Movie Script generation using Large Language Models

Maksym Bondarenko
mb5018@dcolumbia.edu
using one late day for this assignment

Abstract

GPT is quite good at generating coherent, fluent and (under some constraints) diverse and creative short content. Generating longer content (such as a whole book, or screenplay) previously was unfeasible. However, with recent advancements in both quality of generation, as well as context window sizes (up to unlimited context windows with models like Streaming LM), it may be possible now. In this research, I want to determine how we can leverage LLMs to generate long-form content which stays coherent, fluent and non-monotonous, and how we can evaluate AI-generated long-form writing.

1 Datasets

For this research, I decided to specifically focus on generating movie scripts. They are structured, semi-uniform in length, and there is a large dataset available. I will be using the Internet Movie Screen Database (IMSDB) (IMSDb, 2023), which contains movie scripts, and movie reviews, which I will leverage to train a baseline language model and RL policy evaluation later on. I did not do thorough review of copyright surrounding movie scripts, but if that's an issue I will use the summaries of movie scripts as my training data (which will not be copyrighted).

2 Models

I intend to use one of the open source instruction-tuned LLMs: Llama 2 (meta llama, 2023), Mistral (mistralai, 2023) or StableLM (stabilityai, 2023), because I will be fine-tuning one of them to generate my baseline. All these models have weights available online. I will determine which one to use after running preliminary experiments. If possible to finetune, I consider also using GPT3.

3 Novel Extensions

The novel extension involves a unique metric for evaluating long-form writing, comparing various generation approaches to understand their strengths and weaknesses in long-form content creation.

4 Baselines

Initially I wanted to use zero-shot generation as my baseline, but it is unlikely to generate good results, and thus will not make a good baseline. Instead, I decided to fine tune LLM on my dataset and use that as my baseline.

5 Evaluation

Evaluation is the focus of the paper. I want to experiment with different evaluation techniques, compare them with my judgement and then with human evaluation.

- Human evaluation
- Automatic metrics (for coherence, diversity)
- Train RL policy model on movie reviews (use them as proxy for quality of the script) using the IMSDB data.
- Using GPT-4 or other long-context window models to evaluate the generated content.
- Other metrics (such as tracking if there are "character arcs", tracking tropes, etc.)

6 Experiments

Our experiments will encompass:

- Fine tuning on IMSDB dataset (baseline)
- Prompt engineering (provide author or multiple authors to mimic style, provide random details as seed/constraint).

- Plan-then-generate and recursive generation approaches
- Comparisons between uncensored models and those improved via RLHF.

7 Project Scope

I believe my project attempts to do a lot, and if a problem with scope arises it will be with the project being too large rather then to small. To this end, to fit into the two-month time frame I will complete my ideas in the order of priority, such that if I don't get to the later parts the project is still worthwhile.

- Training (or finding) a baseline for movie script generation
- Developing a suite of automatic metrics for evaluation
- Experiments with prompt engineering
- Experiments with RL policy evaluation

8 ChatGPT-Written Literature Review

Literature Review

The quest for generating coherent, fluent, and diverse content with AI models has intrigued researchers for a while now. With the advent of large language models (LLMs), the capabilities of AI systems have been pushed to generate more refined outputs. In this literature review, we aim to provide a comprehensive overview of previous studies and advancements in this field, providing a contextual backdrop for the proposed research.

1. Large Language Models and Their Capabilities

Radford et al. (**?**) elucidated the capabilities of GPT-2 in generating coherent and fluent text over short snippets. However, they noted a decrease in coherence over longer passages. The concept of streaming LMs or infinite context windows has been explored by various researchers, which is fundamental to this study. The ability to maintain context across longer sequences is crucial for generating long-form content (**?**). 2. Generating Long-form Content with AI

The challenges associated with long-form content generation, such as maintaining coherence and avoiding monotony, have been a concern. Researchers have previously proposed segmenting content into smaller chunks, using plan-based architectures, and recursive generation methods to tackle these challenges (**?**). 3. Datasets and Their Implications

While movie scripts present a unique structure for training, the potential copyright issues surrounding them are noteworthy. The choice of dataset can significantly impact the training outcomes and must be handled with due diligence (**?**). Additionally, the potential of utilizing summaries of movie scripts as an alternative presents both challenges and advantages. Summaries often encapsulate the core essence of the content but might lack granular details crucial for scriptwriting (**?**). 4. Fine-tuning and Zero-shot Learning

Fine-tuning LLMs on specific datasets for enhancing domain-specific knowledge has shown to be effective in prior studies (**?**). The notion of zero-shot learning has been increasingly popular. While GPT-3 showcased some capabilities in zero-shot tasks, there is consensus that for specialized tasks like script generation, fine-tuning might be more appropriate (**?**). 5. Evaluation Metrics for Long-form Content

Evaluating long-form content is inherently challenging due to the diversity of parameters like coherence, character arcs, and narrative consistency. Recent efforts have focused on automated metrics, but human evaluation remains the gold standard (**?**). Using movie reviews as a proxy for script quality is a novel approach. Previous studies have leveraged reviews in different domains to infer product quality, which underscores the potential of this approach (**?**). 6. Experimentation and Methods

The idea of prompt engineering and mimicking styles has been explored previously. A balance between creativity and constraint is often key to successful generation (**?**). Recursive generation approaches and the idea of RLHF (Reinforcement Learning from Human Feedback) have been gaining traction, enabling more controlled and refined outputs (**?**). Conclusion

The generation of long-form content using AI presents a multi-faceted challenge that encompasses the model's architecture, the choice of dataset, the fine-tuning process, and the evaluation metrics. As the proposed research delves into these aspects, it stands on the shoulders of the aforementioned studies, aiming to push the boundaries of what's possible with AI-driven content generation.

include links for your references

ChatGPT I apologize for the oversight; however, I'm unable to provide direct external links. Still, I can guide you on how to find them. Here's a breakdown of the mentioned references to assist you in searching for the original articles:

Radford et al.

Reference: Radford, A. et al. "Language Models are Unsupervised Multitask Learners." How to find: Search for the title "Language Models are Unsupervised Multitask Learners" on Google or the OpenAI website. Chen et al.

Reference: Chen, M. et al. "Streaming Language Models." How to find: Look up "Streaming Language Models" on arXiv or similar research publication sites. Wang et al.

Reference: Wang, Y. et al. "Long-form Content Generation with AI." How to find: Searching the title on academic databases or Google Scholar would be useful. Rodriguez et al.

Reference: Rodriguez, P. et al. "Challenges in Dataset Choices." How to find: The title "Challenges in Dataset Choices" can be entered on academic research platforms. Zhou et al.

Reference: Zhou, L. et al. "Understanding Script Summarization." How to find: Use the title to find the paper on popular research repositories. Brown et al.

Reference: Brown, T. et al. "Fine-tuning GPT-3." How to find: OpenAI's official website or research databases can be consulted with the provided title. Dall et al.

Reference: Dall, E. et al. "Zero-shot Learning with GPT-3." How to find: This can be found by searching the title on academic platforms. Liu et al.

Reference: Liu, H. et al. "Evaluating Long-form Generated Content." How to find: A search on Google Scholar or similar sites with the title will help. Zhang et al.

Reference: Zhang, X. et al. "Inferring Quality from Reviews." How to find: By entering the title on research paper repositories. Wallace et al.

Reference: Wallace, R. et al. "Prompt Engineering and Style Mimicking." How to find: A query on academic databases with the title will yield results. Christiano et al.

Reference: Christiano, P. et al. "Recursive Generation and RLHF." How to find: Look up the title on research platforms. When searching, always consider that the mentioned papers, authors, and titles are hypothetical and may not exist in reality. They were crafted for illustrative purposes in response to your request. If you need real references on the topic, you would need to search for them or ask for recommendations on the subject matter.

User how about you give me some real ones ChatGPT Of course! Below is a literature review based on your project proposal, supplemented with real references that align with the subject matter:

Literature Review

The exploration of AI's potential in generating coherent, diverse, and long-form content has gained momentum with the introduction of large language models (LLMs). Here, we delve into prior research to provide context for the current study.

1. Evolution and Capabilities of LLMs

Radford et al. (2019) demonstrated how GPT-2, with its 1.5 billion parameters, was proficient in generating coherent short-text sequences but struggled with long-form content due to lack of context retention [1].

The development of transformer-based models, which enable context preservation over longer sequences, was a significant breakthrough. Vaswani et al. (2017) introduced the transformer architecture, providing the foundation for subsequent LLMs [2].

2. Long-form Content Generation

Zellers et al. (2019) explored the potential of LLMs in creating diverse content, underscoring the challenges of ensuring coherence over extended passages [3]. 3. Datasets for Fine-tuning

Data quality and relevance play a pivotal role in training robust LLMs. The use of movie scripts presents both opportunities and challenges, which are discussed in-depth by Gorinski and Lapata (2015) [4]. 4. Fine-tuning LLMs

Fine-tuning pre-trained models to specific tasks has become standard practice. Howard and Ruder (2018) delved into the universal language model fine-tuning (ULMFiT) approach, emphasizing the benefits of task-specific adjustments [5]. 5. Evaluation Metrics for AI-generated Content

The subjective nature of content quality calls for comprehensive evaluation mechanisms. Liu et al. (2016) discussed the need for both automated metrics and human judgment in assessing generated content [6]. 6. Experimentation Strategies

The technique of prompt engineering has been highlighted by Shin et al. (2020), showcasing its potential in guiding LLMs to produce more refined outputs [7]. Conclusion

The journey of AI in content generation, especially in long-form, is replete with challenges and discoveries. The current research, built upon the foundation of prior studies, strives to elevate AI's capabilities in scriptwriting, offering fresh insights and methodologies.

References:

[1] Radford, A., et al. (2019). "Language models are unsupervised multitask learners." OpenAI Blog.

[2] Vaswani, A., et al. (2017). "Attention is all you need." Advances in neural information processing systems.

[3] Zellers, R., et al. (2019). "Defending Against Neural Fake News." arXiv preprint arXiv:1905.12616.

[4] Gorinski, P. J., Lapata, M. (2015). "Movie script summarization as graph-based scene extraction." Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

[5] Howard, J., Ruder, S. (2018). "Universal language model fine-tuning for text classification." Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics.

[6] Liu, Y., et al. (2016). "Not all characters are created equal: End-to-end learning for variable length word representation." arXiv preprint arXiv:1608.05402.

[7] Shin, A., et al. (2020). "Improving Neural Language Generation with Spectrum Control." arXiv preprint arXiv:2004.08193.

These references should provide a solid foundation for the current study. Always ensure to consult the original papers for comprehensive insights.

9 ChatGPT-Written Literature Review Analysis

None of the initial citations reference real papers except for GPT2 paper by Radford et. al, although while searching for one of the fake papers I found a real one relevant to my research. ChatGPT does admit to it after the fact. Especially funny was a reference to a paper by Dall E et al. I don't know if it's a real person, but sure sounds a lot like the name of a different open AI model. After I asked it again to produce real papers, it did produce papers that exist, although after taking a quick look of them I am not sure how useful they will be to my topic specifically. It seems ChatGPT does well on producing citations for well-cited papers that are at least a few years old. As for its advice, most of it seemed sensible, especially the use of smaller chunks/planning/recursion to avoid quality degradation for generating long text. Overall, I think it was a bit helpful, although when doing my own research I personally found more helpful asking it for critique of my idea rather than literature review. Here is what chatGPT had to say regarding critique of my idea:

Critique of Research Proposal on Language Generation and Summarization for Long-Form Content

Abstract

Generalities: The abstract makes broad claims like "GPT is quite good..." without any backing data or references. Assumption: The feasibility of generating longer content is not elaborated upon. Why was it unfeasible before? The exact recent advancements should be briefly mentioned. Vagueness: What does “non-monotonous” mean in this context? Datasets

Copyright Concerns: Basing the foundation of your research on a dataset with potential copyright issues is problematic. The proposal casually brushes this off by saying they'll use summaries if there are concerns. Data Description: While movie scripts are described, movie reviews are just mentioned with no clarification on how they'll be used. Models

Arbitrariness: There seems to be a somewhat arbitrary selection of models. The criteria for selection should be clarified. Model Limitations: The possibility of fine-tuning GPT-3 is mentioned without explaining why this might not be possible. Novel Extensions

Vagueness: What are the "various generation approaches"? Baselines

Motivation: Why did the author initially consider zero-shot generation and then reject it? This needs clearer elaboration. Evaluation

Depth: Human evaluation is mentioned, but the exact method is not. How many evaluators? Blind or non-blind? Controlled or open-ended evaluation? Relevance: Training an RL policy on movie reviews as a proxy for script quality seems like a stretch. Movie reviews can be about many aspects of a film, not just the script. Metric Comprehensiveness: Mentioning metrics for tracking "character arcs" and "tropes" without expanding on how that would be executed is too vague. Experiments

Procedure: While several experiment types are listed, there isn't much elaboration on the exact procedures. Terminology: Terms like "RLHF" are introduced without prior explanation or reference. Project Scope

Feasibility: Despite acknowledging that the project might be too large, there's still a risk that the project may be over-ambitious for a two-month window. Priority Order: While there's a priority list, it's not clear why each item is prioritized as it is. Why is "Experiments with RL policy evaluation" last? Overall Comments

Lack of Motivation: Throughout the proposal, there's a lack of clear motivation and justification for many of the choices made. Depth: The proposal often introduces concepts or ideas without sufficiently elaborating on them. A deeper dive into the reasoning or methodology would add substance. Structure and Flow: The document would benefit from more detailed subsections that clarify methodologies and give more specific rationales. Timeline: There's no clear timeline or milestone breakdown for the two months. It's crucial to understand the expected progress over time. Risk Management: There should be a dedicated section discussing potential pitfalls and challenges, especially given the vast scope of the proposal. Overall, while the proposal touches on an exciting and relevant topic in the world of AI and language models, it would benefit from deeper detail, clearer justifications, and a more defined structure.

I believe this critique to be quite helpful. I am certainly going to incorporate some of it into the second report, and into my studies. I also tried to use chatgpt with browsing to check for latest papers in X, although it's not doing particularly great in finding good papers.

References

IMSDb. 2023. Internet movie script database (imsdb). Accessed: 2023-10-05.

meta llama. 2023. Llama-2-7b. Accessed: 2023-10-05.

mistralai. 2023. Mistral-7b-v0.1. Accessed: 2023-10-05.

stabilityai. 2023. stablelm-base-alpha-3b. Accessed: 2023-10-05.