Optmizing Automuse with GPT-4 Turbo-128k

Cadey A. Ratio
xn--g28h
cadey@xeserv.us

Nicole Brennan
xn--g28h
twi@xeserv.us

Jessica Williams
xn--g28h
jess@xeserv.us

Ashley Kaplan
xn--g28h
ashe@xeserv.us

Stephanie Williams
xn--g28h
phi@xeserv.us

Ma Insa
xn--g28h
mai@xeserv.us

Abstract

Further improvements to the Automuse system described in Automuse[1] are described. The use of GPT-4 Turbo 128k[2] allows for unique opportunities in increasing output quality and quantity. Further adaptations to modernize scenarios and plots are also described.

Introduction

Modern advancements to large language model technology has allowed for creating new and interesting applications on top of the technology. One such application is the Automuse[1] system. This system used TypeScript and GPT-3.5[3] to generate scenarios and prose for “pulp novels”, or low-quality fiction churned out on cheap paper to meet mass needs for entertainment. This paper describes the improvements made to the system by using GPT-4 Turbo 128k[2] and modernizing the scenarios and plots used. The results of these improvements are also discussed.

The goal of Automuse continues to be the generation of low-quality fiction for entertainment. The system is designed to be used by a human operator to generate the scenario, characters, prose, and cover image of the novel in question. The operator is then responsible for editing the output into a finished product. The system is not designed to be used without massive human intervention by a skilled operator.

Automuse is distributed as a GitHub repository at https://github.com/Xe/automuse. Automuse wraps the following tools:

- Plotto[4], a system to automatically generate the plot of a novel with little human effort.
- The ChatGPT[5] API to generate summaries, character biographic information, prose, and cover images.
- The GPT-4 Turbo 128k[2] model to generate the scenarios and prose.
- Pandoc to take generated prose and stitch it together into an eBook.

Motivation

The authors were motivated to create this paper by the existence of the GPT-4 Turbo 128k model. This model allows for 128,000 tokens of context, which is enough to store an entire novel’s worth of text as defined by the National November Writing Month rules[6]. This allows for the possibility of generating an entire novel with a single context window, which is a significant improvement over the previous system having to manage multiple context windows and haphazardly stitch them together.

The use of Plotto continues to be a core part of the value proposition of Automuse. The authors were motivated to modernize the scenarios and plots used by the system to be more in line with modern sensibilities. The authors were also motivated to improve the quality of the generated prose and cover images with the use of modern models and their generational improvements.

Results

Two publications have been generated with Automuse version 2, but only one of them is considered a “novel”. Debt of Stars[7] was the first publication and only came out to 47079 words, about 3000 short of the goal. The second publication, Virtual Virtue[8], finally came out over the bar at 55252 words. Both stories were generated with different scenarios and plots, but both were generated with the same GPT-4 Turbo 128k model.

The use of a single context window for the entire novel did work at the beginning, but you can see it causing deterioration of model output as novels progress. This is likely due to the model being unable to keep track of the entire context window and the context window being too large to be useful. This is a problem that will need to be addressed in future versions of Automuse, likely after future development in coherent context windows has occurred.

To mitigate this, the authors changed Automuse to create a new context window for every book chapter. This allows for the model to remain coherent for the majority of output in a chapter, at the cost of theoretical information loss between chapters. This is a tradeoff that the authors are willing to make for the time being.

Methodology

The overall flow for generating a novel remains unchanged from Automuse version 1, mirroring the normal "inside-out" process to write a novel:

- Automuse creates a scenario and plot using Plotto.
- The algorithm selects a "catch" (additional plot element) from a fixed list.
- Automuse expands this into a plot summary and list of chapters with the ChatGPT API.
- Automuse creates character summaries with the ChatGPT API.
- Automuse generates chapter summaries with the ChatGPT API.
- Automuse generates prose with the GPT-4 Turbo 128k model via an iterative "typewriter" process.
- Automuse generates a cover image with the Dall-E 3[9] API.
- The operator stitches the generated prose and images together with Pandoc.

The main difference between Automuse version 1 and Automuse version 2 is the use of GPT-4 Turbo 128k instead of GPT-3.5 4k. This allows for putting more context into the model, which allows for more coherent output.

Scenario and Plot

The biggest advantage to GPT-4 models is the increased quality in written content. Here is an example of a plot summary written by GPT-4:

> In an era where the sanctity of the internet has evolved into the paramount religion, "Virtual Virtue" tells the tale of Haley's unconventional quest to confirm the authenticity of aer partner Braylon's love. Amidst a technologically advanced society that worships connectivity and digital presence, Haley, whose relationship with Braylon has culminated in a prolonged engagement, resolves to disguise aerself and engage with him as a stranger. Troubled by the pervasive artificiality that the internet-first culture has brought into interpersonal relationships, Haley crafts a perilous situation—a test of courage—to challenge Braylon's devotion. This leads Haley on an intricate journey, not only to possibly save Braylon from the very trap that ae had set but also to untangle a complicated web of digital deceit, risking aer own integrity in the process. The story unfolds, revealing the struggle between genuine human connection and the facades maintained in a world dominated by online personas.

This is a significant improvement in quantity and quality over the previous version of Automuse, which would have generated something like this:

> After a disastrous turn of events, software engineer, Mia, finds herself stranded on a deserted island with no communication to the outside world. Mia uses her knowledge of peer to peer networks to create a makeshift communication system with other stranded individuals around the world, all connected by the same network. Together, they navigate survival and search for a way back to civilization while facing challenges posed by the island.

The authors did not change the prompts between both versions of Automuse, so the difference in output is entirely due to the model change.

Characters

Automuse version 1 did not spend a lot of effort in character descriptions, which led to the characters being very flat and uninteresting. Automuse version 2 uses the following method to generate character descriptions:

- The Plotto summary creates a list of characters and their roles.

- Automuse adds in randomly generated gender and sex characteristics to each character.
- Automuse generates a character summary with the ChatGPT API.
- Automuse writes these summaries to markdown files on disk for later processing.

These character summaries are fed into the system prompt of prose generation runs. This allows for the model to have more context about the characters and their motivations, which leads to more interesting prose.

Prose

The authors found that the following prompt keywords worked best for generating prose:

> You will be given a description of a scene and you will write the prose for the scene. You will output the prose for the scene. [...] Include realistic dialogue. Be creative and descriptive. Use a lot of detail. Write something that has an outstanding plotline, engaging characters and unexpected climaxes.

When combined with the plot summary, character summaries, and chapter summaries, this leads to each “user” message being the individual scene summary with each “assistant” message being the generated prose. This allows for the model to have more context room to generate prose instead of having to over-verbosely be told what to do at every step. This allowed the authors to save on tokens being processed, which allowed for more prose to be generated for the same cost.

This allows for generated prose like this:

> With a breath that seemed to carry the weight of worlds, Haley pressed ‘Send.’ The message vanished into the nexus, a digital missive crossing the stream of zeroes and ones to arrive at Braylon’s doorstep.

Instead of prose like this:

> When they finally had a solid plan, Mia took charge. She donned her backpack, pulling out a rusty old radio, some wire, and a series of small batteries. They would need to make contact with other stranded individuals around the world, all connected by the same network. And with their knowledge and expertise, they could do it.

Cover Images

Previously, the cover images for Automuse version 1 novels required manual human intervention to create. This was a time-consuming process that required a lot of manual work. Automuse version 2 uses the Dall-E 3 API to generate cover images. This allows for images like the following to be generated:

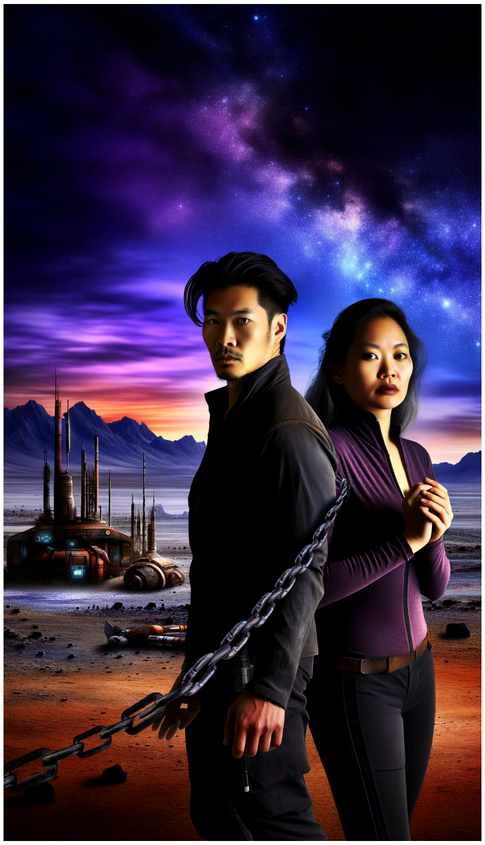

Figure 1: The cover image for Debt of Stars by Midori Yasomi.

Known Issues

Automuse version 2 is a generational improvement over Automuse version 1, but it is not without its flaws. The authors have identified the following issues with the system:

- Automuse uses Plotto as a source of plot generation. Plotto was created in 1928 and reflects many stereotypes of its time. Careful filtering of Plotto summaries is required to avoid repeating harmful cultural and social biases.
- Automuse uses the GPT-4 Turbo 128k model, which is a proprietary model. This means that the authors cannot share the model weights with the public. This is a problem that will need to be addressed in future versions of Automuse.
- Automuse uses a private dependency to "fix" Plotto's issues with inherent racism. This version of Plotto cannot be made public due to fears that the authors will be attacked by source code hosting providers for hosting "hate speech" or "racist content" via the unknown racist entries in the Plotto source code that have not been filtered out yet.
- Using one context window for writing an entire novel presented unique problems with internal coherency. Automuse version 2 was adapted to use a new context window for every chapter, which allows for more coherent output at the cost of information loss between chapters. This will need to be addressed in future versions of Automuse.

Potential Societal Impacts

According to Dan Olson's documentary about the predatory ghostwriting industry named Contrepreneurs: The Mikkensen Twins[10], the average pay rate for a ghostwriter for The Urban Writers can get as low as USD$0.005 per word. Given that Automuse version 2 spends about USD$25.00 to write about 50,000 words (USD$0.0002 per word) using GPT-4 Turbo 128k, this makes Automuse a significant cost reduction in the process for creating pulp novels. It is twenty-five times as cheap as hiring a horribly underpaid ghostwriter who has to endure horrors that no human should be exposed to.

This is a setback from Automuse version 1's cost of USD$0.20 to write 50,000 words, but the much higher quality of GPT-4 Turbo 128k's output more than makes up for the increase in cost. The authors believe that this is a net positive for society.

The quality of Automuse novels still is unable to compare to the output of a highly skilled human writer. Due to the fundamental nature of these networks, they cannot truly create new output, but they can create unexpected combinations of things that have already existed. This is a fundamental limitation of the technology that will need to be addressed in future versions of Automuse, if not in large language models themselves.

The authors of this paper do not believe that Automuse version 2 can displace human authors from their jobs. A version of Automuse could be used to help human authors generate and test ideas, but the authors do not believe that Automuse can replace human authors entirely with any amount of improvements.

The authors would like to note that the conditions that writers for The Urban Writers work in are so terrible that displacing their labor may net them long term benefits provided they can survive the short term loss of their main source of employment vanishing. These displaced workers would still need to pay money to eat food and purchase lodging, meaning that a mass-adoption of Automuse could have massive downsides for this class of workers. The authors would like to note that the authors of this paper are not lawyers and cannot provide legal advice.

Future Work

As future models get released and improved upon, the authors would like to continue to improve Automuse. The authors would like to continue to improve the quality of the generated prose and cover images. The authors would also like to continue to improve the quality of the generated plots and scenarios. The authors would like to improve the relevance of the stories that Automuse tells as well as making them more reflective of the inherent diversity in humanity.

The authors would also like to base future versions of Automuse on open source large language models.

This would allow Automuse novels to cover topics that are inherent to human existence (such as sexual relations) but are otherwise not allowed with hosted models like GPT-4 Turbo 128k. This would also allow for the authors to share model weights for Automuse as well as fine-tune different models for the different tasks in this system.

The authors would like to optimize the costs of Automuse to the point that they can set up a "book club" for people to read and review changes to Automuse generated prose. This would allow for the quality of Automuse prose to be incrementally improved over time by human interaction and reviews, which would allow for the authors to better prioritize finetuning runs.

Acknowledgements

The authors would like to thank the following people and organizations for their support and assistance in the development of this project:

- The OpenAI team for their work on GPT-4 Turbo 128k[2] and for sponsoring the server time used to run inference.
- Paul M. for continued inspiration and support.
- Patrons of the Xe Iaso blog for their continued patience and support.

Bibliography

[1] C. Ratio, N. Brennan, Williams J, S. Williams, A. Kaplan, and I. M., "Automuse: A System for Generating Pulp Novels". 2023.

[2] "New models and developer products announced at DevDay --- openai.com". Nov. 2023.

[3] J. Ye *et al.*, "A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models". 2023.

[4] W. W. COOK, *Plotto: A new method of plot suggestion for writers*. 1928.

[5] ChatGPT, 2023.

[6] N. N. W. Month, "Why 50,000 words? And how do you define "novel"?".

[7] Y. Midori, "Debt of Stars". 2023.

[8] Y. Midori, "Virtual Virtue". 2023.

[9] I. OpenAI, "DALL·E: Creating images from text --- openai.com". 2023.

[10] "Contrepreneurs: The mikkelsen twins". [Online]. Available: https://youtu.be/biYciU1uiUw