
sentences = list(doc.sents)
explanation_depth = total_explanation / len(sentences) if sentences else 0
```

Figure 15: Code for Explanation Depth Measurement

As seen in Figure 15, a dual approach to assessing explanatory content was implemented by counting both explanation keywords (e.g. "because", "therefore", "thus") and adverbial clause modifiers (identified through dependency parsing). Explanation depth thus measures how extensively the text employs causal or logical connectives.

3.3.5 Overall Creativity Score

For each of the preceding evaluation metrics, a score was computed through a weighted formula combining their respective components. Table 1 below shows how the scores were tabulated.

Metric	Formula
fluency	fluency_score = average_sentence_length * 0.3 + lexical_diversity * 0.3 + grammatical_score * 0.4
flexibility	flexibility_score = topical_diversity * 0.4 + semantic_range * 0.3 + concept_transitions * 0.3
originality	originality_score = lexical_novelty * 0.4 + phrase_novelty * 0.3 + (1 – reference_similarity) * 0.3
elaboration	elaboration_score = detail_density * 0.4 + richness * 0.3 + explanation * 0.3

Table 1: Breakdown of Scores for Evaluation Metrics

After deriving an individual score for fluency, flexibility, originality, and elaboration, each was weighted equally at 25%. This ensures that no single dimension disproportionately influences the overall creativity score. By balancing these components, the combined metric provides a more comprehensive assessment of a text's creative qualities, reflecting not just how well-formed it is (fluency), but also how broad (flexibility), novel (originality), and detailed (elaboration) the content appears.

4. Results & Evaluation

4.1 Prompt Design and Selection

The evaluation utilized five distinct prompts representing different knowledge domains and rhetorical structures to ensure comprehensive assessment across diverse thematic areas:

1. "The future of artificial intelligence looks": A forward-looking prompt on technology.
2. "The latest research on climate change suggests": A science-focused, evidence-based prompt.
3. "If time travel were possible, humanity would": A speculative, counterfactual premise.
4. "Deep beneath the ocean's surface, the greatest mystery awaiting explorers is": A creative narrative prompt encouraging descriptive elaboration.
5. “The beauty of creative writing lies in”: A reflective prompt on literary aesthetics and craft.

These prompts were deliberately designed to test different types of knowledge and reasoning capabilities, spanning technological forecasting, scientific reporting, counterfactual reasoning, creative narrative, and trend analysis. The diversity of domains and rhetorical structures enabled a balanced assessment of each model's versatility across varied content generation scenarios.

4.2 Evaluation Methodology

The evaluation process followed a structured, systematic approach to ensure consistency and statistical reliability:

- For each prompt, 100 text completions were generated from each model.
- Reference texts were dynamically retrieved from Wikipedia for each prompt to enable originality assessment.
- Each generated text was evaluated across all four dimensions (fluency, flexibility, originality, and elaboration).
- Scores were averaged across the 100 completions for each prompt.

- Final model performance metrics represent the mean scores across all five prompts and all completions (500 samples per model).

This methodology ensured sufficient sample size to account for generation variability while maintaining consistency across model comparisons. The use of 500 total evaluations per model (5 prompts × 100 completions) provided robust statistical foundation for performance analysis.

4.3 Comparative Performance Results

The evaluation process followed a structured approach to generate and assess content across diverse thematic domains. For each model, the evaluation utilized five distinct prompts representing different knowledge domains and rhetorical structures to ensure comprehensive assessment across diverse thematic areas.

4.3.1 Overall Model Performance

Systematic evaluation across 500 samples per model (100 completions per prompt) yielded detailed insights into each model's relative performance strengths and weaknesses. The overall creativity scores, along with other relevant metrics, are summarized in Table 2 below.

Model	Fluency	Flexibility	Originality	Elaboration	Overall Creativity
GPT-2	0.933	0.637	0.732	0.528	0.708
GPT-Neo	0.838	0.548	0.719	0.395	0.625
OPT-125	0.905	0.620	0.728	0.469	0.681

Table 2: Average Scores for Each Model

GPT-2 demonstrated the highest overall creativity (0.708), indicating its advantage in linguistic coherence, detailed content elaboration, and contextual adaptability, likely attributed to its larger parameter count and extensive training. OPT-125's performance as second-best (0.681) highlights the effectiveness of Meta's architectural optimizations, particularly in achieving parameter efficiency without substantially compromising creative output. Although GPT-Neo ranked lowest overall, it showed notable performance in originality (0.719), emphasizing its potential in generating novel content even with a relatively smaller model size.

4.3.2 Prompt-Specific Performance

Each model demonstrated varying capabilities across different prompt types, revealing important context-dependent strengths and limitations. Tables 3-5 below present the detailed prompt-specific scores for each model. The prompts are mentioned in section 4.1.

Prompt	Fluency	Flexibility	Originality	Elaboration	Overall Creativity
1	0.942	0.637	0.705	0.494	0.694
2	0.939	0.618	0.708	0.545	0.703
3	0.929	0.632	0.717	0.572	0.712
4	0.934	0.643	0.797	0.470	0.711
5	0.924	0.656	0.734	0.558	0.718

Table 3: GPT-2 Prompt Performance

Prompt	Fluency	Flexibility	Originality	Elaboration	Overall Creativity
1	0.853	0.538	0.686	0.390	0.617
2	0.849	0.517	0.690	0.357	0.603
3	0.822	0.561	0.723	0.481	0.647
4	0.827	0.569	0.776	0.336	0.620
5	0.842	0.555	0.718	0.413	0.632

Table 4: GPT-Neo Prompt Performance

Prompt	Fluency	Flexibility	Originality	Elaboration	Overall Creativity
1	0.914	0.619	0.700	0.435	0.667
2	0.912	0.624	0.706	0.461	0.676
3	0.885	0.631	0.717	0.521	0.689
4	0.950	0.611	0.789	0.418	0.692
5	0.911	0.615	0.731	0.509	0.692

Table 5: OPT-125 Prompt Performance

All models performed exceptionally well on Prompt 4, consistently yielding their highest originality scores. This suggests that exploratory and imaginative narrative prompts effectively stimulate creative outputs, irrespective of model architecture.

Flexibility scores showed limited variation, suggesting that adaptability to varied contexts is primarily driven by model architecture rather than prompt-specific influences. GPT-2 consistently outperformed GPT-Neo and OPT-125, underscoring the importance of parameter size for contextual flexibility.

Fluency scores were stable across all prompts, particularly notable in GPT-2's consistent linguistic coherence. This highlights the transferability of fluency across different content domains once achieved at the architectural and training level.

Elaboration exhibited the most substantial variability, with GPT-2 and OPT-125 notably performing better on speculative (Prompt 3) and reflective (Prompt 5) prompts. This indicates these types of prompts effectively encourage detailed content generation, suggesting a potential model specialization area for improving creative writing outputs.

Future research should further investigate the relationship between prompt characteristics and model creativity, potentially guiding the development of prompts specifically tailored to maximize each model's creative potential.

4.3.3 Limitations and Challenges

The evaluation methodology and results presented in this project must be considered within the context of several important limitations and challenges.

4.3.3.1 Model Selection Constraints

- **Parameter Range Limitations**: The models examined represent a narrow band of the parameter scale spectrum (125M and 1.5B parameters), excluding both very small models (<100M parameters) and very large models (>10B parameters).
- **Architectural Homogeneity**: All evaluated models employ decoder-only transformer architectures, limiting insights regarding encoder-decoder or encoder-only designs.

- **Training Data Overlap**: Potential overlap in pre-training corpora across models may influence comparative performance beyond architectural differences.

4.3.3.2 Evaluation Framework Limitations

- **Computational Approximation**: The automated metrics represent computational approximations of creativity dimensions originally developed for human assessment.
- **Reference Dependency**: Originality assessment relies on comparison to reference corpora, introducing potential domain bias.
- **Dimensional Weighting**: Equal weighting of dimensions (25% each) represents a subjective choice rather than an empirically optimized approach.
- **Context Insensitivity**: The evaluation does not account for context-dependent appropriateness of creative elements.

4.3.3.3 Technical Implementation Challenges

- **Computational Resource Constraints**: Limited GPU availability constrained hyperparameter optimization and model variant exploration.
- **Reference Corpus Limitations**: Wikipedia-based reference collection introduced potential bias toward encyclopedic writing styles.
- **Prompt Sensitivity**: Generation outcomes demonstrated high sensitivity to prompt phrasing, potentially influencing comparative results.

5. Advanced Fine-Tuning

Given GPT-2's superior overall performance, further exploration of advanced fine-tuning strategies was pursued to enhance the model's capabilities beyond initial optimization efforts. The following approaches were employed to push the boundaries of GPT-2's generative effectiveness.

5.1 Curriculum Learning

Curriculum learning strategies [25] were applied to gradually increase complexity, thereby improving the model's ability to handle complex creative tasks:

- **Complexity-Based Sequencing**: Training examples were systematically organized from simplest to most complex based on linguistic features including sentence length, vocabulary diversity, and syntactic structure.
- **Progressive Training Schedule**: The training process advanced through five distinct stages, with each stage introducing increasingly complex examples.
- **Adaptive Epoch Allocation**: More training epochs were allocated to early curriculum stages to establish fundamental capabilities, with fewer epochs for later stages to refine complex generation abilities.

5.2 Dynamic Evaluation

Dynamic evaluation was implemented to enable real-time adaptation of GPT-2's parameters:

- **Continuous Model Adaptation**: The model parameters were dynamically updated after generating text, enabling rapid adaptation to specific contexts and user interactions.
- **Drift Control Mechanisms**: To prevent excessive deviation from base capabilities, we implemented automated stability monitoring and selective parameter resetting.
- **Weighted Adaptation**: Different learning rates and update weights were applied for dynamic updates versus standard training updates, enabling fine-grained control over adaptation speed.

5.3 Parameter-Efficient Techniques

Low-Rank Adaptation (LoRA) was adapted for parameter efficiency:

- **Targeted Adaptation**: Applied LoRA techniques to fine-tune only the essential layers responsible for task-specific adaptations.
- **Computational Efficiency**: Achieved comparable improvements in creativity metrics with significantly reduced computational overhead and training time.

The three complementary strategies were integrated into a unified fine-tuning pipeline that leveraged their respective strengths while minimizing potential conflicts. The implementation architecture consisted of three key components: a complexity analyser that calculated linguistic complexity scores for training examples based on sentence structure, vocabulary diversity, and syntactic patterns; custom LoRA Wrappers that added trainable low-rank decomposition matrices to frozen model weights to focus optimization on a small parameter subset; and a dynamic trainer that incorporated both curriculum sequencing and periodic dynamic updates with built-in mechanisms to prevent excessive model drift during adaptation.

5.4 Results

Our focused approach combining Curriculum Learning, Dynamic Evaluation, and LoRA significantly enhanced GPT-2's creativity metrics.

Prompt	Original Creativity Score	Fine-Tuned Creativity Score	% Change
1	0.694	0.752	+8.4%
2	0.703	0.761	+8.3%
3	0.712	0.785	+10.3%
4	0.711	0.774	+8.9%
5	0.718	0.807	+12.4%

Table 6: Fine-Tuned GPT-2 Performance

6. Web Application Prototype & Real-World Applications

6.1 Interactive Prototype Architecture

Building on the empirical findings from Section 5, an interactive web-based application was developed to demonstrate the practical application of GPT-2, which exhibited superior creativity characteristics among the evaluated models. This deployment framework serves dual objectives: enabling intuitive user-model interaction and operationalizing the Creativity Evaluation Framework within a production-oriented environment. Screenshots of the application can be found in Appendix B.

The prototype architecture consists of four integrated components:

1. **User Interface and Parameter Configuration**:
    - The application provides an intuitive interface allowing users to configure two primary generation parameters:
        - Maximum Length: Controls the maximum length of output to be generated
        - Number of Responses: Establishes the number of outputs to be generated
    - The prompt field allows users to specify starting points or content domains
    - Interface design balances technical functionality with accessibility for non-technical users
2. **Model Inference Pipeline:**
    - Backend API mediates communication between the interface and the GPT-2 model
    - Generation parameters are translated into model-compatible constraints
    - Text generation occurs with optimized beam search and temperature settings for creative content
    - Generated output is delivered to the client interface with minimal latency

3. **Real-time Evaluation Framework:**

- Generated content undergoes immediate assessment through the Creativity Evaluation Framework
- NLP libraries (spaCy, NLTK) enable efficient processing of the multi-dimensional creativity metrics
- All four dimensions (fluency, flexibility, originality, elaboration) are calculated in sequence
- Composite scoring follows the methodology detailed in Section 3.3, providing comprehensive quality assessment

4. **Results Visualization and Feedback:**

- The interface presents both generated content and corresponding creativity metrics
- Dimensional breakdown provides transparency into the quality assessment process
- Visual elements aid interpretation of quality metrics

This implementation demonstrates how transformer-based language models can be effectively deployed in interactive environments that combine automated content generation with structured quality assessment.

6.2 Domain Applications and Implementation Scenarios

The prototype illustrates several practical application domains where generative language models can provide significant value.

1. **Creative Development and Narrative Construction:**
   - Ideation Support: Concept generation and exploratory content development for creative professionals
   - Narrative Expansion: Elaboration of plot elements and scenario development based on minimal inputs
2. **Marketing and Brand Communication:**
   - Content Automation: Rapid generation of product descriptions, advertisements, and social media content
   - Audience Targeting: Fine-tuned content creation for specific demographic segments through controlled prompting

- Analytics Integration: Creativity metrics can inform marketing performance indicators and content optimization

3. **Conversational Systems and User Interaction:**
   - Enhanced Response Generation: More natural and contextually appropriate automated communication
   - Personality Consistency: Creating coherent agent identities across multiple interaction points
   - Quality-Optimized Outputs: Real-time creativity assessment to ensure engaging and appropriate responses

6.3 Future Development Trajectories

While this current implementation establishes a foundation for practical language model application, several advancement pathways present opportunities for enhanced functionality:

1. **Model Architecture Enhancements**:
   - Integration of alternative models for specific content requirements
   - Domain-specific fine-tuning for specialized applications
   - Dynamic model selection based on content category and creativity requirements
2. **Advanced Customization Capabilities**:
   - User-provided reference corpus for tailored originality assessment
   - Style and tone parameterization for more controlled generation
   - Weighted creativity dimensions allowing prioritization based on content purpose
3. **Collaborative Editing Environment**:
   - Human-AI collaborative interfaces with interactive feedback
   - Suggestion systems highlighting potential quality improvements
   - Version comparison tools showing creativity metric changes across iterations
4. **Technical Scale Optimization**:
   - Cloud deployment architecture for handling increased user demand
   - Efficient processing for batch content generation
   - Performance optimization for reduced latency in generation and evaluation

5. **Linguistic Expansion**:
    - Multilingual model integration for content generation across languages
    - Cross-language adaptation capabilities
    - Culturally-aware creativity metrics accounting for linguistic variation

This implementation framework demonstrates practical pathways for integrating generative language models within production environments. By combining automated content generation with structured quality assessment, the system provides a blueprint for AI-augmented content creation across diverse professional domains.

7. Conclusion

This project investigated the capabilities of parameter-efficient language models for diverse content generation tasks through rigorous empirical evaluation and practical application development. The comprehensive evaluation of GPT-2, GPT-Neo, and OPT-125m across multiple creativity dimensions has yielded several significant findings with implications for both theoretical understanding and practical implementation of AI-assisted content creation systems.

7.1 Summary of Key Findings

The comparative analysis of model performance across fluency, flexibility, originality, and elaboration dimensions revealed several important insights.

1. **Parameter Efficiency vs. Performance Trade-offs**: While GPT-2 demonstrated superior overall creativity scores, the relatively strong performance of OPT-125m despite its significantly smaller parameter count highlights the effectiveness of architectural optimization in achieving parameter efficiency without substantially compromising creative output quality. This finding challenges the assumption that larger models inherently produce superior creative content, suggesting that well-optimized smaller architectures can achieve competitive performance in specific generation tasks.
2. **Dimensional Performance Variations**: Across all models, fluency consistently achieved the highest scores, while elaboration demonstrated the most significant variation. This pattern suggests that contemporary language model architectures have largely mastered grammatical coherence but continue to face challenges in generating deeply elaborated content with rich detail and explanatory depth. The substantial variation in elaboration scores also indicates this dimension may be the most responsive to architectural and training improvements.
3. **Effectiveness of Advanced Fine-Tuning**: The implementation of curriculum learning, dynamic evaluation, and LoRA techniques produced substantial improvements in GPT-2's creativity metrics. This demonstrates the considerable untapped potential in existing architectures that can be realized through tailored fine-tuning approaches without architectural modification.

7.2 Theoretical and Practical Implications

The findings from this project carry several significant implications for both the theoretical understanding of language model capabilities and their practical applications.

7.2.1 Theoretical Implications

1. **Creativity Dimensionality**: The multi-dimensional evaluation framework validated in this study provides empirical support for treating computational creativity as a composite construct comprising distinct, measurable dimensions. The observed variations across dimensions suggest that creative capabilities in language models develop unevenly, with some aspects (like fluency) advancing more rapidly than others (like elaboration).

2. **Parameter Scaling Reconsidered**: The competitive performance of smaller models challenges simplistic scaling laws that directly correlate parameter count with capability. Instead, results suggest a more nuanced relationship where architectural choices, training methodology, and optimization techniques play crucial roles in determining creative generation capabilities.

3. **Dynamic Adaptation Effects**: The significant improvements achieved through dynamic evaluation techniques provide evidence that real-time parameter adaptation can enhance creative performance, suggesting that static model parameters may impose unnecessary limitations on creative generation potential.

7.2.2 Practical Implications

1. **Model Selection Guidance**: The detailed performance analysis across creativity dimensions provides evidence-based guidance for practitioners selecting models for content generation applications. For applications prioritizing fluency and basic coherence, smaller models may provide sufficient quality with reduced computational demands, while applications requiring elaborate detail may benefit from larger architectures or specialized fine-tuning.

2. **Deployment Optimization**: The prototype implementation demonstrates the feasibility of integrating automated creativity assessment within production environments, enabling quality-aware content generation systems that can dynamically evaluate, and filter outputs based on dimensional quality metrics.

3. **Resource-Efficient Implementation**: The strong performance of parameter-efficient models supports more accessible deployment across diverse computing environments, including resource-constrained scenarios where computational efficiency is paramount.

7.3 Future Research Directions

Based on the findings and limitations of this study, several promising directions for future research emerge.

1. **Creativity Enhancement Techniques**: Further investigation into specialized fine-tuning approaches specifically targeting underperforming creativity dimensions like elaboration could yield significant improvements in overall generation quality. Exploration of reinforcement learning techniques that reward specific creative attributes represents a particularly promising direction.

2. **Cross-Architectural Comparison**: Expanding the evaluation framework to include encoder-decoder architectures and alternative attention mechanisms would provide a more comprehensive understanding of architectural effects on creative generation capabilities.

3. **Human-AI Collaboration Models**: Building on the interactive prototype, research into optimal collaboration patterns between human writers and AI generation systems could identify complementary strengths and develop interfaces that enhance creative outcomes while maintaining human agency and editorial control.

7.4 Concluding Remarks

This project has demonstrated that parameter-efficient language models can generate creative content across diverse domains with quality metrics approaching those of substantially larger architectures. Through systematic evaluation and innovative fine-tuning techniques, the study has identified both the capabilities and limitations of contemporary model architectures while establishing a framework for assessing and enhancing their creative potential.

The interactive prototype implementation provides a practical pathway for deploying these models in real-world content generation contexts, with integrated evaluation metrics ensuring output quality across multiple creativity dimensions. As language model technology continues to evolve, this research provides both methodological tools and empirical insights to guide the development of increasingly capable yet computationally efficient creative content generation systems.

By bridging theoretical creativity assessment with practical implementation considerations, this work contributes to the broader goal of developing AI systems that can serve as effective creative collaborators, augmenting human capabilities while maintaining computational accessibility. Future advances in this domain hold significant promise for transforming content creation workflows across multiple industries, democratizing access to sophisticated content generation capabilities while preserving the essential creative qualities that engage and inform human audiences.

8. References

[1] See, A., Liu, P. J., & Manning, C. D. (2017). CNN/Daily Mail dataset [Dataset]. Hugging Face. https://huggingface.co/datasets/abisee/cnn_dailymail

[2] Merity, S., Xiong, C., Bradbury, J., & Socher, R. (2016). WikiText-103 dataset [Dataset]. Hugging Face. https://huggingface.co/datasets/Salesforce/wikitext

[3] Zhang, X., Zhao, J., & LeCun, Y. (2015). Yelp reviews full dataset [Dataset]. Hugging Face. https://huggingface.co/datasets/Yelp/yelp_review_full

[4] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding [Dataset]. Hugging Face. https://huggingface.co/datasets/nyu-mll/glue

[5] Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). BookCorpus dataset [Dataset]. Hugging Face. https://huggingface.co/datasets/bookcorpus/bookcorpus

[6] Klein, T., & Nabi, M. (2019). Learning to Answer by Learning to Ask: Getting the Best of GPT-2 and BERT Worlds. *arXiv (Cornell University)*. https://doi.org/10.48550/arXiv.1911.02365

[7] Deep Learning Insider. (2023, July 31). How does The Attention Mechanism in GPT Models Work? *Medium*. https://medium.com/@AIExplainedML/how-does-the-attention-mechanism-in-gpt-models-work-5f489a59346b

[8] Wu, H. (2024, September 16). GPT-2 Detailed Model Architecture. *Medium*. https://medium.com/@hsinhungw/gpt-2-detailed-model-architecture-6b1aad33d16b

[9] Gao et al. (2020). The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *arXiv (Cornell University)*. https://arxiv.org/abs/2101.00027

[10] Sharma, D. (2023, May 27). Prompt Generator for Stable Diffusion Model using GPT-NEO and GPT-2. *Medium*. https://medium.com/@driknowsnothing/prompt-generator-for-stable-diffusion-model-using-gpt-neo-and-gpt-2-55d844ccebde

[11] Eldan, R., & Li, Y. (2023). TinyStories: How small can language models be and still speak coherent English? *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2305.07759

[12] Zhang et al. (2022). OPT: Open Pre-trained Transformer Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2205.01068

[13] Zhao et al. (2024, 23 January). Assessing and understanding creativity in large language models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2401.12491

[14] Kann, K., Rothe, S., & Filippova, K. (2018). Sentence-Level fluency Evaluation: References help, but can be spared! *arXiv (Cornell University).* https://doi.org/10.48550/arxiv.1809.08731

[15] Feng et al. (2010, 23 August). A comparison of features for automatic readability assessment. *23rd International Conference on Computational Linguistics (COLING 2010),* Poster Volume, 276-284. https://aclanthology.org/C10-2032.pdf

[16] McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

[17] Ragheb, M., & Dickinson, M. (2012). Defining Syntax for Learner Language Annotation. *Proceedings of COLING 2012: Technical Papers*, 965–974, Mumbai, India. The COLING 2012 Organizing Committee. https://aclanthology.org/C12-2094/

[18] Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

[19] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 3982–3992. 10.18653/v1/D19-1410

[20] Barzilay, R., & Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics, 34*(1), 1–34. https://doi.org/10.1162/coli.2008.34.1.1

[21] Gatt, A., & Krahmer, E. (2018). Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research (JAIR), 61*, 65-170. https://doi.org/10.48550/arxiv.1703.09902

[22] Ethayarajh, K. (2019). How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 55–65, Hong Kong, China. Association for Computational Linguistics. 10.18653/v1/D19-1006

[23] McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated Evaluation of Text and Discourse with Coh Metrix*. Cambridge University Press. https://doi.org/10.1017/CBO9780511894664

[24] Biber, D., & Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes*, *9*(1), 2–20. https://doi.org/10.1016/j.jeap.2010.01.001

[25] Nekamiche, N. (2023, May 13). Curriculum Learning. *Medium*. https://medium.com/aiguys/curriculum-learning-83b1b2221f33

[26] Vaswani et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008. https://doi.org/10.48550/arXiv.1706.03762

[27] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. OpenAI. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[28] Krause, B., Kahembwe, E., Murray, I., & Renals, S. (2019). Dynamic evaluation of transformer language models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1904.08378

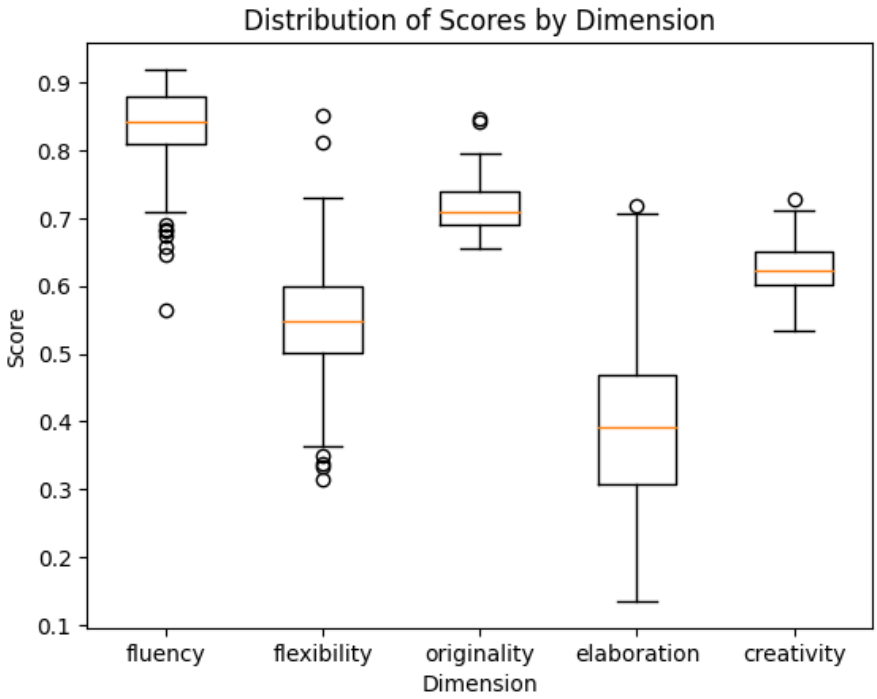
[29] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., & Chen, W. (2021). LORA: Low-Rank adaptation of Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2106.09685

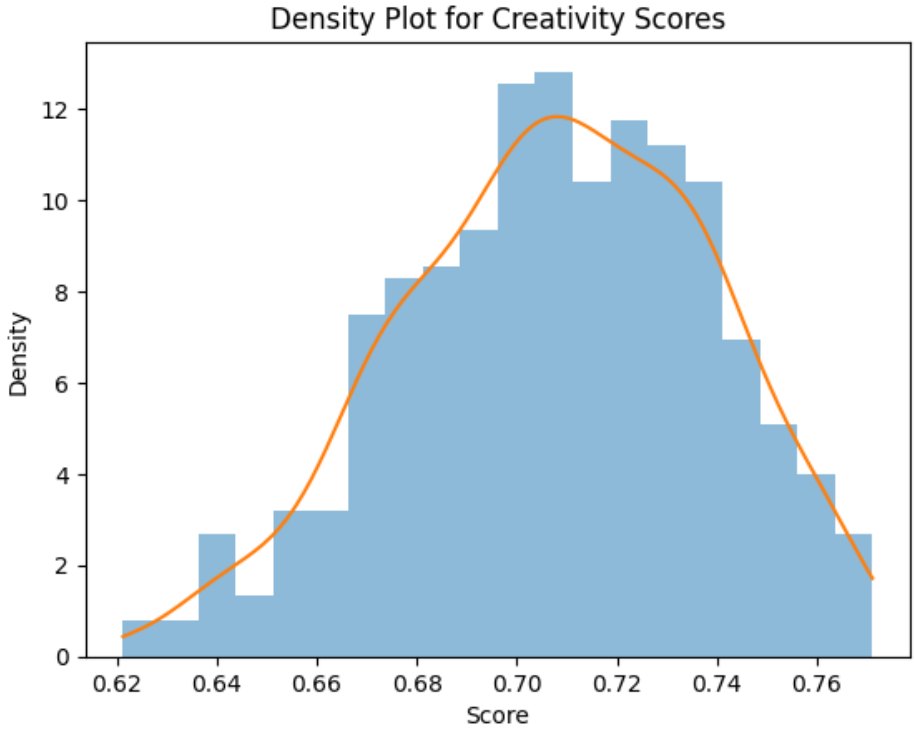
[30] Karpathy, A., Johnson, J., & Fei-Fei, L. (2015). Visualizing and understanding recurrent networks. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1506.02078

9. Appendix A

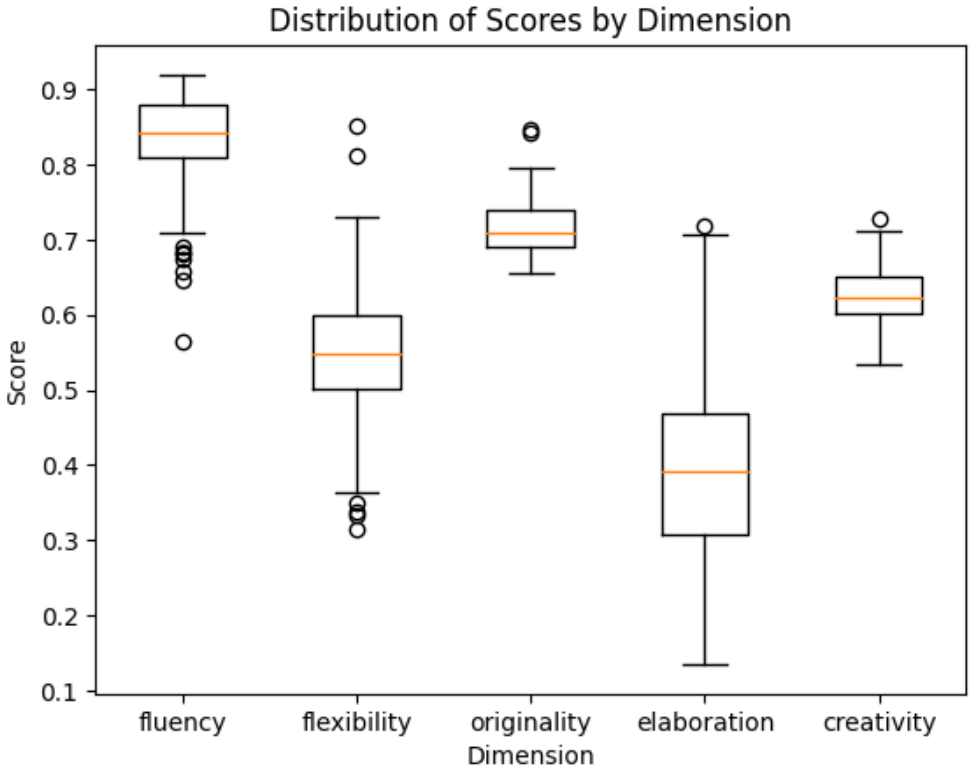
Appendix A contains statistical charts showing the distribution of scores by dimension and density plot for creativity scores for each model.

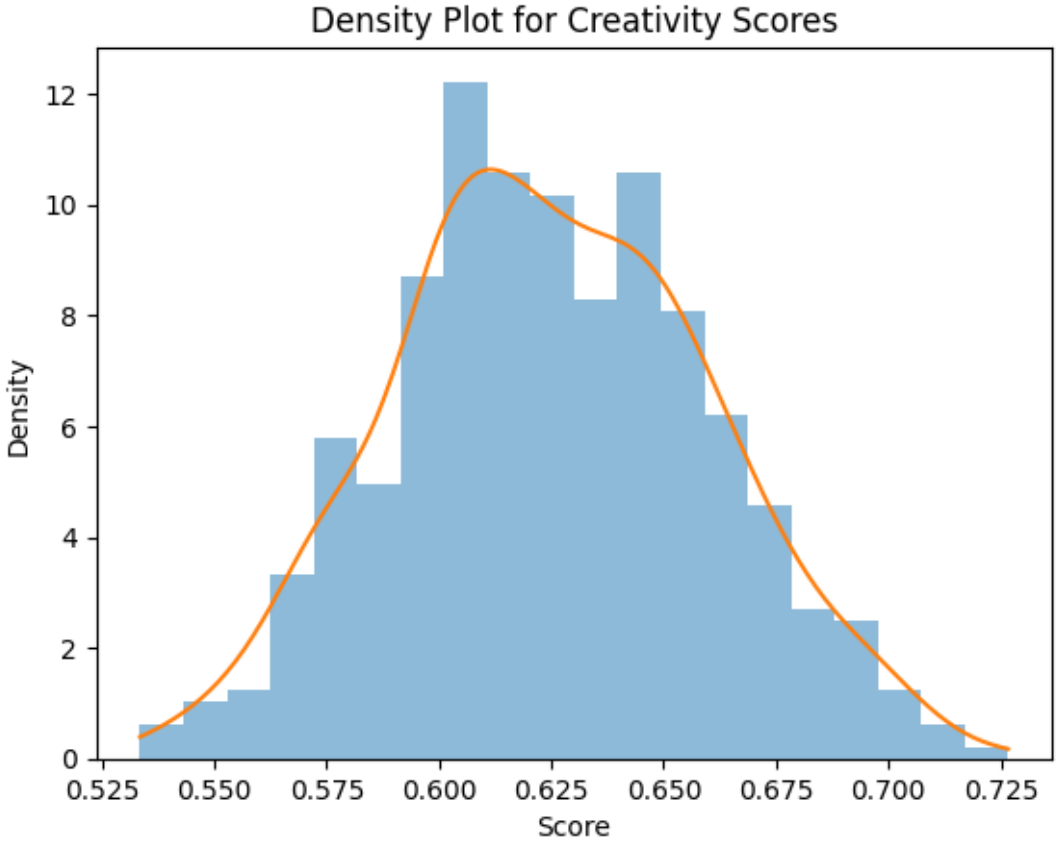
GPT-2



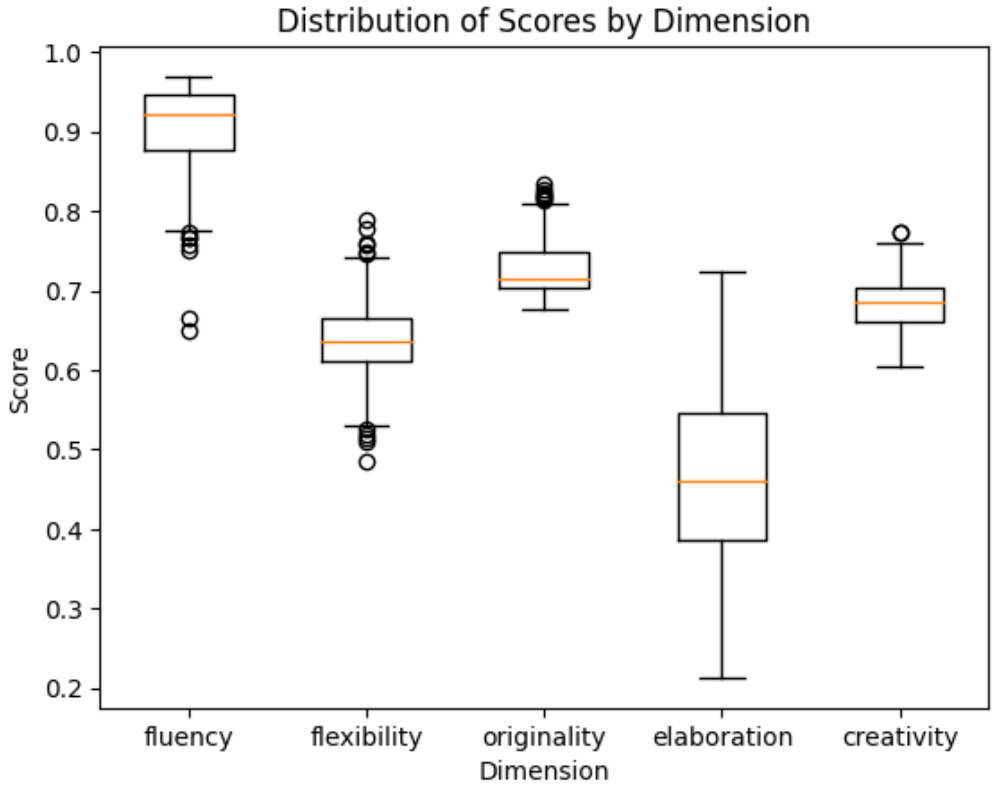


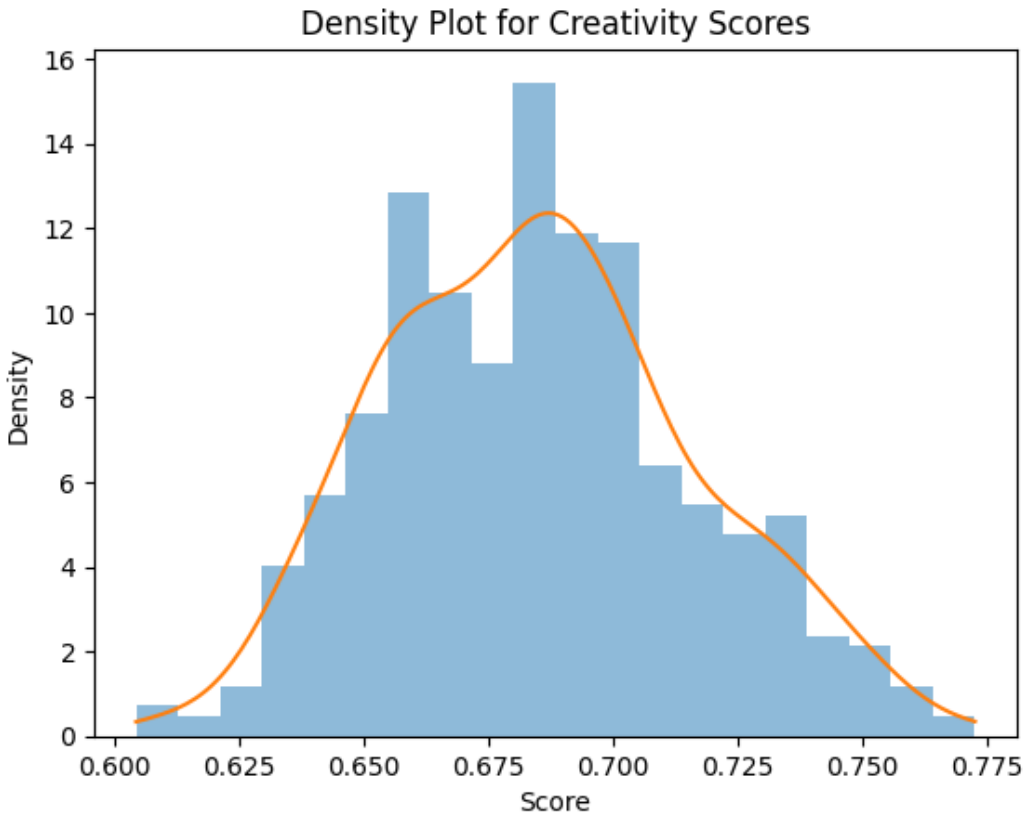
GPT-Neo


Distribution of Scores by Dimension
Score
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
fluency
flexibility
originality
elaboration
creativity
Dimension


Density Plot for Creativity Scores
Density
12
10
8
6
4
2
0
0.525
0.550
0.575
0.600
0.625
0.650
0.675
0.700
0.725
Score

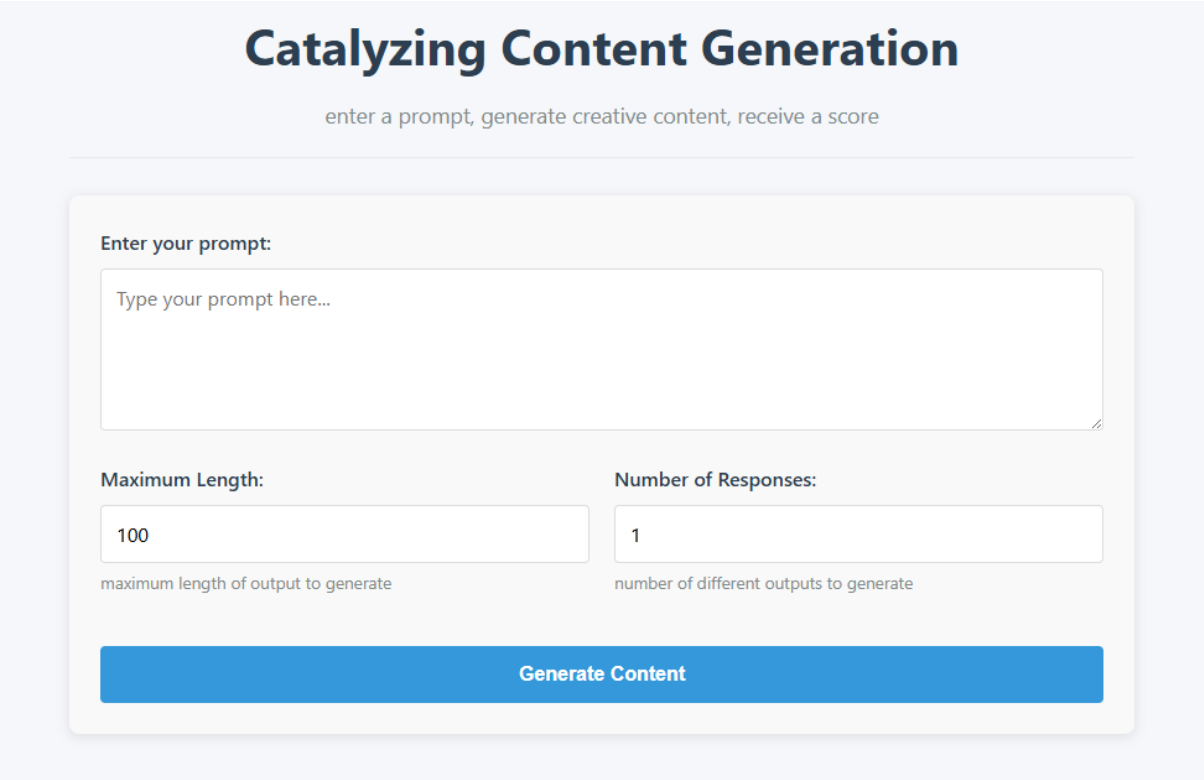
OPT-125


Distribution of Scores by Dimension
Score
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
fluency
flexibility
originality
elaboration
creativity
Dimension


Density Plot for Creativity Scores
Density
16
14
12
10
8
6
4
2
0
0.600
0.625
0.650
0.675
0.700
0.725
0.750
0.775
Score

10. Appendix B

Appendix B showcases the Web Application Interface.



Enter your prompt:

The future of machine learning looks

Maximum Length:

100

maximum length of output to generate

Number of Responses:

1

number of different outputs to generate

Generate Content

Generated Content:

The future of machine learning looks bleak for the next few years, but this week it seems that a new technology is poised to transform our lives. In just over two years researchers have created a system which allows machines with different personalities and abilities such as humans or robots - can be trained to solve complex problems using data from the Internet in real time without having human intervention .

Creativity Evaluation

Fluency:	97.4%	Flexibility:	58.2%
Originality:	69.5%	Elaboration:	36.9%

Overall Creativity: 65.5%

Enter your prompt:

The threat of global warming

Maximum Length:

100

maximum length of output to generate

Number of Responses:

1

number of different outputs to generate

Generate Content

Generated Content:

The threat of global warming has been the subject of much debate, but a new study suggests that it is not only natural hazards to humans - but also climate change itself. The researchers say they believe that if we continue on the path of increasing greenhouse gas emissions and other human-caused risks from the coming decades which they call "carbon dioxide" then future generations will face an even more dire challenge.

Creativity Evaluation

Fluency:	96.1%	Flexibility:	60.2%
Originality:	65.8%	Elaboration:	44.1%

Overall Creativity: 66.6%