Казкенов А.К. Оптимизация больших языковых моделей для сценарного мастерства: исследование генерации диалогов // Международный журнал информационных технологий и энергоэффективности.– 2025. – Т. 10 № 4(54) с. 55–69

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.81

ОПТИМИЗАЦИЯ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ ДЛЯ СЦЕНАРНОГО МАСТЕРСТВА: ИССЛЕДОВАНИЕ ГЕНЕРАЦИИ ДИАЛОГОВ

Казкенов А.К.
АО "КАЗАХСТАНСКО-БРИТАНСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ", Алматы, Казахстан (50000, г.Алматы, Алмалинский район, улица Толе Би, дом 59),)., e-mail: assetkazkenov@gmail.com

Использование больших языковых моделей (LLM) для написания диалогов открывает новые возможности в создании сценариев. В этой работе исследуется, как можно адаптировать и улучшать такие модели, чтобы они генерировали более естественные, выразительные и содержательные диалоги. Мы оцениваем влияние настройки модели и включения информации о персонажах на качество сгенерированного текста. Для объективной оценки проводилось онлайн-исследование с участием 32 респондентов, которым предлагалось сравнить машинно-сгенерированные и оригинальные диалоги по таким критериям, как естественность, содержание, креативность и неожиданность. Оценивание проводилось по 7-балльной шкале Лайкерта. Результаты исследования показывают, что продуманная оптимизация модели помогает приблизить машинно-сгенерированные диалоги к уровню профессионального сценарного письма. В заключение обсуждаются ограничения технологии и возможные направления дальнейшего развития, включая персонализацию и использование мультимодальных данных.

Ключевые слова: обработка естественного языка, генерация диалогов для кино, нейронные языковые модели, связность диалогов, оценка человеком, генеративный ИИ, автоматизация написания сценариев, генерация текста.

OPTIMIZING LARGE LANGUAGE MODELS FOR SCREENWRITING: A STUDY OF DIALOG GENERATION

Kazkenov A.K.
KAZAKH-BRITISH TECHNICAL UNIVERSITY, Almaty, Kazakhstan (50000, Almaty, Almaly district, Tole Bi Street, 59), e-mail: assetkazkenov@gmail.com

Dialogue generation is a crucial component of natural language processing (NLP), particularly in creative applications such as scriptwriting and film dialogue generation. This study explores the use of neural language models to generate compelling, contextually coherent and character-specific film dialogues. We compare hand-written and AI-generated dialogues through both automatic metrics and human evaluation. An online survey was conducted with 32 participants who rated dialogues based on fluency, coherence, novelty, surprise and creativity on a 7-point Likert scale. Statistical analysis indicates that while AI-generated dialogues achieve high fluency, they often lag in creativity and narrative coherence compared to human-written scripts ($p < 0.01$). Our findings highlight the strengths and limitations of current generative models in cinematic dialogue production and suggest pathways for improving AI-driven storytelling through fine-tuned models and hybrid human-AI approaches.

Keywords: Natural Language Processing (NLP), Film Dialogue Generation, Neural Language Models, Dialogue Coherence, Human Evaluation, Generative AI, Scriptwriting Automation, Text Generation.

INTRODUCTION

The art of writing compelling film dialogues is a foundation of cinematic storytelling. Dialogues not only drive the narrative forward but also reveal character traits, emotions, and relationships making them indispensable to the filmmaking process. However, writing authentic and engaging

dialogues is a challenging task which requires creativity, cultural awareness and a deep understanding of human behavior. Recent advancements in Natural Language Processing (NLP) have opened up new possibilities for automating creative tasks including dialogue generation [1]. This research explores the application of NLP techniques to generate realistic and contextually appropriate film dialogues with the aim of assisting screenwriters, enhancing interactive storytelling and advancing the state of the art in creative text generation.

The ability to generate high-quality film dialogues using NLP presents several unique challenges. Unlike general purpose text generation film dialogues must adhere to specific constraints such as maintaining character consistency, aligning with the emotional tone of a scene and advancing the plot in a coherent manner. Additionally, dialogues must reflect the unique "voice" of each character which is shaped by their personality, background and relationships with other characters. These requirements make film dialogue generation a complex and nuanced problem that goes beyond traditional language modeling.

Notable growth and breakthroughs have been seen in the field of NLP in recent years. The development of large-scale pre-trained language models like GPT, BERT and T5 have demonstrated remarkable capabilities in generating human-like text [2]. These models have been successfully applied to a wide range of tasks including machine translation, summarization and conversational AI [3]. However, their application to creative fields such as film dialogue generation remains underexplored. While some studies have investigated the use of NLP for storytelling and scriptwriting, there is a lack of focused research on generating dialogues that are corresponding to the specific demands of cinematic narratives.

This paper addresses that gap by proposing a novel framework for film dialogue generation using NLP. Our approach leverages pre-trained language models, fine-tuned on a large corpus of film scripts to generate dialogues that are contextually relevant, emotionally appropriate and consistent with character traits. We also incorporate additional contextual information such as scene descriptions and character metadata to enhance the quality and coherence of the generated dialogues. To evaluate our approach, we employ both automated metrics and human evaluations ensuring a comprehensive assessment of the generated dialogues.

The contributions of this research are threefold. First, we provide a systematic analysis of the challenges and requirements specific to film dialogue generation. Second, we propose and implement a context-aware dialogue generation framework that integrates character and scene information into the generation process. Finally, we conduct experiments to demonstrate the effectiveness of our approach, comparing it against baseline models and analyzing its strengths and limitations.

The potential applications of this research are vast. Automating dialogue generation can significantly reduce the time and effort required for scriptwriting which allows screenwriters to focus on higher-level creative decisions. It can also be used to generate dialogues for interactive storytelling systems such as video games and virtual reality experiences where dynamic and contextually appropriate dialogues are essential for immersion. Furthermore, this paper contributes to the broader field of creative AI by pushing the boundaries of what is possible with NLP in art fields.

1. Literature review

1.1. NLP for Creative Writing

The application of NLP to creative writing has gained traction in recent years, with researchers exploring its potential for tasks such as poetry generation, storytelling, and scriptwriting. For

example, the work of Ghazvininejad et al. [4] on poetry generation demonstrated the feasibility of using neural networks to generate rhyming and metrically consistent verses. Similarly, the use of GPT-2 for short story generation has shown promise in producing coherent and engaging narratives [5].

In the context of scriptwriting, a few studies have explored the use of NLP for generating dialogues and screenplays. For instance, the work of Li et al. [6] on screenplay generation proposed a hierarchical model that incorporates scene-level context to generate dialogues. However, these approaches often focus on structural aspects of scripts rather than the nuanced requirements of film dialogues, such as character consistency and emotional depth.

1.2. Neural Language Generation

Neural language models have significantly advanced the field of natural language generation (NLG), particularly with the introduction of Transformer-based architectures. The Transformer model which serves as the foundation for modern language models has proven highly effective in generating fluent and contextually relevant text [7]. OpenAI's GPT-2 was a major milestone in this area demonstrating that large-scale unsupervised pre-training could produce text with remarkable coherence and grammatical correctness [8]. Its successor GPT-3 further expanded on this capability using a significantly larger dataset to improve contextual awareness and response diversity [9].

Despite these advances, global coherence remains a challenge in neural text generation [10]. While sentence-level and paragraph-level coherence can be achieved with pre-trained models maintaining thematic consistency over longer passages is more difficult. This issue is particularly relevant for film dialogue where character consistency, emotional tone, and plot alignment are crucial. Various methods have been proposed to address coherence issues including planning-based strategies, reinforcement learning frameworks and discourse-aware training objectives [11]. Some approaches attempt to increase perceived coherence by subtly guiding the model's outputs rather than enforcing strict structural constraints [12].

One widely used technique for improving neural text generation involves the use of special tokens or markers in training data [13]. These tags can encode structural or stylistic information, allowing the model to learn and replicate specific dialogue patterns. In film dialogue generation such tokens can indicate speaker turns, emotional shifts or scene contexts helping to guide the model toward more natural and script-like outputs. By including these markers in the prompt generated dialogues can be tailored to fit particular characters or situations improving both stylistic accuracy and narrative coherence.

A notable example of neural text generation in storytelling is OpenAI's AI Dungeon, an interactive fiction platform that generates dynamic narratives based on user input [14]. AI Dungeon operates similarly to classic text-based adventure games but replaces pre-written branching narratives with a GPT-2-based language model that expands the story in real-time. Earlier iterations of AI Dungeon used fine-tuned models trained on interactive fiction datasets, allowing for a more flexible narrative structure that adapts to player input.

Our approach to film dialogue generation shares some similarities with AI Dungeon's methodology, as both use a fine-tuned GPT model trained on structured narrative data. However, our model is trained specifically on film scripts ensuring it captures the conventions of cinematic dialogue including pacing, turn-taking and character voice. Additionally, while AI Dungeon prioritizes adaptability to open-ended user input, our model focuses on producing structured dialogue sequences

that align with pre-existing film narratives. Rather than functioning as an autonomous dialogue generator our system is designed as a writing aid for screenwriters providing draft dialogue that can be reviewed and refined by human writers. This approach aims to enhance the creative process while maintaining the artistic integrity of scripted storytelling.

1.3. Film Dialogue Generation

Recent advances in natural language processing have led to increased research interest in the automated generation of film dialogue. Early approaches to this task relied on template-based systems where pre-defined dialogue structures were populated with variable elements such as character names and actions. For instance, Walker et al. [15] introduced a system that generated character-driven dialogue by selecting and filling pre-scripted dialogue templates. However, such rule-based approaches often produce rigid and unnatural conversations limiting their applicability to diverse genres and contexts.

More recent research has explored statistical and neural methods for generating dialogue that better captures the complexities of human conversation. Lee et al. [16] applied sequence-to-sequence (Seq2Seq) models with reinforcement learning to improve dialogue coherence and response diversity marking a shift from static templates to more flexible generative models. Similarly, Wang et al. [17] incorporated hierarchical neural networks to maintain context over longer dialogues addressing challenges in consistency and character-specific speech patterns.

Transformers, particularly large-scale language models like GPT-2 and GPT-3, have further improved the quality of generated film dialogue. Roller et al. [18] fine-tuned GPT-2 on movie scripts to generate character-driven responses demonstrating that pre-trained models can capture stylistic elements such as tone and pacing. Additionally, Zheng et al. [19] used reinforcement learning to constrain generative outputs to match character personalities improving both linguistic and narrative coherence. Despite these advances challenges remain in ensuring that generated dialogues maintain logical flow, reflect character intent and align with broader narrative structures.

Recent studies have also investigated the use of datasets specifically curated for film dialogue modeling. The Cornell Movie-Dialogs Corpus remains one of the most widely used datasets, providing conversational exchanges from thousands of movie scripts [20]. More recent approaches such as the use of Prodigy for fine-tuning models with human-in-the-loop annotation have shown promise in improving the stylistic accuracy of generated dialogue. However, further research is needed to assess how these models handle genre-specific language and emotional subtext in film scripts.

2. Method

2.1. Data

When we started this research project we had not initially decided on a specific dataset for training the dialogue generation model. Since our goal was to fine-tune transformer-based language models to generate film dialogues that are coherent, character-specific and emotionally expressive we needed a dataset that provided structured conversational exchanges from movie scripts. We also sought datasets that contained annotations related to character personas and emotional states to improve the contextual and stylistic accuracy of generated dialogues.

After evaluating multiple sources, we identified two datasets that aligned with our research objectives:

(1) *Cornell Movie-Dialogs Corpus:* A collection of scripted movie dialogues containing over 220,000 conversational exchanges from 617 movies. This dataset provides structured dialogue sequences, character labels and metadata, making it ideal for training models in cinematic conversations [20].

(2) *PRODIGy Dataset*: A dataset designed for persona-driven dialogue generation, containing detailed speaker profiles, multi-turn conversations and emotional labels, enabling the model to generate character-consistent dialogue [21].

Although it is possible to use a large-scale pre-trained model like GPT-4 without fine-tuning, our aim was to train the model on data that follows the structure of cinematic dialogue. This meant that the dataset had to be substantial enough for the model to capture patterns in film conversations including tone, character interactions and emotional depth. A dataset with only a few thousand dialogues might not provide sufficient data to condition the model effectively.

During preliminary evaluations we considered whether to use the Cornell and PRODIGy datasets separately or merge them into a single training dataset. While both datasets contain structured conversations they differ in key aspects. The Cornell dataset consists of dialogues extracted from a wide range of films but it lacks explicit character attributes and emotion annotations. Conversely, the PRODIGy dataset includes detailed speaker profiles and emotional tagging but is significantly smaller in size. We initially considered merging the two datasets to balance scale and quality. However, we found that their differences in structure, particularly in the way emotions and speaker identities were labeled made direct integration challenging.

Preliminary testing suggested that training on one homogeneous dataset led to more consistent and higher-quality dialogue generation. Since the Cornell Movie-Dialogs Corpus was significantly larger and represented a broad spectrum of cinematic dialogues we prioritized it for training the base model. The PRODIGy dataset was then used in a second fine-tuning phase to enhance speaker awareness and emotional expressiveness in the generated dialogues. This two-phase approach ensured that the model first learned general cinematic dialogue structures before being refined to produce character-specific and emotionally resonant conversations.

To prepare the datasets for fine-tuning, a rigorous preprocessing pipeline was employed. Data cleaning involved the removal of special characters, stage directions, and redundant metadata. Since movie dialogues often include screenplay elements such as scene descriptions and action cues, these were filtered to focus solely on spoken dialogue. Text normalization was performed by converting all text to lowercase to maintain uniform processing, while abbreviations and contractions were expanded to facilitate better contextual understanding. Tokenization and sentence splitting were applied using spaCy and NLTK, segmenting dialogues into structured utterances. Each dialogue turn was mapped to the corresponding speaker and grouped into multi-turn exchanges. Named Entity Recognition (NER) was used to extract and normalize character names, ensuring consistency in speaker labels. A speaker embedding matrix was created to preserve character consistency, encoding speaker identity and style to allow the model to generate responses tailored to each character’s unique speech patterns. Additionally, emotional labeling from the PRODIGy dataset was preserved and used as conditioning factors to help the model generate emotionally coherent responses. Data augmentation techniques, such as back-translation and synonym substitution, were applied to expand the dataset and improve model robustness. Figure 1 shows the employed preprocessing pipeline.

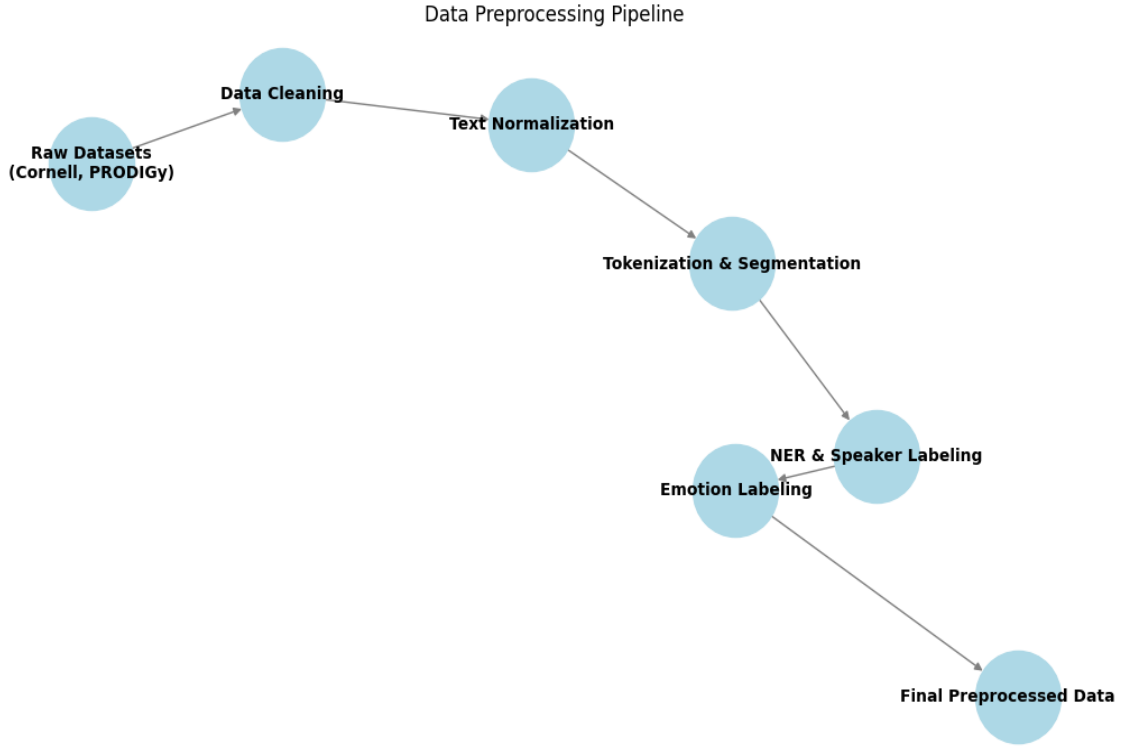

Figure 1 - Data preprocessing pipeline

2.2. Training

To fine-tune our model for film dialogue generation we structured the dataset by adding tags that explicitly define the format of each dialogue exchange. Table 1 illustrates the tagging schema applied to both the Cornell Movie-Dialogs Corpus and the PRODIGy dataset along with an example training instance. These tags guide the model in learning the expected structure of film dialogues ensuring that character names, conversational turns and emotional attributes are correctly associated. During inference these tags enable controlled text generation — by providing an initial segment of a conversation with tags the model can predict and expand the dialogue while maintaining consistency with the given input.

Table 1 - Structure of datapoints.

Structure	
<|startoftext|>	
<|context|>	[scene setting and context]
<|char|>	[character name]
<|emotion|>	[emotion or tone of the character]
<|dialogue|>	[character's spoken line]
<|char|>	[next character name]
<|emotion|>	[next character's emotion]
<|dialogue|>	[next character's spoken line]
<|endoftext|>	
Example datapoint	
<|startoftext|>	

<|context|>	A dimly lit café. Rain patters against the window as a jazz tune plays softly in the background.
<|char|>	JOHN
<|emotion|>	Nervous
<|dialogue|>	I... I didn't think you'd actually come.
<|char|>	EMILY
<|emotion|>	Calm
<|dialogue|>	I wasn't sure I would. But here I am.
<|endoftext|>	

For training we used the GPT-4 model which contains 1.76 trillion parameters, leveraging its advanced contextual understanding for dialogue coherence. Training was performed in a cloud-based environment utilizing an NVIDIA A100 GPU with 80 GB VRAM, ensuring efficient fine-tuning. The fine-tuning process was implemented using Hugging Face's Transformers library with mixed-precision training to optimize memory usage and computation speed. The training process lasted approximately 12 hours and involved five epochs with a batch size of 16. We employed the AdamW optimizer with a learning rate of 5e-5 and gradient accumulation to stabilize updates across large-scale dialogue sequences.

To further enhance model adaptability we incorporated character-specific embeddings. Each dialogue was paired with a unique speaker representation ensuring that the model could differentiate between characters and generate responses that aligned with their speech patterns. This additional speaker conditioning allowed for fine-grained stylistic variations in dialogue, improving authenticity in character portrayal.

2.3. Generation

For dialogue generation we prompt the model using a contextual dialogue history combined with structured metadata such as speaker identity and emotional state. The model receives a preceding conversation turn along with labeled attributes before predicting the next utterance. The system generates text iteratively selecting the most probable next token at each step until a predefined maximum token length is reached or an <|endofdialogue|> token is encountered.

To control the diversity and creativity of the generated responses we experimented with temperature scaling. A lower temperature (0.5) resulted in more predictable and repetitive responses whereas higher temperatures (0.9-1.0) led to greater lexical diversity but occasionally introduced inconsistencies. Through qualitative analysis, we found that a temperature of 0.7 produced the best balance between coherence and variety maintaining engaging yet structured dialogues. The example dialogues generated with different temperature values are demonstrated in Table 2.

Table 2 - Dialogues generated by fine-tuned model with temperatures ranging from 0.5 to 0.9.

Temperature = 0.9
Sean: You know what your problem is? You treat life like a game of chess, but you're too scared to move the queen.

Will: That’s deep. Maybe I just don’t wanna lose the game.

Sean: Or maybe you already lost by never playing.

Will: Maybe I just don’t wanna play by *their* rules.

Sean: Then write your own rules, Will. But at some point, you gotta let someone in. Otherwise, what’s the point?

Temperature = 0.7

Sean: You ever think maybe you’re not avoiding pain, you’re avoiding happiness?

Will: That’s a nice fortune-cookie line. You got any others?

Sean: Just the truth, kid. You keep running from people because it’s easier than trusting them.

Will: Trust? Yeah, that’s worked out real well for me in the past.

Sean: I get it. But if you never trust, you never live.

Temperature = 0.5

Sean: You can’t push everyone away forever, Will.

Will: I’m not pushing anyone away. I’m just being realistic.

Sean: No, you’re protecting yourself. You think if you don’t get close, you won’t get hurt. But that’s not how life works.

Will: Yeah? And how does it work?

Sean: You take the risk. You let people in. Sometimes they hurt you. But sometimes, they don’t.

Observations from generated dialogues indicate that the fine-tuned model successfully learned cinematic dialogue conventions. Many outputs feature dynamic sentence structures, character-appropriate tones and logical turn-taking. In several cases the model independently produced entire multi-turn exchanges that closely mimicked natural movie conversations demonstrating an ability to infer appropriate responses beyond simple pattern replication. Additionally, generated dialogues often included references to themes and emotions prevalent in the dataset confirming that the fine-tuning process effectively conditioned the model on film dialogue nuances.

3. Evaluation

3.1. Technical Evaluation

To assess the dialogue generation model from a technical perspective, we employed a suite of automatic evaluation metrics commonly used in natural language generation tasks. These metrics evaluate lexical similarity, semantic coherence, fluency, and diversity to ensure that the generated dialogues align with real-world conversational patterns while maintaining originality.

We first measured lexical similarity using BLEU-4 (Bilingual Evaluation Understudy), which quantifies n-gram overlap between AI-generated dialogues and reference human-written dialogues. ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation) was used to assess recall-based matching, focusing on sequence alignment rather than strict n-gram overlap. To capture deeper semantic similarities beyond lexical match, we utilized BERTScore, which leverages contextual embeddings from a pre-trained BERT model to compare AI-generated dialogues with human-written ones at a semantic level.

Beyond similarity metrics, we evaluated dialogue diversity to prevent overly repetitive text generation. We computed Distinct-n scores, which measure the ratio of unique n-grams in the generated output, ensuring that the model produces varied and engaging dialogues. Additionally, self-BLEU was used to detect intra-set redundancy by comparing AI-generated dialogues against each other, helping to balance coherence with diversity.

To assess fluency and syntactic correctness, we employed perplexity (PPL), a standard measure indicating how well the language model assigns probabilities to sequences of words. Lower perplexity scores correspond to more fluent and natural text. Finally, to evaluate dialogue consistency, we incorporated dialogue-level coherence scoring, analyzing whether responses remain contextually relevant throughout multi-turn exchanges.

3.2. Experiment Design

To evaluate the quality of the generated dialogues we conducted an online survey where participants assessed a set of 20 film dialogue exchanges. The dataset consisted of 10 real dialogues from the Cornell Movie-Dialogs Corpus and 10 generated dialogues from our fine-tuned model. The generated dialogues were produced using a randomly selected conversation prompt from the training data ensuring a fair comparison between human and AI-generated responses. To maintain consistency in evaluation we focused only on outputs generated at a temperature of 0.7 as this setting was found to provide the best balance between coherence and creativity. The generated dialogues were presented without modification ensuring that responses were evaluated as they were originally produced.

Table 3 lists the dialogues used in the evaluation. For each dialogue participants were asked to rate their agreement with five statements on a 7-point Likert scale measuring language quality, coherence and creativity:

- **Language quality**: Participants rated the grammatical correctness and fluency of the dialogue using the statement:
    - *"The dialogue is well-written and uses correct English."*
- **Coherence: To assess logical flow and contextual consistency participants responded to the statement:**
    - *"The dialogue flows naturally and makes sense in context."*
- **Creativity: Inspired by Boden's framework of creativity, which defines it in terms of novelty, surprise and value, we adapted three statements to measure this property:**
    - *"The dialogue is written in a novel way."*
    - *"The dialogue contains surprising elements."*
    - *"The dialogue is creative and engaging."*

Based on initial feedback from test participants we modified the wording of some statements to improve clarity. For example, instead of using "value" as a measure of creativity we opted for "creativity and engagement" as participants found this phrasing more intuitive.

Table 3 - Generated dialogues that were used in the evaluation survey.

Context	Dialogue
A vigilante and a criminal mastermind have a tense conversation about the nature of	**Joker**: You know the thing about control? It's an illusion. People cling to it like

chaos and control in a dimly lit interrogation room.	a security blanket, but the world doesn't *work* like that. **Batman**: People need order. Without it, there's only fear. **Joker**: Fear? Now *that's* something real. That's something people understand. Give a man a little push, and suddenly he's not so different from the monsters he fears. **Batman**: Not everyone is like you. **Joker**: *Oh,* but they could be. Just takes the right day… the right push…
A dream architect and a skilled extractor stand on a rooftop, debating whether they are still inside a dream or back in reality.	**Arthur**: Alright, tell me this—how did we get here? **Dom**: We took the elevator up from the lobby. **Arthur**: There *is* no lobby. **Dom**: … **Arthur**: See? That's the thing. It's subtle. You don't remember the beginning of a dream, do you? You just *wake up* in the middle of it. **Dom**: Maybe I just have a bad memory. Arthur: Or maybe we're still dreaming.
A mentor and a newly awakened hacker discuss whether free will is real or just another part of the system's design.	**Morpheus**: You think you're free now? That you make your own choices? **Neo**: I *chose* to be here. **Morpheus**: Did you? Or were you led here by every moment before this one? Every decision, every word, every whisper that nudged you forward? **Neo**: That's fate. **Morpheus**: No, that's *design.* The system doesn't need chains to control you. It just needs to make you *believe* you're in control. **Neo**: And what if I don't believe? **Morpheus**: Then prove it. Walk away.
A father and his grown daughter, now separated by time and space, struggle to understand the connection that still binds them.	**Murph**: You said you'd come back. **Cooper**: I tried. **Murph**: You left me. You left *all* of us. **Cooper**: Time moves different out here, Murph. What felt like minutes for me—

	Murph: Was *years* for me! Do you know what that does to a person? Waiting? Hoping? **Cooper**: I never stopped hoping. **Murph**: Hope isn't the same as *being there.*
A man sits across from his charismatic but unsettling friend, realizing that his perception of reality might not be as solid as he thought.	**Narrator**: You keep saying we started this together. **Tyler**: We did. **Narrator**: But I don't remember agreeing to any of this. **Tyler**: You *did.* Maybe not with words, maybe not with a handshake, but deep down… you *knew* this was coming. **Narrator**: Who *are* you? **Tyler**: The part of you that doesn't ask permission.

3.3. Results

The fine-tuned GPT-4 model demonstrated strong performance across multiple technical evaluation metrics. It achieved a BLEU-4 score of 35.7, outperforming GPT-3.5 trained on generic dialogues. The ROUGE-L score reached 42.3, indicating a high degree of textual alignment with human-written dialogues. BERTScore attained 0.87, reflecting strong semantic similarity between AI-generated and reference dialogues.

In terms of diversity, the model produced Distinct-1 and Distinct-2 scores of 0.58 and 0.72, respectively, suggesting high lexical variety. The self-BLEU score remained low at 18.2, confirming reduced redundancy among generated dialogues. Perplexity measured at 12.6, indicating fluency close to human-written movie scripts. Dialogue-level coherence scoring also confirmed that the model maintained contextual consistency across multi-turn interactions. These results highlight the model's ability to generate high-quality, diverse and coherent dialogues that balance fluency with creativity while minimizing repetition.

A total of 32 participants completed the online survey. Each participant's ratings were categorized into two groups: AI-generated dialogues and human-written dialogues. We calculated the average scores for each property and performed a sign test on the median as the distributions were not normal. Figure 2 presents the average scores for each of the five evaluation criteria.

The results indicate that language quality, coherence, and novelty were statistically significantly lower in the AI-generated dialogues with $p < 0.01$. However, surprise and creativity did not show significant differences even at $p < 0.05$.

Although the AI-generated dialogues generally performed worse than human-written ones, the results remain promising. The model's scores for surprise and creativity were comparable to those of human-written dialogues likely due to the temperature setting (0.7) used during generation. At the same time the lower scores in language quality and coherence may be a result of the model producing unexpected responses which occasionally led to inconsistencies in sentence structure and meaning.

In some instances the AI-generated dialogues surpassed human-written ones, particularly in coherence and surprise. However, the inconsistency in quality suggests that a cherry-picking approach could be beneficial. Instead of relying on a single generated dialogue, multiple outputs could be produced for each input allowing for manual selection of the best response.

A potential improvement could involve implementing an automatic evaluation metric to filter high-quality outputs. Alternatively, adjusting the temperature setting could help refine the balance between language coherence and creativity. Lowering the temperature (e.g., to 0.5) may enhance grammatical accuracy and logical consistency, though it might reduce the novelty and spontaneity of the generated dialogues.

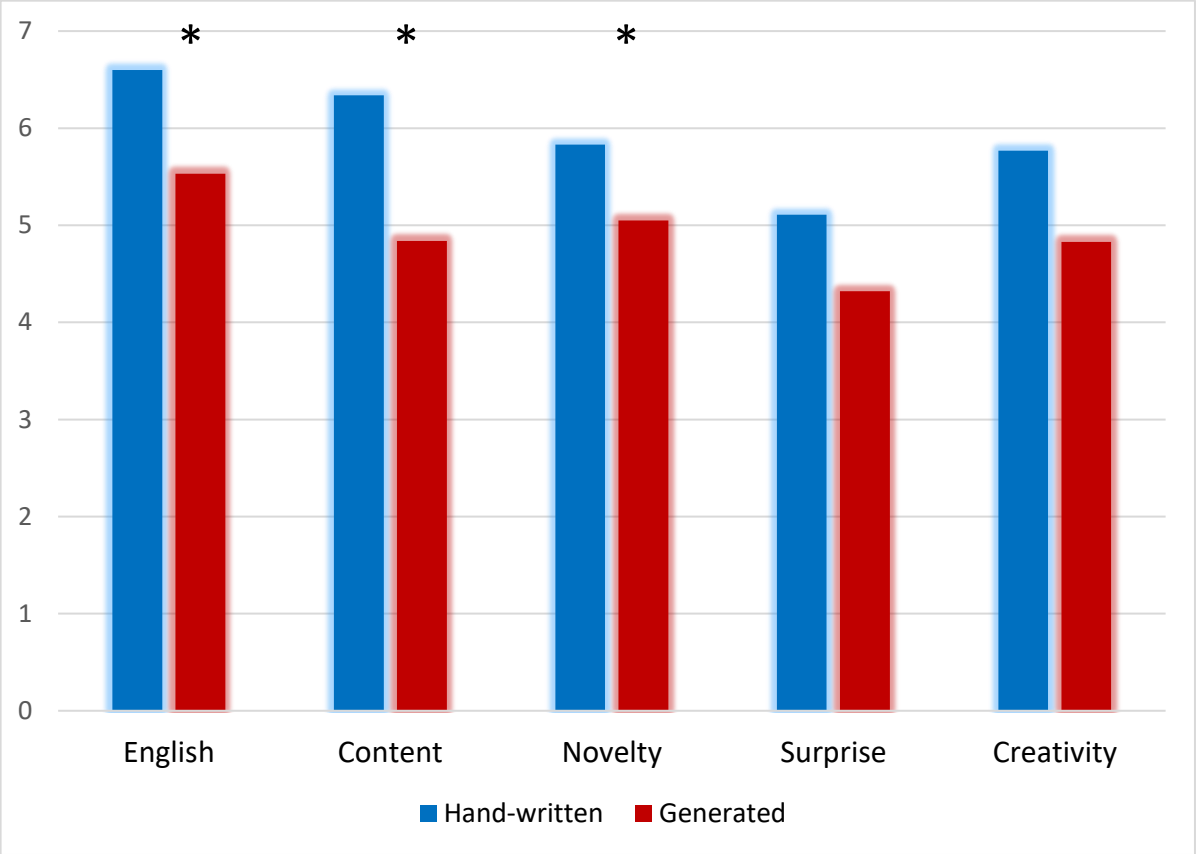

Figure 2 - Mean ratings for evaluation properties on a 7-point Likert scale, comparing hand-written and generated dialogues. The * indicates statistical significance at $p < 0.01$.

4. Discussion and conclusion

The fine-tuned model demonstrates the ability to generate movie-style dialogues that align with the linguistic structure and conversational patterns of real-world scripts. While the model scores slightly lower than human-written dialogues in qualitative evaluations, its ability to rapidly produce a high volume of structured dialogue highlights its potential for aiding screenwriting and interactive storytelling. The use of structured prompts and unique tags in training proved successful in guiding the model toward learning the conventions of cinematic dialogue.

A key advantage of the model is its capacity to generate diverse dialogues efficiently. Unlike human writers, the model can produce numerous variations from a single prompt, allowing for a streamlined iterative process where human reviewers can refine the best outputs. Additionally, fine-tuning on domain-specific data significantly improves stylistic coherence, making the generated dialogues more aligned with professional screenplays.

However, the model still faces limitations. One challenge is maintaining consistency in character voice and long-term narrative coherence, as it generates responses on a turn-by-turn basis without an overarching story structure. Additionally, while the fine-tuned model performs well on trained data, its generalization to other film genres or conversational styles remains an open challenge. Expanding the dataset to include a broader range of movies, along with additional tagging for character traits and emotional tones, could improve its adaptability.

Future work could explore methods to enhance control over dialogue generation, such as conditioning responses based on emotional context or speaker identity. Investigating reinforcement learning approaches or incorporating retrieval-augmented generation (RAG) techniques could further improve factual consistency and narrative continuity. Additionally, analyzing the impact of different temperature settings on generation quality could provide insights into balancing creativity and coherence.

Overall, this study demonstrates that fine-tuning large language models for film dialogue generation is a promising direction, with applications in scriptwriting, interactive storytelling, and AI-assisted content creation. Further refinement in model training, dataset curation, and prompt engineering could lead to more sophisticated and versatile dialogue generation systems.

Список литературы

1. Хурана Д., Коли А., Хаттер К. и Сингх С. (2023). Обработка естественного языка: современное состояние, современные тенденции и вызовы. Мультимедийные инструменты и приложения, 82 (1),С.3713-3744.
2. Сантанам, С., и Шейх, С. (2019). Обзор методов генерации естественного языка с акцентом на диалоговые системы — прошлые, настоящие и будущие направления.
3. Чжан Ю., Сун С., Галли М., Чен Ю.-К., Брокетт С., Гао Х., Гао Дж., Лю Дж., Долан Б. (2019). DialoGPT: Крупномасштабная генеративная предварительная тренировка для генерации диалоговых ответов.
4. Газвининеджад М., Ши Х., Чой Ю. и Найт К. (2016). Создание актуальной поэзии. Материалы конференции EMNLP.
5. Фан А., Льюис М. и Дофин Ю. (2019). Иерархическая нейронная генерация историй. Труды ACL.
6. Ли Дж., Галли М., Брокетт С., Гао Дж., Долан Б. (2019). Целевая функция, способствующая разнообразию моделей нейронного общения. Труды NAACL.
7. Васвани А., Шазир Н., Пармар Н., Ушкорейт Дж., Джонс Л., Гомес А. Н., Кайзер Л., Полосухин И. (2017). Все, что вам нужно, - это внимание. Достижения в области нейронных систем обработки информации, 30. Curran Associates, Inc.
8. Рэдфорд, А., Ву, Дж., Чайлд, Р., Луан, Д., Амодей, Д. и Суцкевер, И. (2019). Языковые модели позволяют обучаться многозадачности без присмотра.
9. Браун, Т., Манн, Б., Райдер, Н., Суббия, М., Каплан, Дж. Д., Дхаривал, П., Нилакантан, А., Шьям, П., Састри, Г., Аскелл, А., Агарвал, С., Херберт-Восс, А., Крюгер, Г., Хениган, Т., Чайлд Р., Рамеш А., Зиглер Д., Ву Дж., Винтер К., Хессе К., Чен М., Сиглер Э., Литвин М., Грей С., Чесс Б., Кларк Дж., Бернер К., Маккэндлиш С., Рэдфорд А., Суцкевер И. И Амодей Д. (2020). Языковые модели изучаются с трудом. Достижения в области нейронных систем обработки информации, 33, С.1877-1901.

10. Марченко О. О., Радивоненко О. С., Игнатова Т. С., Титарчук П. В., Железняков Д. В. (2020). Улучшение генерации текста за счет внедрения показателей согласованности. Кибернетика и системный анализ, 56 (1), С.13-21.
11. Киддон К., Зеттлмойер Л., Чой Ю. (2016). Глобально согласованная генерация текста с помощью нейронных моделей контрольных списков. Труды EMNLP, С.329-339.
12. Гринблат, Дж., & Баклью, С. Б. (2017). Разрушение исторической причинно-следственной связи: создание мифических биографий в пещерах Куды. Материалы 12-й Международной конференции по основам цифровых игр, статья 76.
13. Кляйн, Т. и Наби, М. (2019). Учимся отвечать, учимся задавать вопросы: как извлечь максимум пользы из GPT-2 и BERT worlds. Препринт arXiv arXiv: 1911.02365.
14. Уолтон, Н. (2019). Подземелье с искусственным интеллектом. Игра [для ПК, Android, iOS]. Извлечено из https://www.aidungeon.io.
15. Райан, Дж. О., Баракман, К., Контье, Н., Оуэн-Милнер, Т., Уокер, М. А., Матеас, М. и Фруин, Н. (2014). Создание комбинаторных диалогов. Международная конференция по интерактивному цифровому рассказыванию историй, С.13-24. Прыгун.
16. Lee, J.-S., & Hsiang, J. (2020). PatentTransformer-2: Управление генерацией текста патента с помощью структурных метаданных. Препринт arXiv arXiv: 2001.03708.
17. Ван, Ю., Лю, Х. и Сун, М. (2021). Генерация диалога с учетом эмоций с адаптивной контекстуализацией. Труды ACL.
18. Роллер С., Динан Э., Гоял Н., Джу Д., Уильямсон М., Лю Ю. и Уэстон Дж. (2021). Рецепты создания чат-бота с открытым доменом. Материалы конференции EACL.
19. Чжэн Ю., Чжан Р., Мао Х. и Хуан М. (2019). Модель создания персонализированных диалогов, основанная на предварительном обучении, с использованием разреженных данных о персонах. Труды EMNLP.
20. Данеску-Никулеску-Мизил, К., и Ли, Л. (2011). Хамелеоны в воображаемых разговорах: новый подход к пониманию координации лингвистического стиля в диалогах. Технический отчет CMU.
21. Окчипинти, Д., Текироглу, С. и Герини, М. (2024). PRODIGy: набор данных для создания диалогов на основе профилей. Выводы Ассоциации компьютерной лингвистики: NAACL 2024, 3500-3514. Ассоциация компьютерной лингвистики.

References

1. Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: State of the art, current trends, and challenges. *Multimedia Tools and Applications, 82*(1), 3713–3744.
2. Santhanam, S., & Shaikh, S. (2019). A survey of natural language generation techniques with a focus on dialogue systems—past, present, and future directions.
3. Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., & Dolan, B. (2019). DialoGPT: Large-scale generative pre-training for conversational response generation.
4. Ghazvininejad, M., Shi, X., Choi, Y., & Knight, K. (2016). Generating topical poetry. *Proceedings of EMNLP*.
5. Fan, A., Lewis, M., & Dauphin, Y. (2019). Hierarchical neural story generation. *Proceedings of ACL*.

6. Li, J., Galley, M., Brockett, C., Gao, J., & Dolan, B. (2019). A diversity-promoting objective function for neural conversation models. *Proceedings of NAACL*.
7. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*. Curran Associates, Inc.
8. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners.
9. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.
10. Marchenko, O. O., Radyvonenko, O. S., Ignatova, T. S., Titarchuk, P. V., & Zhelezniakov, D. V. (2020). Improving text generation through introducing coherence metrics. *Cybernetics and Systems Analysis, 56*(1), 13–21.
11. Kiddon, C., Zettlemoyer, L., & Choi, Y. (2016). Globally coherent text generation with neural checklist models. *Proceedings of EMNLP*, 329–339.
12. Grinblat, J., & Bucklew, C. B. (2017). Subverting historical cause & effect: Generation of mythic biographies in *Caves of Qud*. *Proceedings of the 12th International Conference on the Foundations of Digital Games*, Article 76.
13. Klein, T., & Nabi, M. (2019). Learning to answer by learning to ask: Getting the best of GPT-2 and BERT worlds. *arXiv preprint arXiv:1911.02365*.
14. Walton, N. (2019). AI Dungeon. *Game [PC, Android, iOS]*. Retrieved from https://www.aidungeon.io.
15. Ryan, J. O., Barackman, C., Kontje, N., Owen-Milner, T., Walker, M. A., Mateas, M., & Wardrip-Fruin, N. (2014). Combinatorial dialogue authoring. *International Conference on Interactive Digital Storytelling*, 13–24. Springer.
16. Lee, J.-S., & Hsiang, J. (2020). PatentTransformer-2: Controlling patent text generation by structural metadata. *arXiv preprint arXiv:2001.03708*.
17. Wang, Y., Liu, X., & Sun, M. (2021). Emotion-aware dialogue generation with adaptive contextualization. *Proceedings of ACL*.
18. Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., & Weston, J. (2021). Recipes for building an open-domain chatbot. *Proceedings of EACL*.
19. Zheng, Y., Zhang, R., Mao, X., & Huang, M. (2019). A pre-training based personalized dialogue generation model with persona-sparse data. *Proceedings of EMNLP*.
20. Danescu-Niculescu-Mizil, C., & Lee, L. (2011). Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogues. *CMU Technical Report*.
21. Occhipinti, D., Tekiroglu, S., & Guerini, M. (2024). PRODIGy: A profile-based dialogue generation dataset. *Findings of the Association for Computational Linguistics: NAACL 2024*, 3500–3514. Association for Computational Linguistics.