Domain-Specific Small Language Models for Creative Text Generation: A Case Study on the Backrooms

Roxanne Silva Julia

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Uberlândia
2026

Roxanne Silva Julia

Domain-Specific Small Language Models for Creative Text Generation: A Case Study on the Backrooms

Dissertação de mestrado apresentada ao Programa de Pós-graduação da Faculdade de Computação da Universidade Federal de Uberlândia como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Inteligência Artificial

Orientador: Elaine Ribeiro de Faria Paiva
Coorientador: Marcelo Zanchetta do Nascimento

Uberlândia
2026

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da UFU, MG, Brasil.

J94d
2026

Julia, Roxanne Silva, 1999-
Domain-Specific Small Language Models for Creative Text Generation [recurso eletrônico] : a case study on the Backrooms / Roxanne Silva Julia. - 2026.

Orientadora: Elaine Ribeiro de Faria Paiva.
Coorientador: Marcelo Zanchetta do Nascimento.
Dissertação (Mestrado) - Universidade Federal de Uberlândia, Programa de Pós-graduação em Ciência da Computação.
Modo de acesso: Internet.
Disponível em: http://doi.org/10.14393/ufu.di.2026.10021
Inclui bibliografia.
Inclui ilustrações.

1. Computação. I. Paiva, Elaine Ribeiro de Faria, 1980-, (Orient.). II. Nascimento, Marcelo Zanchetta do, 1976-, (Coorient). III. Universidade Federal de Uberlândia. Programa de Pós-graduação em Ciência da Computação. IV. Título.

CDU: 681.3

Nelson Marcos Ferreira
Bibliotecário(a)-Documentalista - CRB-6/3074

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Coordenação do Programa de Pós-Graduação em Computação
Av. João Naves de Ávila, 2121, Bloco 1A, Sala 243 - Bairro Santa Mônica, Uberlândia-MG, CEP 38400-902
Telefone: (34) 3239-4470 - www.ppgco.facom.ufu.br - cpgfacom@ufu.br

ATA DE DEFESA - PÓS-GRADUAÇÃO

Programa de Pós-Graduação em:	Ciência da Computação				
Defesa de:	Dissertação, 13/2026, PPGCO				
Data:	18 de Junho de 2026	Hora de início:	14:30	Hora de encerramento:	16:25
Matrícula do Discente:	12322CCP025				
Nome do Discente:	Roxanne Silva Julia				
Título do Trabalho:	Domain-Specific Small Language Models for Creative Text Generation: A Case Study on the Backrooms				
Área de concentração:	Ciência da Computação				
Linha de pesquisa:	Inteligência Artificial				
Projeto de Pesquisa de vinculação:	----------				

Reuniu-se por videoconferência, a Banca Examinadora, designada pelo Colegiado do Programa de Pós-graduação em Ciência da Computação, assim composta: Professores Doutores: Marcelo Zanchetta do Nascimento (Coorientador) - FACOM/UFU, Fabíola Souza Fernandes Pereira - FACOM/UFU, Ronaldo Cristiano Prati - UFABC e Elaine Ribeiro de Faria Paiva - FACOM/UFU, orientador do(a) candidato(a).

Os examinadores participaram desde as seguintes localidades: Ronaldo Cristiano Prati - Santo André/SP. O aluno(a) e os outros membros participaram da cidade de Uberlândia.

Iniciando os trabalhos o(a) presidente da mesa, Prof. Dr. Elaine Ribeiro de Faria Paiva, apresentou a Comissão Examinadora e o(a) candidato(a), agradeceu a presença do público, e concedeu ao(à) Discente a palavra para a exposição do seu trabalho.

A seguir o senhora presidente concedeu a palavra, pela ordem sucessivamente, aos examinadores, que passaram a arguir o(a) candidato(a). Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando o candidato(a):

Aprovado

Esta defesa faz parte dos requisitos necessários à obtenção do título de Mestre.

O competente diploma será expedido após cumprimento dos demais requisitos, conforme as normas do Programa, a legislação pertinente e a regulamentação

interna da UFU.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme foi assinada pela Banca Examinadora.

Documento assinado eletronicamente por **Elaine Ribeiro de Faria Paiva**, **Professor(a) do Magistério Superior**, em 19/06/2026, às 10:27, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Fabíola Souza Fernandes Pereira**, **Professor(a) do Magistério Superior**, em 19/06/2026, às 11:08, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Ronaldo Cristiano Prati**, **Usuário Externo**, em 23/06/2026, às 16:09, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Marcelo Zanchetta do Nascimento**, **Professor(a) do Magistério Superior**, em 23/06/2026, às 16:36, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **7377716** e o código CRC **638F5F2C**.

Referência: Processo nº 23117.037429/2026-98 SEI nº 7377716

I dedicate this work to my family, my advisors, my friends and to all stories that have made me dream.

Agradecimentos

If you are not careful and noclip out of academia in the wrong areas, you may end up in a master's degree dissertation composed of fluorescent overleaf lights, endless revisions, mysterious bugs, and the constant background hum of deadlines. Somehow, after wandering these halls for nearly three years, I finally found an exit, thanks to the help of a lot of really dear people who I want to thank here.

I would like to express my sincere and deep gratitude to my advisor, Prof. Elaine Ribeiro de Faria Paiva, for the guidance, patience, and support provided throughout this master's research. Thank you for helping me transform my strange ideas about small language models and the Backrooms into an actual scientific work.

I also want to thank my co-advisor Prof. Marcelo Zanchetta do Nascimento and Prof. Rita Maria da Silva Julia, whose comments, suggestions, support and excitement with this research contributed to improving this work immensely.

Special thanks to my colleagues, friends, and everyone who listened to me talk about language models, datasets, training logs, GPUs, and surreal liminal spaces far more often than any normal person should. Thank you also for taking my mind out of all of this in times when I really needed it and for filling my days with laughter and happy memories in this long difficult journey.

Thank you to all Human Judges who gave their time to evaluate all these crazy Backroom level generations. You have really helped me enrich my research.

I would also like to acknowledge the open-source community, whose tools, models, datasets, libraries, and research made this work possible. Modern research in Machine Learning and Large Language Models is deeply collaborative, and this dissertation stands on the work of countless researchers and developers around the world.

A special and liminal thank you to everyone who worked and still works on the Backrooms wikis. This is a beautiful creative collective project and I couldn't admire you more. I also could never have created the datasets used in this research without you.

To my mother, father, brother, grandmother and the rest of my beautiful family, thank you for the unconditional support, encouragement, and understanding during this adventure. You

know how crucial you were in this process and how your patience and love have gotten me trough the darkest patches of this journey. Your support was the biggest backbone of this work and have kept me going even when the path ahead looked endless.

Finally, I thank The Backrooms themselves and every one in the community who daily contributes to enriching it's lore: because of you, I was able to somehow turn an internet urban legend about empty yellow rooms into a legitimate academic case study. Please never stop having these crazy wonderful minds of yours.

“If you’re not careful and noclip out of reality in the wrong areas, you’ll end up in the Backrooms, where it’s nothing but the stink of old moist carpet, the madness of mono-yellow, the endless background noise of fluorescent lights at maximum hum-buzz, and approximately six hundred million square miles of randomly segmented empty rooms to be trapped in. God save you if you hear something wandering around nearby, because it sure as hell has heard you.”

(Unknown)

Abstract

Large Language Models (LLMs) occupy a prominent place among Machine Learning models that have been revolutionizing modern life as an indispensable support tool in the most diverse areas of human activity. Nevertheless, the popularization of large LLMs faces a significant limiting challenge: the high cost associated with their use, whether through expensive hardware, cloud processing or paid APIs. In this light, the main objective of this work is to propose an approach to generate smaller language models, which can be trained in more modest hardware and that are effective in generating texts related to specific domains. The case study chosen for that are The Backrooms, an online legend which describes an alternative reality with different *levels* that are described with very unique, surreal and rich descriptions. For this, a Backrooms dataset, was created and used to produce two small language models (SLMs) to generate Backrooms level descriptions: Backrooms-Llama, conceived from the fine-tuning of the open source pretrained model Llama 1B; and Tiny Backrooms, a decoder-only architecture pretrained from scratch and then fine-tuned. The performance of these two models was evaluated against the famous and much larger DeepSeek-V3. The generated level descriptions of the three models were evaluated by using chatGPT 4o-mini as judge and human judges. Additionally, a brand new evaluation method, which consists on generating pictures with the output of each competitor model and, then, evaluating these pictures alongside their descriptions with chatGPT 4o, was proposed. The results found are very promising, showing that, even though DeepSeek performed better in general, Backrooms-Llama and Tiny Backrooms, operating in a much more modest architecture, also got mainly relevant evaluations and were able to learn how to generate Backrooms level descriptions. Such results indicate that, in creative generative tasks, a lower cost domain-specific SLM, when properly optimized, can achieve performance comparable to that of consecrated LLMs.

Keywords: Small Language Models, Natural Language Processing, Large Language Models, Computational Storytelling, Generative Models, Creative Tasks, The Backrooms.

List of Figures

List of Tables

Lista de siglas

Contents

CAPÍTULO 1

Introduction

In the current fast-paced scientific world, Large Language Models (LLMs) have become one of the main exploration subjects in the artificial intelligence (AI) community (Kasneci et al. 2023), (Chang et al. 2024). Among the most widely used LLMs today are, for instance, ChatGPT, DeepSeek (DeepSeek-AI et al. 2025), Gemini (McDonald e Emami 2024). These LLMs are remarkable in that they achieve very high performance across a wide range of specific domains. Such admirable success in performance is the result of exhaustive training using appropriate techniques and an immeasurable volume of data sourced from numerous origins, all within extremely powerful processing environments.

Given this, the popularization of LLMs faces a significant limiting challenge: the high cost associated with their use and training process, whether through expensive hardware, cloud processing, or paid APIs (Bender et al. 2021) (Kaplan et al. 2020). This bottleneck in the widespread adoption of more powerful and generic LLMs stems from the fact that most of the global scientific community and the general public do not have access to the costly hardware and infrastructure needed actually to experiment with them.

With this in mind, Small Language Models (SLMs) have been developed with the aim of running on less complex machine architectures, such as the smaller Llama models, GPT-2 (Radford et al. 2019) and GPT-Neo (Black et al. 2021). Even though these open-source smaller models are more accessible to the general public, most of them usually are not nearly as good at performing multiple specific tasks as the well-succeeded larger generic chat models aforementioned, since they simply consist of either pre-trained backbones or on models fine-tuned in a much smaller range of tasks.

With regard to using APIs such as Deepseek or chatGPT, even though they offer smaller models available at affordable costs for small tasks, the user depends on their endpoint to run inference, which can introduce problems and instabilities (Bommasani et al. 2021) (Breck et al. 2017). Additionally, if used for larger tasks or for longer periods of time, these models might become expensive. Also, by not running these models locally, architectural changes can not be explored, which can be limiting research-wise. Finally, there is the question of privacy concerns and content moderation policies. For dealing with private data and not being limited by modera-

tion policies, it would be a lot safer and practical to run a model locally (Carlini et al. 2021) (Carlini et al. 2021).

One of the most interesting aspects to explore in this research field is the performance and potential of these models in creative tasks such as writing stories (34), songs (Marco et al. 2024) and poetry (Chen et al. 2024). Considering the modern context in which creativity is increasingly becoming an essential attribute in everyone's life, the proven effectiveness of LLMs in creative activities justifies the need for their popularization as support tools for such purposes. This becomes even more relevant when it comes to creative tasks involving text production, since this represents the most universal and common form of communication and expression.

When it comes to gathering creative public textual data, wikis, in particular, can be very interesting, since they are essentially online encyclopedias covering a large range of subjects. This richness in text, in addition to the fact that they are free and public, makes wikis a very suitable data source to train Transformer-based architectures, as well as to investigate the generative ability of LLMs and SLMs. In this light, many works have created datasets based on wikis and/or used such datasets in studies related to LLMs (Hou et al. 2024), (Caffagni et al. 2024).

Two very original, unique, and large Wikis are dedicated to The Backrooms [1]. The Backrooms is a concept that originated from a *creepypasta* (a kind of horror fiction shared online), which describes an alternative reality that people supposedly slip into, usually without warning. It is an endless labyrinth of unsettling rooms located on different levels. Each level has specific characteristics, rules, and monsters, which are called 'entities'. The concept has grown into a popular online urban legend and has inspired a variety of stories, games, and media. It has also become the study subject of many works exploring socio-cultural and psychological aspects of urban legends and why people are so fascinated by them (Stephen 2022), (Zawacki 2024). Considering the size and creative nature of these Backrooms wikis, they represent an appropriate case study for evaluating LLMs in specific creative generative tasks.

Investigating how SLMs behave in creative generative tasks when compared to larger ones and even to human generation is a very relevant and on-trend study field, as can be seen in several recent state-of-the-art works. In fact, in (... 2024) the authors explored two different approaches to generate role-playing game (RPG) video game quest descriptions: one, in which they fine-tuned GPT-2; another one in which they just generated quest descriptions with the vanilla GPT-3 model, which is larger than GPT-2. In (Marco et al. 2024), the authors carried out a contest between an awarded novelist called Patricio Pron and GPT-4 in the task of writing short stories based on a title.

Better understanding of LLMs' strengths, weaknesses, and which is the best strategy to use them in specific generative tasks, such as creative tasks can be the starting point for many useful research and applications, such as creating games, including educational games (Huber et al. 2024), creating characters, stories, levels, and quests. Despite the fact that many works have started investigating the performance of LLMs for these kinds of tasks, there is still room for a lot more

[1] Wiki, Wikidot

investigation, especially when it comes to comparing the performance of huge generic models, like GPT and Mistral models, which were trained with billions of parameters and with very diverse data that covers almost every subject in specific tasks, against the performance of smaller, more specific models, which were trained for specific tasks with specific data. Furthermore, it could be really beneficial, hardware and money-wise, to investigate the convenience of using much smaller Transformers-based models, trained for a specific task, instead of depending on these massive, generic LLMs.

Another challenge, when it comes to LLMs, is performance evaluation. Evaluating LLMs is not a simple, objective, and straightforward task, especially when the evaluated output consists of Free-Form Text, which is the case here. In this kind of research, humans are often used as evaluators, as in (... 2024), in which people evaluated the generated RPG quests by answering an online questionnaire. The authors in (Lanzi e Loiacono 2023) also used human subjects to work on and evaluate their proposed game design framework. Another famous strategy for evaluating LLMs' outputs is using the LLMs themselves as judges. This is what was proposed in (Badshah e Sajjad 2024): a reference-guided verdict method that automates the evaluation process of Free-Form Text by leveraging multiple LLMs-as-judges.

In this light, this work investigates the possibility of producing models with smaller architectures (therefore, with lower financial and computational costs), trained from specific domain data, that achieve satisfactory performance compared to that of LLMs with much larger architectures and trained from generic domain data, like DeepSeek and chatGPT. Furthermore, this work aims to investigate the feasibility of building an end-to-end language model pipeline, from pre-training to task-specific fine-tuning, enabling full control over the model development process rather than relying exclusively on externally pre-trained foundation models.

Moreover, to enrich the performance evaluation scenery of LLMs, in the experiments carried out here, chatGPT 4o-mini and human evaluators were used to evaluate the quality of the Backrooms level description produced by each strategy. Additionally, a new evaluation method that analyses the coherence and visual quality of images generated from such descriptions by the well-known DL-based architecture Dall-E, was proposed.

1.1 Objectives and Challenges

The main goal of this work is to investigate if, in specific tasks, a smaller model, trained more specifically for the task in question, can perform similarly or even better than these current famous massive general Large Language Chat Models. Additionally, this work aims to investigate the feasibility of building a complete end-to-end language model pipeline, in which the entire model development process—from pre-training to downstream fine-tuning—is carried out within the proposed framework, rather than relying exclusively on externally available foundation models. To achieve this, three different strategies for generating new Backrooms level descriptions with LLMs will be explored: the first one simply consists of some prompt

engineering for generating new level descriptions with chatGPT 4o, which is the most famous, powerful and current LLM; The second one will consist of fine-tuning Llama 1B, for the Backrooms level description generation task; The third one will consist of pre-training a vanilla Decoder-only architecture with Backrooms lore found online to then fine-tune this pre-trained backbone for the Backrooms level description generation task. Moreover, the famous deep learning model Dall-E, designed for generating images from textual descriptions, will also be used for generating images based on the output of the three strategies described above. The images generated, alongside the Backrooms level descriptions outputted by each strategy, will be used in the evaluation process, which will be performed by both human and LLM judges. This main goal can be divided in the following specific goals:

- ❏ Create a public Backrooms Dataset with Backrooms lore text data found online;
- ❏ Produce Backrooms-llama, a fine-tuned Llama 1B for the Backrooms level description generation task;
- ❏ Produce Tiny Backrooms, a vanilla Decoder-only architecture developed through an end-to-end pipeline, including domain-specific pre-training and subsequent fine-tuning for the Backrooms level description generation task;
- ❏ Compare the performance of Backrooms-llama and Tiny Backrooms against the famous Large Language Chat model Deepseek for the Backrooms level description generation task;
- ❏ Evaluate the level descriptions and images generated by the three strategies with human and LLM judges and determine which strategy produced better results.

1.2 Research Question

Current very famous and general Large Language Chat Models, like chatGPT and Llama, not only have been used for a very large variety of tasks nowadays, but also hold the state of the art for many of these tasks (35), (Schryver 2023), (Wang et al. 2023), (Zhao 2023). The question that is being posed in this work is whether, for a specific task, in this case a creative generative task (generating new Backrooms level descriptions), a smaller model, also with a general backbone, fine-tuned for this task, or a vanilla Decoder-only architecture, also much smaller, trained from scratch for this task and then fine-tuned, can achieve a competitive result, based on human and LLM evaluation, when compared to these massive general Large Language Chat Models.

1.3 Dissertation Organization

This chapter presented the goals, challenges and motivation for this work. The next chapters will be organized in the following manner:

- **Chapter 2 - Theoretical Background:** Presents the fundamental theoretical concepts that relate to this work, more specifically: The Backrooms, Transformers, Large Language Models and Small Language Models;
- **Chapter 3 - Related Works:** Presents an overview of the literature related to this work, especially in the context of: LLMs in creative tasks and Evaluation of LLMs in Free-Form Text Tasks;
- **Chapter 4 - Backrooms Level Description Generation** Described the strategies explored in this work for Backrooms Level Description Generation;
- **Chapter 5 - Results** This chapter describes and compares the results obtained by the strategies described in Chapter 4.
- **Chapter 6 - Conclusion:** Presents the conclusions that were reached through this work, its contributions, limitations and future works.

CAPÍTULO 2

Theoretical Background

In this section, the necessary Theoretical Background to understand the proposed work is presented. It was divided in: The Backrooms; Transformers; Large Language Models and Small Language Models; Procedural Content generation .

2.1 The Backrooms

2.1.1 What are the Backrooms?

The Backrooms are a fictional location originated from a 2019 4chan, which is an anonymous English-language image board website which hosts boards dedicated to a wide variety of topics, from video games and television to literature, cooking, weapons, music, history, technology, anime, physical fitness, politics, and sports, among others thread (Wikipedia contributors 2024). This started because of a posted picture, illustrated in Figure 1, which led an anonymous user to start a thread on 4chan's paranormal-themed board, asking users to "post disquieting images that just feel off,"accompanying the thread with the photograph of what is now known as level 0 of the Backrooms. The lore that was created ever since, called The Backrooms, is one of the most popular examples of a concept called liminal spaces. Broadly, the term liminal space is used to describe a place or state of transition that can be physical (like a doorway, corridor or bridge) or psychological (like childhood, adolescence or an existential crisis). Liminal space imagery often depicts this sense of ín-betweenb́y capturing transitional places such as stairwells, roads, corridors or hotels. These 'dreamlike' places are unsettlingly devoid of people and may convey moods of eeriness, surrealism, nostalgia, discomfort, loneliness and sadness (Wikipedia contributors 2024).

In the mentioned 4chan thread, an anonymous user replied to the original post and gave the image its name and supplied the first description of the Backrooms: 'If you're not careful and you noclip out of reality in the wrong areas, you'll end up in the Backrooms, where it's nothing but the stink of old moist carpet, the madness of mono-yellow, the endless background noise of fluorescent lights at maximum hum-buzz, and approximately six hundred mil-

lion square miles of randomly segmented empty rooms to be trapped in God save you if you hear something wandering around nearby, because it sure as hell has heard you'. This was the origin of the Backrooms, a creepypasta. Inspired by it, users began to share stories about the Backrooms on subreddits and a fandom began to develop around it, in which creators expanded upon the original iteration of the creepypasta by creating additional floors or *levels* and entities (monsters that inhabit the backrooms and which origin is unknown) that populate them (Wikipedia contributors 2024).

The fandom steadily expanded onto other platforms like Twitter and TikTok. In 2022, an american YouTuber called Kane Parsons started a series of Backrooms short films on YouTube, which went viral and even got Parsons to be invited to direct a film adaptation of his series produced by A24 (Wikipedia contributors 2024). Games based on the Backrooms have also been made. A huge part of the Backrooms universe was created in the Wikis hosted on Fandom and Wikidot Wiki 1, Wiki 2.

In essence, the Backrooms are usually portrayed as an impossibly large extra dimensional maze of empty, familiar but also unsettling, dream-like rooms, accessed by exiting ("no-clipping out of") reality (which is called the Frontrooms in the Backrooms lore) (Wikipedia contributors 2024). These rooms (or levels) can be inhabited by monster-like creatures called entities and there are a series of rules and legends about how to navigate, survive and even find other humans in the Backrooms. Nobody knows a sure way to leave the Backrooms, but there are also a lot of legends and tales around people who have supposedly done it.

Figure 1 – The picture that started the Backrooms lore, also known as Level 0 of the Backrooms.

2.1.2 The structure of the Backrooms levels

The main pillar of the Backrooms mythology are the levels. As mentioned above, The Backrooms consists of this alternative reality made up by infinite, very surreal levels, which may or may not contain monsters called entities. These levels are like floors on a building, where each floor is supposedly infinite, mostly devoid of human life (it is possible to stumble upon other human in the Backrooms, but very difficult) and contain specific characteristics of familiar places with a surreal twist to them. A few example of famous levels are:

- ❏ Level 0, also known as "The Lobby" is the most famous level of the Backrooms and the

one which started the entire lore. I is an endless maze of yellow, damp, moldy office rooms with buzzing fluorescent lights and stained carpet. It has a disorienting and monotonous atmosphere, with humming lights and the smell of wet carpet. There are few entities, but extreme isolation and sensory fatigue can lead a wanderer to madness;

- ❏ The Poolrooms is an immense complex of tiled rooms, hallways, and chambers partially submerged in still, crystal-clear water. The entire area glows faintly from hidden light sources, creating a serene yet uncanny calm. Quiet, echoing, and dreamlike, it's like an infinite public pool frozen in time. The air smells faintly of chlorine, and sound travels strangely through the humid space;

- ❏ The Suburbs can be described by an endless neighborhood bathed in perpetual sunset, filled with identical pastel houses, white fences, and perfectly mowed lawns. No people are ever seen, yet curtains occasionally twitch and lights flicker on inside empty homes. Comforting and nostalgic at first glance, but the stillness and lack of life quickly become unnerving and prolonged stays induce nostalgia, amnesia, and dissociation;

- ❏ The Crimson Forest is a vast, dense forest bathed in a deep red hue — leaves, soil, and even the mist itself tinted crimson. The trees are unnaturally tall and often seem to shift positions when unobserved. Oppressive and otherworldly, the forest hums faintly as if alive. The red light never fades, making it impossible to tell day from night.

It is important to highlight that there is no official source for the Backrooms story. It's an online legend which grew and became a huge collaborative creative project with many sources and materials spread throughout different platforms. As with any legend, there are many different interpretations and versions of it and the levels listed above are just one version of these famous levels.

These level descriptions show the essence of the concept of The Backrooms, which is the surrealist and dream-like aspect they carry. They resemble a dream one can have, in which they wander a familiar place. This place seems to be empty and infinite, creating a feeling of loneliness, nostalgia and strangeness to the wanderer. The level may contain dislodged things which should not be there and may not follow classic physics laws, but one usually doesn't question these things in a dream.

When it comes to the representation of a Backrooms level, there are four aspects of the level which are usually always mentioned: The main description of the level, the entities of the level, the Entrances and Exits of the level and the Bases, Outposts, and Communities of the level. In the wikis, these aspects became sessions that compose the standard structure of a Backroom level. Here is the full description of the level Crimson Forest, taken from Wiki 1:

The Crimson Forest

Survival Difficulty: Class 0

- Safe
- Stable
- Devoid of entities

The Crimson Forest is a safe haven for the wanderers.

Description

Touched by heaven, a tear of God; the forest is a sacred place, a secure and serene realm. It is devoid of any dangers, protected by the Perfected ones of heaven itself, sent down below to protect those who within the forest reside.Covered in a dense layer of fog—the evaporated tears of the exalted numerous—a forest resides and a community thrives; soil and flora of the same are tinted in ravishing, bewitching shades of crimson and red, giving the level its unique aura and its moniker.Although there is a high presence of ultisols and ferric oxides, the strong coloration of the level is still a mystery—but it is thought to come from the spilled blood of the sacred believers, who died for the safeness of this idyll.The true size of the forest is unknown; delimited by only imagination itself, its borders remain unexplored, untouched; a largely unmapped geography.

Areas that have been explored are home to many different forms of landscapes and terrain formations, with some being a signature of the level, like valleys, dense woods, large mountains, and snaking streams cutting through it all, ending in boggy wetlands. Underneath the forest, no tectonic plate lays flat. These crumple and fold into each other, much like crinkled paper.

In some regions, looking into the sky reveals more of the forest above you, accessible only if one keeps on walking for long enough to reach the fold; however, it should be noted that gravity and direction become meaningless and subjective, the concept of what is 'downwards' loses all sense, as reaching high altitudes will reverse your gravity, causing one to fall upwards onto the fold above. Although these properties are strange, those who reside within this level have become accustomed to them, given that these properties are stable and reliable.

The aforementioned evaporated cries of the holy guardians are present at any time in the level, keeping the temperatures around a chilly 8 °C constantly. Bringing warm clothing is recommended. Much of the water running down the rivers in the forest is potable, yet acrid; nevertheless, boiling the water is still recommended. Wanderers have attempted to combat the unsettling natural taste by collecting the water in containers and letting it settle, allowing to remove the silt that forms on the bottom.

Entities

With a rich ecosystem, the Crimson Forest is home to a wide range of both fauna and flora, ranging from insects to large mammals, new species of plants and fruits, and everything in between. Many of these creatures resemble, and in close similarity, animals common to The Frontrooms yet nothing that resides here was previously known to mankind.

Eight-eyed–deer, rust-colored–moose, squirrels with prehensile tails, colonies of mice that group together in the dozens; even haunting, silent owls with four wings have been documented among hundreds of other animals. Strange, but not unnatural or anomalous, many of these creatures are benevolent; however, there is still a large collection of creatures that are hostile towards other lifeforms, including wanderers. One of these entities are the Trollges, they entered the level via teleportation and chasing of victims sometime around the time of the Fun War. They will brutally murder anyone who disturbs them.

These hostile creatures are part of the stable ecosystem of the level and should not be disturbed; the Perfected ones will not attack these, much like they would not attack wanderers. Hulking creatures resembling bears, lithe wild cats, and packs of six-legged–canines are known to live here in harmony with the ecosystem, just to name a few. There is no Wi-Fi in the Crimson Forest. Wanderers who wish to connect to it and send information out have to pass through Level 9.1 and access a zone with enough coverage. In order to keep information steady and reliable, and to ensure that those who reside outside of the forest still keep up with those living in it, the U.A.E. has created a monthly cycle of data upload and download; however, these cycles are sometimes interrupted by occasional hazards, forced to be postponed. The longest period of radio silence from the Crimson Forest was seven months and sixteen days of standard time.

Since traversing through Level 9 and Level 9.1 is highly risky and dangerous, most wanderers who reach the forest decide to stay, considering it their new home; a permanent residence. Anyone is free to leave the forest at any time, but it should be noted that the Perfected ones only protect the entirety of Crimson Forest, and anything outside of it is out of their jurisdiction.

The Perfected Ones

White, ring-like orbs; these are the Perfected: marvelous creations of God sent down to us to protect the many lost in the depths of the Backrooms.

Patrolling the level at all times, they will ignore wanderers and carry on with their duty. Attempting to convey or interact with a Perfected will result in it teleporting away in an instant. Their exact properties and biology are still a mystery, but it should be noted that their actions are not to be interrupted at any time; given that they are what keeps the level safe, alive, and well. These ring-like creatures are exclusive to the Crimson Forest, with no other sighting of them in any other level being documented so far. Any hostile entities that attempt to traverse into the forest will be met by the sheer power of the Perfected, who will disintegrate the entities in mere seconds; a power they possess that is far too dangerous, far too powerful, and far too sacred to be explained. Hidden in the depths of the forest, engraved on an enormous tree in a relatively open area and underneath the leaves of the surrounding trees, a note can be found about these entities.

This design, something extraordinary. Forged with silver, fire, and the soul of the sacred many. Volatile, fast, succinct creatures; aerodynamics to die for. A floating ring, a perfected circumference. A dance of the sun and the moon, a dance of both ends of the spectrum: the light and the dark. This, dear wanderer, is no creature to underestimate. These are the Perfected.

Bases, Outposts, and Communities

Rosewood & Independent Cabins

Rosewood is a community of cabins grouped along the banks of Bigrock Lake. Home of the primary outpost of the United Azeutian Empire. for the entire level, it acts as a sanctuary and a safe zone for those traveling in or out.

Most of the community has established itself with the premises of a day-to-day living, but they are also a huge benefactor for the U.A.E., sourcing seed cultivation, soil sampling, mineral panning, level research, and deep forest surveying; as well as the manufacturing of goods, such as preserved foods and natural medicines.

Shipments of goods from the forest are highly celebrated by the receiving communities—their traditional pickled eggs are treated as an especially precious delicacy.

Not every wanderer resides in Rosewood, with many having traveled deep into the thicket in order to seek out their own, quiet life, presumably in a cabin of their own. Some live by themselves, venturing on their own and surviving based on the local fauna and flora, but small groups also exist deep inside the forest. These cabins are scattered all across the forest, but some are located relatively close to Rosewood, with some even in hiking distance within the village.

Those who reside outside of Rosewood are dubbed "Lost". However, the cabins on the outer skirts of the village are not deemed far enough to be classified as "Lost", and therefore, approximately 40 individuals reside in enough proximity not to be classed as such.

Cabins of the Lost

It should be noted that some members of the U.A.E who have traveled to the Crimson Forest disappear deep into the woods, only to get lost entirely—presumably hoping to forget the troublesome, cruel, and eerie nature of the other levels.

Under the watchful care of the Perfected, it comes with no surprise that this level has gained the reputation of being a utopia: a mythical, serene, mysterious yet familiar land with promises of a calm and safe life, and since its discovery, it is estimated that no fewer than 200 individuals, who were once aligned with the U.A.E., have disappeared into the beguiling world of the forest, only to never return. Field research teams, exploration parties, and search parties have reported sightings of humans with crimson-tinted–skin and pale, milky eyes. These strange sightings—which have not been backed by any photographic or reliable document—have been deduced to nothing but a rumor. Settlers of Rosewood, however, believe that those who wander into the woods become changed by the forest. The souls of the many believers, seekers of a precious life among the trees, become "Clay Men", empty husks of their former selves, who do nothing but wander the woods silently until they too become one with the woods.

Entrances and Exits
The only known entrance and exit to the Crimson Forest is through Level 9.1. One has to traverse through the dangerous expanse and locate the barrier between the levels. Recently found, wander off too much in Level 370 to get here. Find an forest in Level 11.11 to get here. Locating the barrier can be a hit-or-miss, as no reliable way of finding it has yet been found or documented, but one can assume they are in close proximity to the forest if they spot any sharp, bright white lights in the distance: the Perfected. As soon as spotted, it is recommended to run towards them but not interact with them; under their presence, no wanderer is prone to any attacks from dangerous entities, and they are safe to walk into the forest without having to look back. When exiting, one should remind themselves that they will no longer be under the watch of the Perfected, shattering all immunity against the dangerous entities of Level 9.1. If in search of Wi-Fi, the settlers of Rosewood have marked a commonly-used trail through the expanse of the Crimson Field, with large posts flagged with cloth in various colors. Should you find yourself on this path, it is safe to assume that you will be safe during your journey, as these paths are mostly devoid of any hostile encounters.
Entrances
- Going to the outside of Level 250 leads here.
- Wandering into the forest of Level -2140 supposedly leads here.
Exits
- Wander back to get to Level 370
- Another exit just found, walking deep through the forest will send you to Level 280
- If you enter a Black cabin i heard you get to A Old Level...
- No-clipping through a tree Might Lead to Level 9.666 Or Bloody Crimson Fields
- Reaching the very heart of the ominous red glow at the end of the grass field leads to Level 3778.

2.1.3 Challenges in automatic generation of Backrooms levels

When it comes to automatic generation of Backroom levels, there are a lot of challenges to be faced. The more objective ones are narrative coherence, syntax and clarity of text. These are highly imaginative long texts of a very descriptive nature that contain element of a broad fictional world with its own rules and 'logic' (Stephen 2022). Obtaining as model which can maintain syntactic and semantic coherence in such a task, while respecting the lore of a fictional world, is a very difficult task which requires gathering lot of good quality data.

The subjective challenges are even harder to achieve, considering the difficulty of even defining them. Some of the most important components of a Backrooms level descriptions are the creativity, originality and imaginative nature they contain. But if defining what is creativity is

already difficult, teaching it to a model and then evaluating how good it became at it becomes an almost impossible task, especially considering how something may appear original to a person, but completely unoriginal to another. These were the main challenges faced and discussed in this work.

2.2 Transformers

The Vanilla Transformer architecture (Vaswani et al. 2017) is a recent and very famous deep learning model that has been widely adopted in various AI fields, such as natural language processing (NLP), computer vision (CV) and speech processing (Lin et al. 2022). It consists of a sequence of encoder and decoder blocks.

An encoder block is composed of a stack of N identical layers. Each layer has two sub-layers. The first is a multi-head self-attention mechanism, while the second is a simple, position-wise fully connected feed-forward network. Residual connections are applied around each of the two sub-layers, followed by layer normalization (Vaswani et al. 2017).

A decoder block is also composed of a stack of N identical layers, but, in addition to the two sub-layers that are also present in the encoder layer, the decoder inserts a third sub-layer, which performs multi-head attention over the output of the encoder stack with the output of the masked multi-head self-attention layer present inside each decoder layer. Similar to the encoder, residual connections are applied around each of the sub-layers, followed by layer normalization (Vaswani et al. 2017).

The attention mechanism, essential to the Transformer architecture, is a query–key–value (QKV) model. The dot products of the query with all keys are computed, divided each by d_k, which is the dimension of the keys, and, then, the softmax function is applied to obtain the weights on the values. The keys, queries and values are matrix representations created with the embedded inputs of the network and the scaled dot-product attention used by Transformer is given by Equation 1 (Vaswani et al. 2017), (Lin et al. 2022):

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V. \tag{1}$$

Instead of simply applying a single attention function, the Transformer architecture uses multi-head attention, in which the d_m-dimensional original queries, keys and values are projected into d_k, d_k and d_v dimensions, respectively, generating H different sets of learned projections. For each one of the projected queries, keys and values, an output is computed according to Equation 1. The model then concatenates all the outputs and projects them back to a d_k-dimensional representation (Lin et al. 2022).

There are three types of attention used in the complete Transformers architecture (Vaswani et al. 2017), (Lin et al. 2022):

- ❏ Self-attention: in a self-attention layer, all of the keys, values and queries come from the same place, which, in this case, is the output of the previous layer in the encoder. Each position in the encoder can attend to all positions in the previous layer of the encoder;
- ❏ Masked Self-attention: similarly to the self-attention layers, in the decoder there are Masked Self-attention layers, which allow each position in the decoder to attend to all positions in the decoder up to a certain point. The leftward information flow needs to be blocked in the decoder to preserve the auto-regressive property. This is implemented inside of the scaled dot-product attention with a mask that masks all values (setting to -∞) in the input of the softmax which correspond to illegal connections;
- ❏ Cross-attention: in the Cross-attention layers, the queries are projected from the outputs of the previous (decoder) layer, whereas the keys and values are projected using the outputs of the encoder.

The full classic Transformers architecture can be seen in Figure 2. There are two main variations of the Vanilla Transformer architecture: the encoder-only Transformer and the decoder-only Transformer. The encoder-only architecture only consists of a stack of encoder layers. These models usually process the input text in a bidirectional manner, which means that they consider the context from both the left and right sides of a token within the input sequence. This allows the model to capture a thorough grasp of the context surrounding each word. Encoder-only models are usually used in tasks which involve understanding the entire relationship between the tokens on an input sequence, such as classification, sentimental analysis and summary (Liu 2024). A famous example of an encoder-only model is BERT (Devlin et al. 2019).

On the other hand, decoder-only models generally consist of multiple layers Transformer decoder blocks stacked on top of each other. In contrast to encoder models, decoder-only models typically generate text in an unidirectional or auto-regressive manner, which means that each token is generated based on the previously generated tokens without seeing future tokens in the sequence. This auto-regressive nature allows this kind of model to be used in generative tasks, like question-answer and any sort of task which involves generating new text based on the previous tokens inputted to the network. The most famous example of a decoder-only model is OpenAI's GPT (Radford et al. 2019), (Liu 2024). In this work, considering the generative nature of the proposed task, which is generating new backrooms level descriptions, decoder-only architectures will be used.

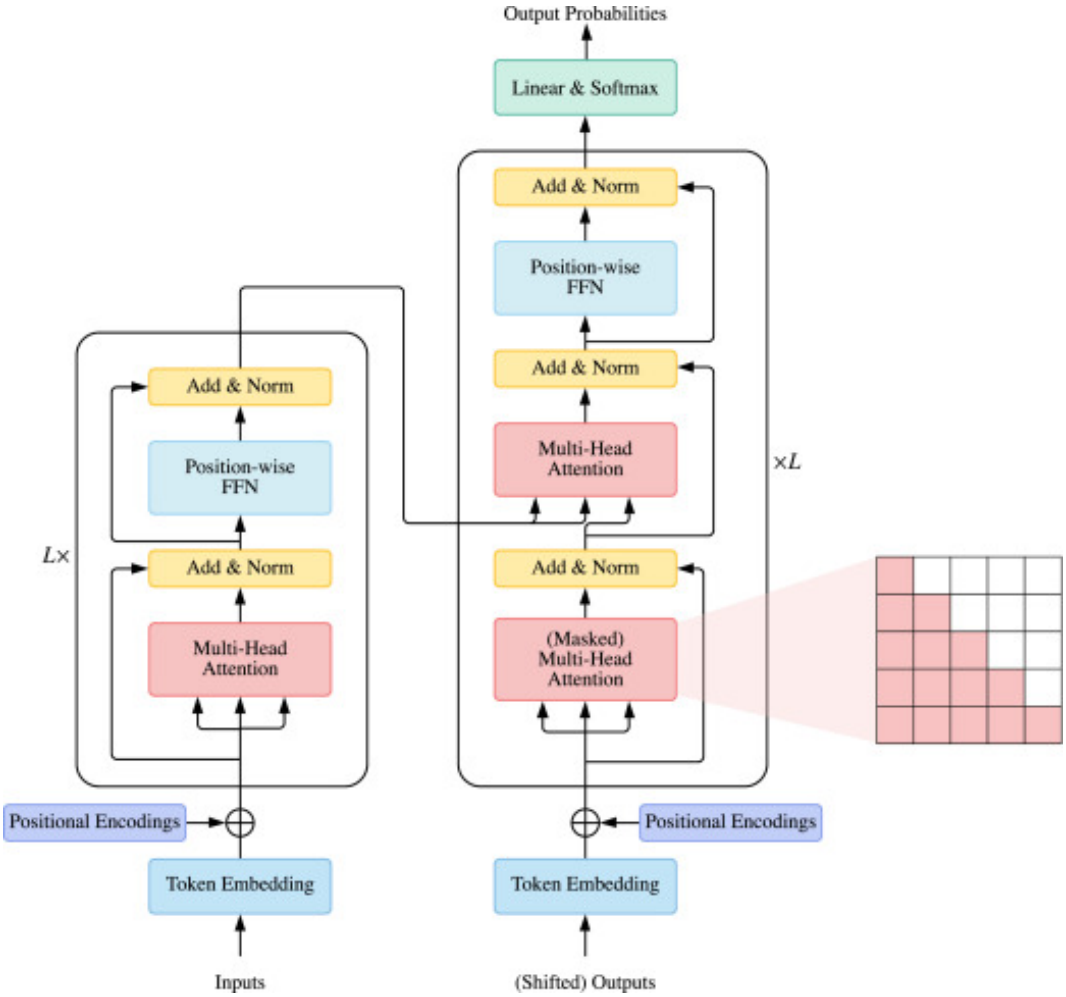

Figure 2 – Vanilla Transformers architecture extracted from (Lin et al. 2022).

2.3 Large Language Models and Small Language Models

Typically, Large Language Models (LLMs) are Transformer language models containing hundreds of billions (or more) of parameters, which are trained on massive text data that usually approaches a very large variety of subjects. A few examples of very famous LLMs are GPT (Radford et al. 2019), PaLM (Chowdhery et al. 2023), Galactica (Taylor et al. 2022) and LLaMA (Touvron et al. 2023). LLMs are known to show strong capacities to understand natural language and solve complex and diverse textual tasks (Zhao et al. 2024). A timeline of the most famous and influential LLMs can be found in Figure 3.

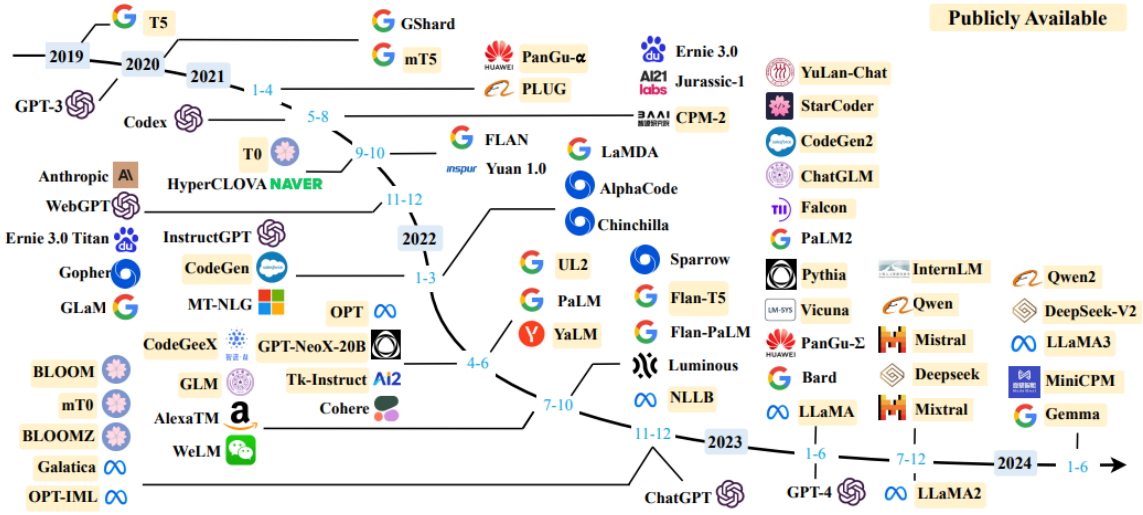

Figure 3 – LLMs timeline extracted from (Zhao et al. 2024).

Another term that is frequently used in the literature is Small Language Models (SLMs). There is no official, globally agreed-upon definition separating LLM from SLM. Neither term has a formal threshold from academia, but, popularly speaking, a LLM can be defined as a model containing billions to hundreds of billions of parameters, with a deep, wide transformer architecture trained on multi-trillion-token corpora, which requires GPUs/TPUs to train and run efficiently, capable of advanced reasoning. An SLM, on the other hand, is popularly defined as a model with millions to low billions of parameters, with a narrower architecture , trained on smaller corpora, optimized for edge devices, on-device inference, low compute use, or domain-specific tasks. It is typically less capable in general reasoning, but often very strong in targeted tasks. A few famous SLMs are TinyLlama (Zhang et al. 2024), Gemma 2B (Riviere et al. 2024) and GPT-Neo (26).

When it comes to training a Language Model, be it a small or large one, there are two main steps involved: Pre-training and Post-training. Each of these processes is described below, followed by a section which describes the process of evaluating a trained LLM.

2.3.1 Pre-training a Language Model

Pre-training is the process which builds the backbone of a LM. It establishes the basic abilities and knowledge of a model. By pre-training a model on large-scale corpora, it can acquire essential language understanding and generation skills. In this process, the size, diversity and quality of the pre-training corpus are critical for the model to obtain basic powerful capabilities later on. Collecting a large amount of natural language corpus from various data sources is essential to build this pre-training corpus. Most existing LMs mainly leverage a mixture of diverse public textual datasets as the pre-training corpus (Zhao et al. 2024).

The data used in a pre-training corpus can be roughly divided into two categories: general data, which consists of webpages, books, and conversational text and specialized data, which consists, for example, of multilingual text, scientific text and code. General data not only has

an accessible nature, but also enhances the generalization abilities of a LLM, while specialized data powers a LLM with specific task-solving capabilities (Zhao et al. 2024).

Pre-training plays a key role in the process of encoding general knowledge from the large-scale corpus into the massive model parameters. In this pre-training process, the main used pre-training task is called Language Modeling. Given a sequence of tokens (a token is a small unit of text, such as a word, part of a word, or even just a character, depending on the tokenizing strategy, that a model uses to understand and process language) $x = x_1, ..., x_n$, the model aims to auto-regressively predict the target tokens x_i based on the preceding tokens $x < i$ in a sequence (Zhao et al. 2024).

Once a LM is pre-trained, it becomes a 'most-probable-next-token generator'. This means that the model is capable of, based on a text input, generating the tokens that most probably follow this input. It is important to highlight that, at this point, the model hasn't learned any specific task (like question-answer or sentiment analysis) and is only capable of completing the input text with what it thinks are the most appropriate next tokens. Any sort of specific task training shall be performed in the post-training stage, which is described below.

2.3.2 Post-training a Language Model

After pre-training, LMs can be further trained with the aim of having a model fit to perform a more specific generative task. This is part of what is called the post-training stage. The main approach when it comes to post-training LMs is called instruction tuning. In essence, instruction tuning is the approach to fine-tune a pre-trained LM in a supervised way on a collection of natural language formatted instances. A fine-tuning dataset will contain instances formed by an input and a corresponding target output, which should represent the nature of the desired task (Zhao et al. 2024). A few examples of famous fine-tuning tasks in LMs are question-answer, sentiment-analysis, spam detection (Devlin et al. 2019), topic categorization, creative writing, dialogue generation, content summarization (Radford et al. 2019), named entity recognition, language translation (Raffel et al. 2020) and code generation (Yin e Neubig 2017).

When it comes to fine-tuning a LM, there are two very important therms that are often mentioned: Distillation Knowledge and Catastrophic Forgetting.

Catastrophic Forgetting is the abrupt and severe loss of previously acquired knowledge that occurs when a neural network is trained on new tasks or datasets without access to the earlier data. Its when a model forgets previously learned information when it is presented with new data. Since neural networks learn by adjusting their weights, when training happens sequentially, the new data can cause weight updates that overwrite, distort or interfere with representations learned earlier. In summary, since the model has no reason to "protect" old knowledge, learning a new task ends up destroying the internal structure needed for older tasks (McCloskey e Cohen 1989).

Distillation Knowledge is a training method in which a smaller model (the student) is trained to replicate the behavior of a larger or more complex model (the teacher) by minimizing a loss

function that matches the teacher's softened probability distribution over classes. This softened distribution captures knowledge, enabling the student to learn richer decision boundaries than from hard labels alone. Formally, Knowledge distillation is the process of transferring the function learned by a high-capacity model to a smaller model by matching the softened output distributions produced by the teacher at an elevated temperature (21).

2.3.3 Evaluating a Language Model

With the rapid progress of LMs and the NLP field in general, the need for reliable evaluation methods has become increasingly critical. Evaluating these large models is very important to ensure they meet quality standards, align with human expectations and maintain safety and reliability in various applications and scenarios (Chang et al. 2024), (Badshah e Sajjad 2024).

Famous automated metrics such as BLEU (Papineni et al. 2002), (Lin 2004), METEOR (Banerjee e Lavie 2005), BERT score (Zhang et al. 2020), the Levenshtein distance and even the F1-score have long and frequently been employed to evaluate the performance of model generated text. Nevertheless, these metrics primarily focus on token and surface-form similarity and often fail to identify similarities in semantically equivalent outputs which don't use the exact or almost exact words. In addition to that, automated metrics struggle in evaluating creative and subjective generation or free-form text, in which a wide range of acceptable outputs exists. Though benchmarks such as MMLU (Hendrycks et al. 2021) evaluate if models generate controlled outputs, they fall short in assessing the complexity and variability of open-ended generation. This limitation becomes particularly clear in instruction-tuned chat models (Badshah e Sajjad 2024).

Another interesting point to observe is that the correlation between these automated metrics and human evaluation is relatively weak. That is exactly why human evaluation in the evaluation process of LMs can play a crucial role. Humans are usually better at assessing aspects that automated metrics often miss, such as coherence and contextual relevance (Badshah e Sajjad 2024). Also, when it comes to subjective evaluation, such as originality and creativity, human evaluation can bring many more interesting points to the table. That is why many works involving text generation use humans as evaluators (... 2024), (Lanzi e Loiacono 2023).

While human evaluation can still be considered the "gold standard" for evaluating the quality of generated text, it has several limitations, such as being financially demanding and time-consuming. These limitations highlight the need for an automated evaluation method that aligns closely with human judgments while being more automatic, efficient and scalable (Badshah e Sajjad 2024). In this light, many works have started using LLMs themselves as judges and shown promising results (Badshah e Sajjad 2024), (Tan et al. 2024). The main idea of this strategy is to develop instruction prompts, which will be passed to one or more LLMs along with the generated text, where the evaluation criteria and desired output format for the evaluation are described to the judge model.

2.4 Procedural Content Generation

Procedural Content Generation (PCG) refers to the algorithmic creation of game content with limited or no human input (47). It allows games (or simulations) to produce multiple, diverse and playable aspects of a game without manually designing every piece. PCG can be used to create a variety of game-related content, but it is manly used to create art assets, levels and maps, textures and graphics, creatures and characters, game mechanics, dialogue, lore, story paths and music for games. The algorithms used can greatly vary depending on the content they are supposed to generate, but they can generally be categorized under these categories: (1) search-based methods, such as Monte Carlo Tree Search (MCTS) (Browne et al. 2012), which focus mainly on optimizations; (2) learning-based methods, including traditional machine learning and deep learning (DL), such as generative adversarial networks (GAN) and reinforcement learning (RL) (18); (3) other methods, such as noise functions and generative grammars (Perlin 1985); (4) the newest and most popular player on the team, LLMs (Vaswani et al. 2017) (47).

Although The Backrooms isn't a game (even if there are Backrooms-related games), the research proposed in this work is very related to PCG, especially considering the story-telling applications of it, such as creating dialogue, story-lines, quest descriptions and background for characters in games.

CAPÍTULO 3

Related Works

In this chapter, recent works that approach using Large Language Models (LLMs) for different creative generative purposes, especially in literature and in the game-making process, will be presented. Additionally, a section covering evaluation methods usually used in these kind of tasks, in which free text is being evaluated, will also be presented. The research method used to find these works consisted in using the keywords "Large Language Models", "Games", "Generative Models", "Creative Tasks", "Evaluation of Large Language Models", "Evaluation of Free Text"in a Google-Scholar search and, then, filtering the works that seemed more interesting, recent (from 2020 to 2026) and close to the proposal of this work.

3.1 LLMs in creative tasks

Studying different application of LLMs in creative generative tasks is a research field which has been very explored lately. To the knowledge of the authors of this work, there is not another study in which The Backrooms were used combined with LLMs. The closest work in the literature to the one here proposed is the one by (... 2024), in which they gathered and openly published a dataset of 978 quests and their descriptions from six RPGs. With this dataset, they leveraged two strategies to generate RPG video game quest descriptions: i) one in which they fine-tuned GPT-2; ii) one in which they just generated quest descriptions with the vanilla GPT-3 model. Not only does our work leverage these two strategies with more recent and better evaluated models (GPT-Neo and GPT 4o), but we also added a third strategy, which consists of pre-training a vanilla Decoder-only architecture with Backrooms data before performing fine-tuning on the model.

In work proposed by **??**), the authors proposed a study in which a contest was held between Patricio Pron (an awarded novelist) and GPT-4, very much in the spirit of AI-human duels such as DeepBlue vs Kasparov and AlphaGo vs Lee Sidol. As in these previous AI duels, they did not try to compare a top AI model with average humans. Instead, they focused on a one-on-one comparison between two (AI and human) top performers. The contest involved each contestant (Pron and GPT-4) initially designing thirty short story titles each. Then, both contenders wrote

synopses (approximately 600 words) for each of the 60 titles. Similarly to our proposed work, this work evaluated the performance of a famous LLM in a highly creative-driven task. But not only they didn't perform any kind of training, but also, they were more focused on comparing a famous LLM against a human professional writer, whereas our work aims to compare the performance of different LLM techniques in a a highly creative-driven task.

Similar to the previous described related work, the authors in **??**) performed two different experiments to compare model's writing and human's writing in the following task: given a potential movie title, to write an imaginary synopsis for a movie with that title. The first one consisted in evaluating how far the texts written by a fine-tuned (to the down-stream task of synopsis generation from titles) BART-large, called SLM, are from those written by humans. The second experiment consisted in qualitatively analyzing the linguistic similarities and differences between the SLM texts and GPT-3.51 and GPT-4o. Similarly to this work, they fine-tuned a smaller model than GPT for a specific task and compared it with GPT-4o, but they didn't perform any pre-training.

Another very interesting work was presented by (Todd et al. 2024), which proposed an automated game design system called GAVEL (Games via Evolution and Language Models), based on large language models and evolutionary computation, that explores the generation of novel board games descriptions in the comparatively expansive *Ludii* game description language (L-GDL) (Lehman et al. 2024). They did this by fine-tuning CodeLlama-13b (Rozi‘ere et al. 2023) to operate in L-GDL and then training this model to act as a mutation operator over programs by using a fill-in-the-middle (FITM) (Bavarian et al. 2022) training objective. Differently from our goal, they proposed a method to create full board game descriptions, instead of focusing on a more specific task inside a game or fictional universe, like creating Backroom level descriptions. Their training method was also different from the one proposed in this work, since they only fine-tuned CodeLlama-13b, while our work also proposes to pre-train a Transformer model. Additionally, they used L-GDL as training data, which resembles code language, whereas, our work uses free-form text.

The authors in (Lanzi e Loiacono 2023), also inspired by LLMs and evolutionary computation, proposed a collaborative game design framework that combines interactive evolution and Large Language Models to simulate the typical human design game-making process. They use users' feedback for selecting the most promising ideas and LLMs for the recombination and variation of ideas. In the proposed framework, the process starts with a set of candidate game design ideas (game concepts), either generated using a language model or proposed by the users. Next, users collaborate on the design process by providing feedback to an interactive genetic algorithm that selects, recombines, and mutates the most promising designs. Similar to our work, the data used was represented as free-form texts. But they didn't perform any training with this data, instead proposing this interactive evolution prompt engineering method for creating new game descriptions.

In (23), the authors also based on a type of video game description language (VGDL)

(Schaul 2013), proposes an LLM-based framework to generate game rules and levels simultaneously using mainly prompt engineering. They used LLMS such as GPT-3.5, GPT-5 and Gemma 7B to generate *Maze* games and levels in VGDL. Similar to (Todd et al. 2024), (23) used LLMs to generate games using some sort of game description language, thou the used method was simpler, only consisting on finding the best prompts possible, without any training, as opposed to this work.

3.2 Different methods of applying LMs for specific tasks

With the rapid evolution of the NLP field, many different methods ans strategies of applying LMs for specific tasks have been studied, explored and compared. The authors in (Yao et al. 2019) proposed a plan-and-write hierarchical generation framework to explore open-domain story generation, in which the framework writes stories given a title as input by first planning a storyline and, then, generating a story based on the storyline. For this, two methods that use neural generation models were adopted: a Dynamic Schema, in which a bidirectional gated recurrent unit (BiGRU) first encodes context into a vector and then incorporates the auxiliary information, in this case the previous word in the storyline, into the decoding process. Then, a story sentence is generated based on both the context and an additional storyline word as cue; a Static Schema, where a whole storyline, which does not change during story writing, is first generated with a sequence-to-sequence (Seq2Seq) conditional generation model that first encodes the title into a vector using a bidirectional long short-term memory network (BiLSTM), and then generates words in the storyline using another single-directional LSTM. Based on that generated storyline, the full story is created. A Seq2Seq model that encodes both the title and the planned storyline into a low-dimensional vector by first concatenating them with a special symbol <EOT> in between, and encode them with BiLSTMs was trained for that.

In (Gururangan et al. 2020) the authors investigate if the latest large pretrained models work universally or if it is still helpful to build separate pretrained models for specific domains. This was done by continuing the pretraining process of the famous model RoBERTa on a large corpus of unlabeled domain-specific text on four different areas (biomedical and computer science publications, news, and reviews) and comparing the results with the performance of the original model.

In (Shin et al. 2025), they evaluated GPT-4 by using three prompt engineering strategies (basic prompting, in-context learning, and task specific prompting) and compared it against 17 finetuned models across three code-related tasks: code summarization, generation and translation.

??) took a different approach where they introduce something they call Trace-of-Thought Prompting, a novel framework, which is an intersection between prompt engineering and knowledge distillation, designed to distill critical reasoning capabilities from large teacher models to much smaller student models. This approach utilizes in-context learning (ICL) to emulate

traditional distillation knowledge within the accessible framework of LLM prompting. The proposed technique was tested on the GSM8K and MATH datasets.

This work investigates and compare the results of three of the techniques mentioned in the works above: prompt engineering, finetuning and pretraining. The work proposed by (Yao et al. 2019) consists of a different, less recent approach, but it is still interesting, specially considering the creative story-telling aspect of it, which very much relates to our work. Even though knowledge distillation wasn't explored in this work, like in (McDonald e Emami 2024), it is still a very promising method when it comes to training a small model with the goal of it being capable of competing with a much larger one in a specific task. It would be very interesting to apply it in the Backrooms context in future works.

3.3 Evaluation of LLMs in Free-Form Text Tasks

Evaluating LLMs can be quite tricky, specially if the evaluated output consists of Free-Form Text, which is the case in the proposed work. Often, in this kind of work, humans are used as evaluators. This is the case for (... 2024), which used humans as evaluators of the generated RPG quests. They created a randomized mixed design user study in the form of an online questionnaire in which participants were presented with quests and asked to rate corresponding quest descriptions. The participants were recruited from various RPG sub-communities on Reddit and r/SampleSize. The study was advertised toward everyone aged over 18 years with RPG playing experience. In the work proposed by(Lanzi e Loiacono 2023), the authors also used human subjects (junior and senior game designers and people participating in the 2023 Global Game Jam) to work on and evaluate their proposed game design framework. In (Marco et al. 2024), three experts in pedagogy, psychometrics, literature and NLP designed a rubric design (a table of criteria and standards for a task) to evaluate the synopsis written by the author and the ones written by GPT-4 encompassing quality dimensions like Attractiveness, Originality and Creativity. Then, six literary experts, all different from those who developed the rubric actually evaluated the synopsis with the rubric design. Similarly, the study presented by (34) also had human evaluators comparing synopsis written by the SLM and GPT and texts written by humans by answering two quizzes for each experiment. The evaluators consisted of 68 volunteers, all students in an international MBA master's program. This is quite similar to one of the evaluation methods proposed to evaluate our work, which consists of using human judges to evaluate the generated Backrooms levels.

Another strategy that is currently being used in evaluating LLMs' outputs is using LLMs themselves as judges. This is exactly what (Badshah e Sajjad 2024) proposed: a reference-guided verdict method that automates the evaluation process of Free-Form Text by leveraging multiple LLMs-as-judges. By performing experiments on three open-ended question-answering tasks, they demonstrated that combining multiple LLMs-as-judges significantly improves the reliability and accuracy of evaluations, particularly in complex tasks, like generating text. They

also show that the LLMs evaluation reveal a strong correlation with human evaluations, which indicates that this is a viable and effective alternative to traditionally used metrics and human judgments, particularly in the context of LLM-based chat assistants, where the complexity and diversity of responses challenge existing benchmarks. LLMs as evaluators will also be used in the proposed work.

CAPÍTULO 4

Backrooms Level Description Generation

Unfortunately, the impressive power of the larger LLMs available for textual generative tasks in the most varied types of specific domains is not usually accessible to a large part of the scientific community and people in general, since they require large and expensive hardware to be executed or involve paying for expensive plans. In this light, the main objective of this work is to propose an approach that allows for generating new, smaller LLMs that can be executed and trained in more modest machine architectures, which are effective in generating texts related to specific domains. Such an approach consists of investigating adequate techniques that make it possible to improve existing smaller pretrained models, which despite running on smaller hardware, usually do not present good performance in generating structured text related to specific domains.

In this chapter, the data and experiments of this research will be described. It was divided into: section 4.1, where an overview of the proposed work will be given; section 4.2, in which not only the data used in this work, but also how this data was gathered will be described; section 4.3, in which the three proposed generation strategies will be discussed and section 4.4, which describes the evaluation methods of this work.

4.1 Overview of the Proposed Approach for Backrooms Level Generation

The main purpose of this work is to investigate the potential of custom SLMs for specific tasks when compared against the renowned Large Language Model DeepSeek. The specific task chosen for this study was Backrooms level generation. In addition to comparing model performance, this work also investigates the feasibility of building a complete end-to-end language model development pipeline, in which the entire process—from dataset creation and pretraining to task-specific fine-tuning—is carried out without relying exclusively on externally pre-trained foundation models. To this end, three experiments were developed and compared. In the first one, only the test dataset was used, and the strategy simply consisted of prompting

the test inputs to DeepSeek with the goal of generating new Backrooms level descriptions. The second experiment consisted of fine-tuning Llama 1B, using a custom-made fine-tuning dataset, for the Backrooms level description generation task. The third experiment consisted of pre-training a vanilla decoder-only architecture using a custom-made pre-training dataset and subsequently fine-tuning this pre-trained backbone on the custom fine-tuning dataset for the Backrooms level description generation task, thus demonstrating a complete end-to-end training pipeline.

The overall schema of the proposed work was described in three diagrams, where each of them, respectively, details one of the three proposed experiments: Figures 4, 5, and 6.

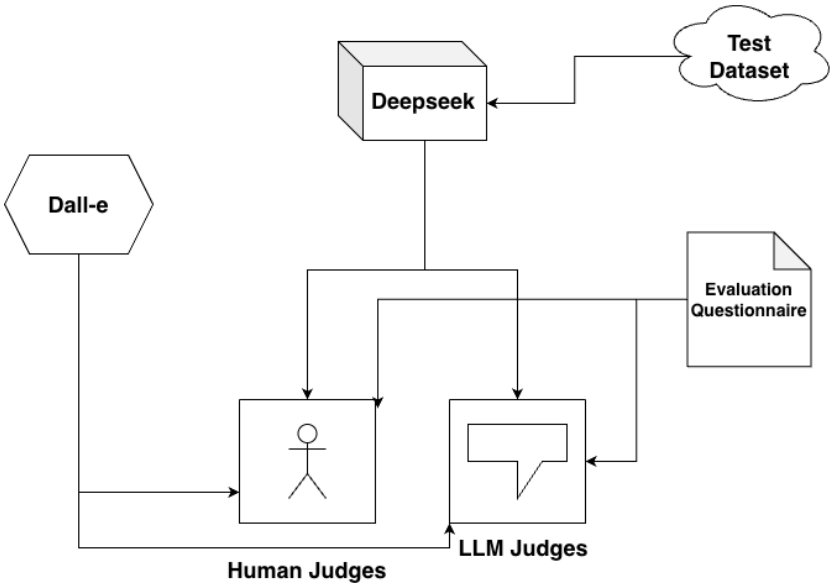

Figure 4 – Diagram of the first proposed experiment.

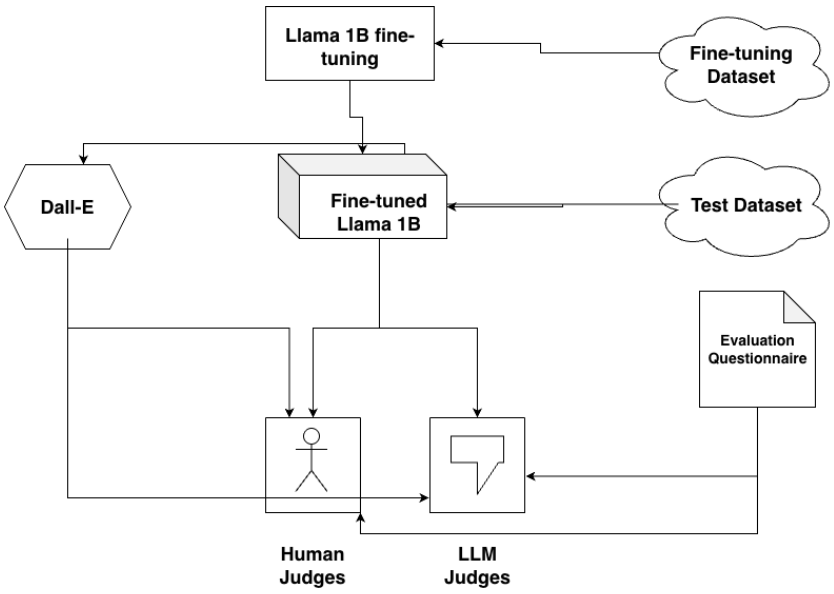

Figure 5 – Diagram of the second proposed experiment.

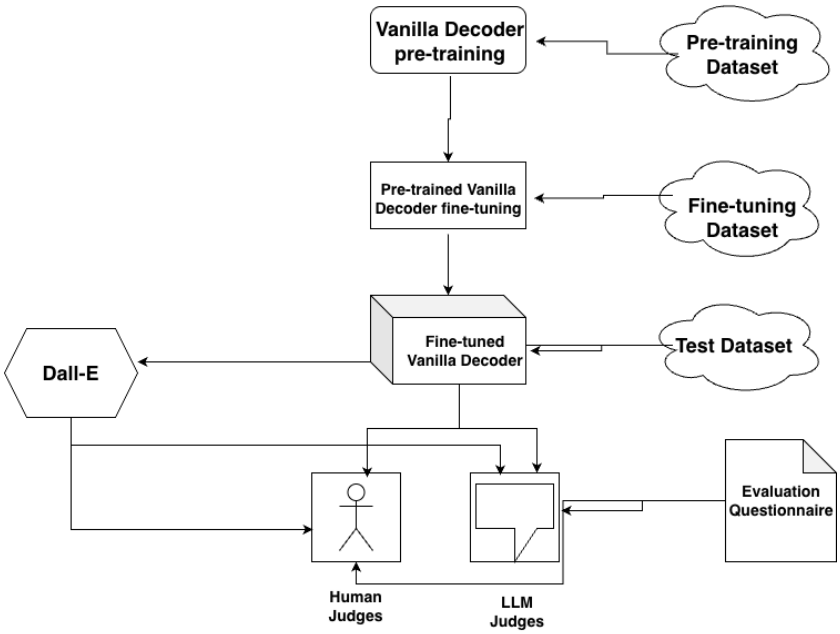

Figure 6 – Diagram of the third proposed experiment.

Considering the amount of high-quality textual data required to train a language model, especially in a rich, creative task such as creating Backrooms levels, a large portion of this work consisted of gathering and formatting the necessary data for the experiments proposed here. The structure of each dataset and how they were created will be detailed in section 4.2.

Evaluating a creative task, like creating Backrooms levels description, can be quite challenging, since originality and creativity, which are two desired aspects in the outputs of tasks like this, are very subjective metrics that can vary drastically depending on the evaluator (... 2024). In this light, three evaluation strategies were used: In the first one, ChatGPT was used as a judge; the second consisted of human evaluation; and the third one represents a new assessment strategy in which the quality of the textual outputs of the competing models is evaluated through the compatibility between these outputs and pictures that are generated from them. These strategies were described in 4.4.

4.2 Datasets

Three datasets were created in this work: i) a Pre-training dataset; ii) a Fine Tuning dataset; iii) a Test dataset. Since the Backrooms wikis were the main source for these datasets, they will also be discussed here.

4.2.1 Backrooms Wikis

There is a lot of Backrooms text data available online, which is very convenient for creating a pre-training dataset, since pre-training a transformer backbone requires a lot of textual raw data. When it comes to Backrooms lore, there are two main online wikis: Wiki 1 and Wiki 2. They

contain a very large variety of descriptions concerning many aspects of the Backrooms, like level descriptions, entities (monsters of the Backrooms world) descriptions, object descriptions and Tales of the Backrooms. All this rich textual corpus was used to create a large portion of the pre-training dataset.

A substantial part of The Backrooms wikis are the level descriptions. There is a whole section in each wiki composed of a long list of level descriptions. A standard level description usually contains the following sections: name of the level; whether the level is safe, stable and deprived of entities or not; the textual detailed description of the level; the level entities descriptions; The Colonies and Outposts of the level; the entrances and exits of the level. These descriptions were used to build the Fine-Tuning dataset.

4.2.2 Pre-training Dataset

Training a language model from scratch, even a small one, can be a very difficult task, mainly because building a large-scale language modeling dataset requires significant effort in collecting, filtering and documenting diverse high-quality sources (Gao et al. 2020).

Initially, a large text corpus, created from the Backrooms wikis, made up the pre-training dataset. A web scraper was used to create this text corpus by extracting all raw text from the wikis. The irrelevant pages (like FAQ pages) were removed from this corpus and chatGPT 4o mini was used as a data cleaning assistant, since the scraped data contained a lot of noise. The following prompt was used for that: "You are a data cleaning assistant. Your task is to process the following text by removing all HTML tags, XML artifacts, URL remnants, CSS, encoding noise (like '#D7EPE'), Japanese characters, Arabic characters, characters and any other non-linguistic or non-textual elements. Do NOT modify, summarize, or paraphrase the actual textual content—only remove junk. Keep all meaningful sentences, paragraphs, and punctuation intact. Return only the cleaned text, without additional commentary. Additionally, remove any repetitive wiki-specific boilerplate (e.g., 'Edit this page', 'Table of Contents', 'Jump to navigation', 'Article written by','Author:', 'ADULT CONTENT', 'Click here to edit contents of this page') but preserve all factual and narrative content. Here is the text to clean: text"

The biggest challenge that surfaced in creating this dataset was the realization that, after cleaning all the raw text scraped from the Backrooms wikis, there was a lot less data than what was expected at first, which also meant that there wasn't enough data to train a small language model big enough to handle the complexity of a task like generating Backrooms levels. With only the Backrooms wiki, there was only a total of 15.497 training samples of 2.048 tokens each (31.75 million tokens) to train a 151.87M parameters model, which is very little. This kind of model, when trained from scratch, typically requires on the order of at least 10-30 billion tokens to achieve strong language modeling capacity (depending on data quality). With such an unbalanced data-to-parameter ratio, the model is dramatically underexposed to sufficient data to learn good statistical regularities of language, which leads to a very week pre-trained model.

After a few experiments it became quite clear that only the Backrooms wikis data wouldn't be enough to build a strong pretraining dataset. In the light of that, the first method of gathering more data that was explored consisted in finding four other wikis, which aren't Backrooms related, but also revolve around supernatural themes, one even being focused on liminal spaces, were scraped cleaned and added to the data corpus: Library of Babel; Liminal Archives; SCP Foundation and The Wanderers Library. Unfortunately, not only did the gathered new text was still not enough, but also, since wiki scraped raw text contains a lot of problems, all the scraped text had to be cleaned by again using chatGPT 4o mini.

After a few experiments, it became clear that the Backrooms wikis data alone would not be sufficient to build a strong pretraining dataset. So, with the goal of enlarging the dataset and enriching the model's vocabulary, two very famous open source textual dataset was also used: The first one is a part of OpenWebText-Dataset, which is a large-scale text dataset scraped from the open web, created to roughly replicate the data used to train GPT-2; and the second one is WritingPrompts, which is a creative writing dataset built from the r/WritingPrompts subreddit. A few Wikipedia pages, manly related to places descriptions and paranormal phenoms were also added to the final data corpus. A diagram summarizing all data sourced used to create the pretraining dataset was detailed in Figure 7

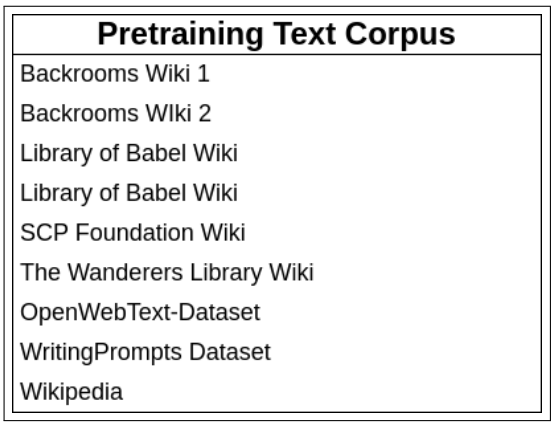

Figure 7 – Full pretraining text corpus

All data was tokenized before training with a custom trained BPE tokenizer and saved in chunks of sequence length size, so that each chunk corresponds to a model instance. This was done in order to accelerate the training process. The details of how many instances came from which source and if there was any oversampling (represented in the table with *nx*, in which *n* is the size of the oversampling) were described in Table 1. The goal of the oversampling was to try to make the model converge more toward the Backrooms/supernatural data, instead of mainly learning from the general data, which composes the largest chunk of the pretraining corpus.

4.2.3 Fine-Tuning Dataset

As mentioned above, there is an entire section in each wiki composed of multiple pages of of level descriptions. These pages were used to build the fine-tuning dataset. Each sample of the fine tuning

Table 1 – Pre-training Dataset details

Subset	Total instances	Wiki	OpenWebText	WritingPrompts
Training	1588166	107550 (10x)	1024826	455790 (5x)
Validation	5567	567	2500	2500

dataset is composed of a short description of the level as input (an idea for a level) and the full level description as the target output. This short description as input was generated by summarizing the raw level descriptions in a couple of short sentences with Llama 7B. For creating the target output, chatGPT 4o mini was used to format the raw text description of each level into a desired sample format which follows the standard Backrooms level description structure mentioned above. An example of input and corresponding target output was distributed in Figures 8, 9 and 10, respectively.

This fine tuning dataset splits were described in Table 2. Is is important to highlight that the validation set was kept small with the goal of not removing too many samples from the training set, since the full dataset is already quite small.

Table 2 – Fine-tuning dataset splits

Subset	Number of Samples	Percentage
Training	1190	88%
Validation	162	12%
Total	1352	100%

Input:

An expansive complex of interconnected rooms and corridors filled with white tiles and pools, slightly submerged in undulating, lukewarm water.

Figure 8 – Fine-tuning dataset input sample

Target Output:

Level Name: Level 37: 'Sublimity'

Safety: Safe, Unstable, Devoid of Entities

Level Description: Level 37, commonly referred to as the Poolrooms, is an expansive complex of interconnected rooms and corridors slightly submerged in undulating, lukewarm water. Each area of the level varies greatly in size and structure, ranging from uniform pools and hallways to more open, abnormally-shaped areas. The walls, ceilings, and floors of the level all appear to be constructed from the same white ceramic tile, with the only deviation from this color being the blue-green hue of the water. Such tiles are eerily pristine in condition, all identical to one another, without a single hint of damage on their shiny surfaces. The architecture of Level 37 is varied, but strict in design; all areas in the level connect to one another in a senseless manner, with none having an easily identifiable purpose. Although the intended use of each area is difficult to determine, they are much too large to properly serve the function of a pool. The abnormality of Level 37's geometry negates any purported purpose it might serve: pillars protrude out of the water in certain areas, much too large in quantity to be practical; light is cast from irregular angles, leaving some areas completely cast in darkness; and certain staircases descend directly into deep pits of water. Entrances to these deeper areas are typically found in the form of descending staircases or direct underwater holes. It is advised to avoid exploring such areas due to their inherent danger; without the proper equipment, one could easily drown in the level's depths due to their lack of lighting and unpredictable currents. Upon investigation, traces of magnesium sulfate (Epsom salt) have been discovered in the waters of Level 37. This compound, when engulfing one's body for extended periods of time, has a naturally relaxing effect, relieving muscles and alleviating bodily pains. Alongside this, traces of multiple unidentified compounds were also found present in the water; the origin and properties of these substances currently remain unknown, but due to their coexistence with the aforementioned Epsom salt, it is believed that they enhance the water's relieving effects via anomalous means. By unknown means, a minimal and constant rippling is present in all liquid surfaces within Level 37, despite them being otherwise undisturbed. It is not known whether this rippling originates from a certain point in the level or if it is a property inherent in the water itself, though some wanderers theorize the level may be so large that it has developed its own tidal forces. Sound is severely altered in Level 37. Examples include the sudden dropping-off of noises produced in the level, soundwaves not reverberating as they should, and speech feeling "muted" or "dampened". It is suspected that this primarily stems from the sound of the water, which behaves in a subtly unorthodox manner. In certain explorations, the sound of the water has been described as quiet, almost calming, whilst in other cases, it has been reported to be obtrusive, drowning out most other sounds. The cause of these effects is still undergoing research, but no concrete conclusion has been drawn thus far.

Figure 9 – Fine-tuning dataset: target output sample part 1

Entities:

Level 37 is thought to be entirely devoid of life. No encounters with entities or other wanderers have been recorded in the level as of yet; it is unknown if this results from some isolating property of Level 37, or if the level is so unimaginably large that one could not possibly come across another within it

Colonies and Outposts:

Due to a lack of general incentive, its possible isolating properties, and its terrain not easily allowing for permanent settlements, no outposts currently exist within Level 37. The level's immense size and its inability to sustain life also do not entice the creation of permanent housing, and community-building would prove difficult due to its strange auditory properties.

Entrances and Exits:

Entrances: Certain locations under the water of Level 7 can lead one to this level; Oases in Level 46 may lead to Level 37; Entering inflatable pools in Level 54 will lead one to this level, with it being more common on the first floor; Entrances to Level 37 may be found in the Exitus and Terminus areas of Level 30.

Exits: Wading through the dark, dingy areas of this level may lead one to Level 37.1, a sublevel of Level 37, typically without one taking notice; Immersing in the water of the level has a rare chance of reemerging in the pools of Level 33; Long corridors in Level 37 may rarely lead to the workrooms of Level 4; Enclosed tunnels in this level may lead one to Level 43; Certain staircases may lead to Level 34.

Figure 10 – Fine-tuning dataset: target output sample part 2

4.2.4 Test Dataset

The test dataset samples follow the same format as the inputs of the fine tuning dataset. It is a dataset composed of 161 new level ideas where 40 were generated by chatGPT 4o, 50 by DeepSeek, 50 by Gemini and 21 by a human. A few samples of the test dataset can be found in Figure 17.

Test Dataset samples:

Level 0012, also known as 'Cocktailand', contains a complex system of intertwined colorful cocktail bars. Most cocktails are harmless, but it is said that some of them can have unpredictable effects on the human body

---

Level Zhen is an infinite, very clean spa with glass walls through which you can see a beautiful, peaceful garden. Relaxing music is always playing, and the air smells like vervain, jasmine, and lemon. A very dangerous entity that looks and laughs like a child roams its hallways.

A level that recreates your memories, but twisted and wrong.

Figure 11 – Test dataset samples

4.3 Backrooms Level Description Generation Strategies

Considering the datasets described above, three strategies were explored: in the first one, only the test dataset was used and the strategy simply consisted of prompting the test inputs into Deepseek with the goal of generating new Backrooms level descriptions; The second one consisted of fine-tuning Llama 1B, using the fine-tuning dataset, for the Backrooms level description generation task; The third one consisted of pre-training a vanilla decoder-only architecture with the pre-training dataset to then fine-tune this pre-trained backbone for the Backrooms level description generation task with the fine-tuning dataset.

4.3.1 Using Deepseek

The first experiment simply consisted in passing prompts containing the test dataset samples to DeepSeek to generate level descriptions in the desired output format. The choice of DeepSeek as a competitor LLM to face the small Backrooms-Llama is due to the fact that it is a very recent and large chat model (236 billion parameters against Llama's 1.3 billion parameters) which performs really well in a large variety of tasks (DeepSeek-AI et al. 2025). It is interesting to note that ChatGPT 4o was also considered for the role. However, as it will be used as a judge in the experiments, in order to avoid any bias in the judgment process, it is better to use a different model as contestant. A prompt example of this experiment is illustrated in Figure 12 and an output example of a DeepSeek level generation is illustrated in Figure 13.

DeepSeek Prompt example:

Please create a new Backrooms level with the following description:Level 002, also known as 'Daddy Issues', consists of an art museum where every single painting portrays different dads with their children. They all seem happy, but sometimes there is something weird about the painting.
Please divide the created level in only the following sections: Level Name; Safety (Whether the level is stable and deprived of entities or not); Detailed Level Description; Entities (The level entities descriptions); Entrances and Exits; Colonies and Outposts (The Colonies and Outposts descriptions of the level)

Figure 12 – A DeepSeek prompt example

DeepSeek output example:

Level 774: "Videodrom"

Safety:

Unsafe – The level is mostly stable, but prolonged exposure to the televisions can cause psychological distress, hallucinations, and eventual mental breakdown. Hostile entities are present, though they are not immediately aggressive unless provoked.

Detailed Level Description:

Level 774 resembles an endless, dimly lit electronics store from the late 1990s, with rows upon rows of CRT televisions stacked on shelves or mounted on walls. Every screen is tuned to a different reality show, talk show, or infomercial, all playing simultaneously at varying volumes. The broadcasts appear to depict distorted versions of real people—some of whom may resemble lost wanderers. The shows often feature grotesque or surreal scenarios, such as contestants being subjected to bizarre challenges or hosts delivering unsettling monologues.

The televisions cannot be turned off, unplugged, or destroyed. Attempting to do so causes the screens to glitch violently before resuming playback, sometimes with the wanderer now appearing in the broadcast. The air hums with static, and the flickering screens cast erratic shadows, making navigation disorienting.

Hidden among the aisles are VHS tapes labeled with cryptic titles. Inserting these into certain VCRs found in the store can trigger anomalous effects, ranging from temporary glimpses into other levels to dangerous manifestations of entities.

Entities:

- **The Broadcasters** – Humanoid figures that appear on the televisions, often speaking directly to wanderers. They may offer cryptic advice or attempt to lure victims into the screens. Those who interact too closely risk being "pulled in," disappearing from the level and becoming part of the broadcasts.
- **Static Phantoms** – Semi-corporeal entities that flicker in and out of existence like TV static. They lurk in the peripheral vision and may attack if a wanderer stares at a screen for too long.
- **The Audience** – Shadowy figures seated in unseen rows, occasionally laughing or applauding in response to the broadcasts. They are passive unless a wanderer disrupts the "show," at which point they may become hostile.

Entrances and Exits:

- **Entrances:**
  - No-clipping through a television screen in Level 11 or Level 434.
  - Finding a VHS tape labeled "Videodrom" in other levels and playing it in a functional VCR.
  - Waking up inside a display model TV in Level 37.
- **Exits:**
  - Locating a "Test Screen" TV and staring at it for exactly 74 seconds transports the wanderer to Level 774.1 (a grainy, black-and-white version of the Backrooms).
  - Being "invited" by a Broadcaster to join their show may eject the wanderer into a random level, often with psychological scars.
  - Rarely, a hidden door marked "EMPLOYEES ONLY" leads to Level 8.

Colonies and Outposts:

- **The Channel Surfers** – A small group of wanderers who have adapted to the constant noise and flickering. They wear makeshift blindfolds with small slits to limit screen exposure and communicate in whispers to avoid attracting entities. They trade VHS tapes with useful (or dangerous) properties.
- **The Network** – A mysterious faction that believes the broadcasts contain hidden messages about escaping the Backrooms. They obsessively document every show, often losing members to the screens. Their base is a fortified section behind the store's checkout counters.

Level 774 is a disorienting prison of endless noise and flickering screens, where reality blurs with whatever night playing next. Stay too long, and you might just become part of the show.

Figure 13 – A DeepSeek level generation output for a test sample

4.3.2 Fine Tuning Llama 1B

The smallest, most famous, recent, pretrained and open source language models available are: GPT-2, which, even though is not that recent anymore, is frequently used in this kind of research for it's small sizes (... 2024); the smaller Llama models; as well as GPT-Neo, which is a GPT-3-like language model that comes in 125M, 1.3B, and 2.7B parameter variants (Black et al. 2021). With this in mind, in this work, three models for Backrooms level description generation were fine-tuned and investigated (by using Nvidia RTX 4070 12GB GPU). They consisted of two versions of GPT-Neo (considering the hardware limitations, only the 125M and 1.3B could be tested) and Llama 1B (1.3 billion parameters). As

the Llama 1B model performed much better than both versions of GPT-Neo investigated in preliminary empirical tests, it was selected as base-model for Backrooms-Llama. Then, this section details the fine-tuning of Llama 1B in order to produce the SLM Backrooms-LLama.

In order to train Llama 1B in the available hardware (Nvidia RTX 4070 12GB GPU), the pre-trained model had to be loaded with *bfloat16* precision. At first, the model was overfitting really fast, which makes sense since the dataset isn't that large, but by lowering the learning rate, adding a dropout of 0.1 and changing the optimizer to *'cosine_with_restarts'*, the model learned to generalize a lot better. Moreover, adding an 'end of level' tag in the end of each dataset level description really helped the model to learn when and how to end the level. Before adding the tag, very often, the model would just keep generating text for the level until it reached max tokens. An experiment with a custom tokenizer (trained with the training data) was also performed, but using Llama's tokenizer proved to be a lot better.

The parameters used in the best produced model will be detailed in Table 3. These parameters were achieved after a lot of tests and hyperparameter tuning. The main metric used to choose the best model was the validation loss. Then, some manual testing proved that the model actually was creating good Backrooms level descriptions. Once this model was obtained, just like in the DeepSeek experiment, the test dataset was fed to it in order to generate new level descriptions. An output example of this model was illustrated in Figure 14.

Table 3 – Backrooms-Llama fine-tuning configuration

Parameter	**Value**
Batch Size	1
Gradient Accumulation Steps	6
Effective Batch Size	6
Learning Rate	5×10^{-6}
Number of Epochs	74
Optimizer	Cosine with Restarts
Dropout	0.1
Final Validation Loss	1.91
Final Validation Perplexity	6.75

LLama 1B fine tuned model output example:

Backrooms level idea: Level 774, also known as 'Videodrom', is a TV store where all the televisions are playing some kind of reality show or talk show. You can't turn them off.:
Level Name: Level -774: Videodrome
Safety: Unstable and dangerous due to the presence of entities.
Detailed Level Description:
Level -74 is characterized by a vast TV showroom that spans an infinite expanse. The walls are covered in a thick layer of dust, and the air is thick with the smell of stale popcorn. As one enters, they are greeted by an endless array of telephones, each displaying a different reality TV show, ranging from talk shows to reality shows. These telecasts are often accompanied by advertisements for various products, including food, clothing, electronics, etc. However, the shows themselves are not real, as they do not depict any actual events. Instead, viewers are presented with fabricated scenarios that may include characters from the show's cast, but they never interact with each other. This creates a sense of isolation and confusion, leading to feelings of anxiety and paranoia among those who enter.
The level is filled with a constant hum of static, which can be disorienting and overwhelming. It is advised to avoid looking directly at the screens, especially when they begin to flicker, because this can lead to a disorientation that can last for hours.
Additionally, it is important to remain calm and avoid engaging with any of the characters on the screen, particularly if they appear to be in distress. If a character does appear distressed, one should refrain from intervening and instead wait for help to arrive.
It is crucial to remember that this level exists solely as a simulation, with no actual entities present. While the displays may appear real and familiar, there is no physical connection between the viewer and any characters depicted on them. Attempting to interact or engage with these characters can result in severe consequences, such as being sucked into the TV screen or being transported to another reality.
Additionally, a strange phenomenon occurs when a viewer attempts to leave the level. They may find themselves in an empty room, seemingly trapped, unable to move or communicate. To escape, individuals must find a way to exit the room and re-enter the main area of Level-74. Once inside, those attempting to escape may experience a sensation of being pulled back into reality, although this is not guaranteed.
Overall, Level -74 presents a surreal and disconcerting environment that is both visually appealing and psychologically unsettling. Its primary purpose is to induce feelings among its visitors, creating a state of confusion and anxiety that mirrors the experiences of those trapped within. Those who find it difficult to navigate this environment should seek assistance from others or seek refuge in the safety of other levels.

Entities:
- **The Host:** A male character who appears on all telecast screens. He is the host of all reality and talk-shows, appearing as an actor in various roles. His voice is distorted and his presence is unsettling, often leading viewers to feel anxious or paranoid.
- The Host's Wife: A female character that appears alongside The Husband. She is also an actress, portraying various characters in different roles, contributing to an unsettling atmosphere.

Figure 14 – A level generation output of the fine tuned Llama 1B (Backrooms-Llama) for a test sample

4.3.3 Pre-training and Fine Tuning a vanilla Decoder-only Transformer

The last experiment of this work consisted in implementing a vanilla Decoder-only architecture and training it from scratch with the custom pretraining dataset described in section 4.2.2. All model parameters were detailed in table 4, while training parameters of the best model were detailed in Table 5. Considering this configuration, this model has approximately 190.6M parameters in total.

The hardware used was also a Nvidia RTX 4070 12GB GPU and, just to train a little more than one epoch it took around 5 days.

Table 4 – Vanilla Decoder-only model parameters

Parameter	**Value**
Number of layers	16
Number of attention heads	16
Embedding size	768
Vocabulary size	50257
Sequence length	2048

Table 5 – Pretraining configuration

Parameter	**Value**
Batch Size	2
Gradient Accumulation Steps	32
Effective Batch Size	64
Learning Rate	5×10^{-4}
Number of Epochs	1
Optimizer	Cosine with Restarts
Dropout	0
Final Validation Loss	3.12

After verifying that this custom pre-trained model was actually able to generate consistent text through manual testing, it was then fine-tuned with the fine-tune dataset, which is the same one used to train Backrooms-Llama. The same training parameters from Backrooms-Llama were also used and all fine-tuning details can be found in 6. This model was called and will be referred to as Tiny Backrooms from now on and an output example of it was illustrated in Figure 15.

Table 6 – Fine-tuning configuration

Parameter	**Value**
Batch Size	2
Gradient Accumulation Steps	3
Effective Batch Size	6
Learning Rate	5×10^{-6}
Number of Epochs	70
Optimizer	Cosine with Restarts
Dropout	0.1
Final Validation Loss	2.27
Final Validation Perplexity	9.66

Tiny Backrooms output example:

level name: this store resembles a fancy house, adorned with gold accents and gold decor. the room is equipped with a variety of furniture, including an old fashioned chair and fireplace. a table and chairs can be found within the store, while a wooden table is present. however, there are a few areas known as 'the living room's hallway' where the televisions are playing something. a door labeled 'the living room hallway' leads to the storage area, a small hallway with a kitchen. if a wanderer does not enter through the door, they will fall unconscious and wake up on level 71. additionally, the bedrooms consist of various types of video games available, while the living room is furnished with some standard and personal items.

the laundry room this laundry room is furnished with a variety of appliances, including a small closet and a wardrobe. the laundry room is furnished with various items from various brand staples. the laundry room does not have a functional washing machine and is devoid of any employees. the bathroom is also devoid of entities.

entities: there are no known entities residing in this level.

detailed level description:

level 774 is characterized by an infinite number of rooms that vary in size and shape, often featuring only a handful of rooms.

at the bottom of the laundry room lies a door marked 'level 69'. upon entering, every single individual inside the washing machine will experience a different set of events, including being forcibly evicted and being forcibly evicted from the level. while the washing machine is typically present, it is not consistently present. the laundry room has the same layout, with the same walls and mirrors, although the room is completely devoid of appliances. a door labeled 'the living room' leads to the storage area, a spacious living room, and an empty kitchen. however, no-clipping into the laundry room or living room may lead to level 774. additionally, a door labeled 'videodrom' can also be found in the living room and kitchen.

entrances and exits:

- **entrances:** level 774 can be accessed through various doors located in various levels.
- **exits:** the only known exit is the washing machine door to reach level 70.
- **exits:** the only known exit from level 774 is to go through the washing machine in the living room. the washing machine is only present during the washing. however, this exit is not infinite, as

wanderers are unable to encounter another individual until either exiting level 69 or returning to the original location.

colonies and outposts:

there are no known colonies or outposts within level 774. however, it is rumored that one wanderer may accidentally enter by falling asleep on the washing machine to wake up at level 71.

Figure 15 – A Tiny Backrooms level generation output for a test sample

4.4 Evaluation Strategies

As mentioned above, three evaluation strategies were used: chatGPT as a judge; Human Evaluation; and a new evaluation strategy based on pictures generated with the level descriptions.

4.5 LLMs as judges

With the goal of better understanding each model's strengths, weakness and peculiarities with an automated evaluation method, chatGPT 4o mini was used as judge. This choice was made because, further than offering a paid API that allows for automating the evaluation intended here and being frequently and successfully used for evaluation purposes (Wang et al. 2023), ChatGPT is pretty familiar with The Backrooms universe and with the standard structure of level descriptions.

It is important to note that LLM-based evaluations are inherently influenced by the context provided to the evaluator. Besides the generated level description itself, the instructions, evaluation criteria, prompt design, and any additional contextual information supplied during evaluation may affect the model's judgments. In this work, the original level idea was intentionally provided together with the generated description so that the evaluator could assess how well the output matched the requested concept. Although this additional context enables a more informed evaluation, it may also influence the scores assigned by the LLM, representing an inherent characteristic and limitation of LLM-as-a-judge methodologies.

A questionnaire following on a 4-point Likert scale (Strongly Disagree—Strongly Agree) was created, in which four of the affirmations are of a more objective nature and the last two are of a subjective nature. With each grade, a small justification was required. The prompt which accompanied this questionnaire was: *"Evaluate this Backrooms level description, based on a given idea, by answering the following questionnaire. Grade each one of the affirmations in the questionnaire with one of these grades and offer a brief justification for the grade."*. The affirmations of the questionnaire are detailed below:

- ❏ The syntax of the text is correct. It doesn't contain any grammatical errors;
- ❏ The text is clear, which means the ideas presented in it are well described and easy to understand;
- ❏ The level description is consistent with what was asked in the idea;
- ❏ The description follows the standard Backrooms level structure;
- ❏ The level description suits the Backrooms universe and seems like an actual Backrooms level;
- ❏ The level description is original and creative. It contains new and interesting ideas;

For each input of the Test dataset, each output of the two strategies was passed accompanied by the questionnaire, the input itself (the level idea) and instructions about how the evaluation response should be formatted (a *.json* file containing the grade of each statement and the respective justification for the grade).

4.6 Human Evaluation

To enrich the evaluation process and not depend solely on LLM judges, a group of six human judges, who are familiar with The Backrooms universe, performed evaluation on a subset of the test set. For that, 30 random level ideas were selected and the generations from Backrooms-Llama and DeepSeek for these levels were evaluated (totalizing 60 evaluated levels - 30 from Backrooms-Llama and 30 from DeepSeek).

Each human judge evaluated 10 levels. In order to avoid judges evaluating the same idea, which could become repetitive and boring for them and affect their judgment, each judge evaluated levels from either only Backrooms-Llama or only DeepSeek. To avoid any kind of bias, they were not told which model they were evaluating. For each level they had to answer the same 4-point Likert scale questionnaire used in section 4.5 and offer a brief written justification for the given grades.

4.7 Evaluation trough generated pictures

This section proposes a new evaluation method based on images for evaluating LLMs in descriptive tasks, such as generating Backroom level descriptions. In the context of this research, this was done by randomly selecting 30 samples from the the test dataset and, with the output of each of the two strategies for these samples, generating images of the Backrooms level descriptions with the deep-learning text-image model Dall-E.

Each image was then passed to chatGPT 4o, which accepts both text and pictures in the same prompt, accompanied by the corresponding level description an another 4-point Likert scale, with the following prompt: "*Evaluate this Backrooms level description and the picture that was generated based on it, by answering the following questionnaire using a 4-point Likert scale (Strongly Disagree - Disagree - Agree - Strongly Agree). Grade each one of the affirmations in the questionnaire with one of these grades and offer a brief justification.*". The affirmations of the questionnaire are detailed below:

- ❏ The picture captured well what was described in the level;
- ❏ There are no elements in the picture that were not mentioned in the level description;
- ❏ The level description is clear enough so as to describe what should be in the picture;
- ❏ The picture looks like a Backrooms level and suits the Backrooms universe;

CAPÍTULO 5

Results

This chapter describes the tests carried out to assess the results obtained and compare the performance of the three proposed strategies. The first one consisted of prompting the test inputs into Deepseek with the goal of generating new Backrooms level descriptions; The second one consisted of fine-tuning Llama 1B for the Backrooms level description generation task and the third one consisted of pre-training a vanilla decoder-only architecture to then finetune this pre-trained backbone for the Backrooms level description generation task.

As described in Chapter 4, three evaluation strategies were used: In the first one, ChatGPT was used as a judge; the second consisted of Human Evaluation; and the third one represents a new assessment strategy in which the quality of the textual outputs of the competing models is evaluated through the compatibility between these outputs and pictures that are generated from them, as presented in subsection 5.3.

5.1 Evaluation using LLMs as judges

As detailed in chapter 4, chatGPT 4o mini was used as judge. A 4-point Likert scale questionnaire was passed to it and, with each grade, a small justification was required. For each input of the Test Dataset, each output of the three strategies was passed accompanied by the questionnaire, the input itself (the level idea) and instructions about how the evaluation response should be formatted (a file .json containing the grade of each statement and the respective justification for the grade). The results of this evaluation were detailed in Table 7.

Criteria	DeepSeek				B-Llama				Tiny Backrooms			
	++	+	-	–	++	+	-	–	++	+	-	–
1. The syntax of the text is correct. It doesn't contain any grammatical errors	161	0	0	0	42	106	13	0	0	79	81	1
2. The text is clear, which means the ideas presented in it are well described and easy to understand	161	0	0	0	0	158	3	0	0	126	34	1
3. The level description is consistent with what was asked in the idea	161	0	0	0	129	16	16	0	75	82	4	0
4. The description follows the standard Backrooms level structure	4	157	0	0	0	153	8	0	0	104	57	0
5. The level description suits the Backrooms universe and seems like an actual Backrooms level	161	0	0	0	131	30	0	0	68	92	1	0
6. The level description is original and creative. It contains new and interesting ideas	117	44	0	0	2	158	1	0	0	151	10	0

Table 7 – GPT's evaluation of Deep Seek vs. Backrooms-Llama vs. Tiny Backrooms. Columns: ++: Strongly Agree, +: Agree, -: Disagree, –: Strongly Disagree.

The first interesting observation that can be made based on the results of the judgment is that, for all affirmations of all 161 test samples, there was not a single 'Strongly Disagree' for either DeepSeek's levels or Backrooms-Llama's levels. Tiny Backrooms only got 2. DeepSeek's levels also didn't receive any 'Disagree' grades, contrary to Backrooms-Llama, which did receive a few 'Disagree' grades and Tiny Backrooms, which got a considerable amount of 'Disagree' grades. A more in depth analyses of the grades for each affirmation was detailed below.

It can be observed on the results for affirmation (1) that Deepseek did not commit any grammatical errors (see Table 7). Backrooms-Llama, on the other hand, had some trouble with inconsistent level numbering or naming (it made reference to the level itself with different numbers or names throughout the description) and a few spelling errors such as 'wandereres' instead of 'wanderers'. There were also a few problems with word capitalization, such as 'xyLOPHONE' instead of 'xyophone'. The main problem with Tiny Backrooms was inconsistent capitalization. It didn't really learn how to use capitalization correctly so it mainly left everything in lowercase. It also had some awkward phrasing and some repeated phrases and therms.

For affirmation (2), the results show that DeepSeek generated very clear levels for all test samples, while Backrooms-Llama generated clear levels for most of the test samples. As justification for the levels that scored 'Disagree', GPT justified that they were overly verbose and for the reason why the other levels scored 'Agree', instead of 'Strongly Agree', it justified that some phrases could be simplified for better understanding. The bad grades Tiny Backrooms got were mainly justified because of the repetition and redundancy present in the descriptions.

For affirmation (3), once again, DeepSeek scored only 'Strongly Agree'. Backrooms-Llama, overall, was able to generate levels which were consistent with what was asked in the idea, with the exception of three cases, in which the generated level did not really match the original idea. All other given negative grades were due to the confusion Backrooms-Llama sometimes make by referring to the same level by different names (Level -33 instead of Level -32, for example). For this affirmation Tiny Backrooms also got mostly positive grades and the negative ones were very punctual cases in which the generated level did nott fully align with the original concept.

For affirmation (4), the main reason for the 'Disagree' grades for Backrooms-Llama were due to the

confusion the models sometime makes with the level name. As for the other grades, the main reason why they were not higher were the sections entities, entries/exits and colonies. It was observed that, quite often, Backrooms-Llama does not generate these sections, which might not be a problem, since the main theory for this is that, in the wiki (and, consequentially, in the fine-tuning dataset), there are a lot of cases in which these sections are not detailed, so it is believed that the model has learned this pattern. Similarly to Backrooms-Llama, Tiny Llama often did not generate the entities, entrances, and exits sessions, which reinforces the theory that this was a behavior learned because of the dataset. Repetitions and redundancies also affected this grade.

For affirmations (5) and (6), the three models performed really well, which shows that they were able to generate original levels which suit the Backrooms universe. This is very interesting, since creativity and originality are such subjective metrics.

By analyzing all these grades, it is possible to observe that, even though DeepSeek clearly performed better, Backrooms-Llama was able to get pretty good grades in every affirmation, scoring mostly 'Agree', a lot of 'Strongly Agree', few 'Disagree' and not a single 'Strongly Disagree'. That is quite impressive, considering that DeepSeek has 236 times more parameter than Backrooms-llama. Additionally, it was also possible to observe that the main problem with Backrooms-Llama is the mixing-up it does with level names. But, when it comes to the coherence, consistency, creativity, and fidelity to the Backrooms universe and its levels structures, Backrooms-Llama is actually performing really well. Though, the fact that it does not often generate the sections entities, entries/exits and colonies would also be worth investigating, since it could be something it learned from the dataset or an actual learning problem.

Even though Tiny Backrooms got the worst grade out of the three models, it did get mostly positive grades. The main problems it presented were related to capitalization, which really was something the model did not quite learn. Redundancy was also frequently present in the text, which is a quite common problem in language models. However, considering that this is a 190M-parameter model trained from scratch on a custom dataset, it performs well and is capable of generating original structured Backrooms levels, especially, if considering that Backrooms-llama is 1B and Deepsek 236B.

5.2 Human Evaluation

To enrich even more the evaluation process, a group of six human judges, who are familiar with The Backrooms universe, performed evaluation on a subset of the test set. For that, 30 random level ideas were selected and the generations from Backrooms-Llama and DeepSeek for these levels were evaluated (totalizing 90 evaluated levels - 30 from Backrooms-Llama, 30 from DeepSeek and 30 from Tiny Backrooms).

Each human judge evaluated 10 levels. To avoid any kind of bias, they were not told which model they were evaluating. For each level, they had to answer the same 4-point Likert scale questionnaire used in section **??** and offer a brief written justification for the given grades. The results of this evaluation are detailed in Table 8

Table 8 – Human evaluation of Deep Seek, Backrooms-Llama, and Tiny Backrooms. Columns: ++: Strongly Agree, +: Agree, -: Disagree, –: Strongly Disagree.

	DeepSeek				B-Llama				Tiny Backrooms			
Criteria	++	+	-	–	++	+	-	–	++	+	-	–
1. The syntax of the text is correct. It doesn't contain any grammatical errors	28	2	0	0	7	14	9	0	11	3	16	0
2. The text is clear, which means the ideas presented in it are well described and easy to understand	22	8	0	0	11	12	7	0	1	12	13	4
3. The level description is consistent with what was asked in the idea	22	8	0	0	14	10	4	2	6	13	5	6
4. The description follows the standard Backrooms level structure	25	5	0	0	9	16	5	0	12	11	4	3
5. The level description suits the Backrooms universe and seems like an actual Backrooms level	17	10	3	0	22	8	0	0	11	13	5	1
6. The level description is original and creative. It contains new and interesting ideas	8	16	6	0	19	11	0	0	9	16	4	1

There was a lot more variability of judgment in the human evaluation than in GPT's evaluation. Some evaluators were much harsher than others, but that is expected in a human evaluation, and that is exactly what makes it so interesting. DeepSeek performed considerably better based on these results, considering that it got 9 negative grades, while Backrooms-Llama got 27 and Tiny Backrooms 62. By comparing these results with the LLM judge ones, it appears that the human judges were stricter than the GPT.

The criticisms received by Backrooms-Llama from the human judges were very similar to the ones given by the LLM judge. The main problem raised was the mixing-up the model does with level names. Also, the lack of the section entries/exits and colonies in some levels was pointed out. There were very few comments calling out some contradictions in the descriptions. The main criticisms DeepSeek received were regarding the lack of originality in the generated levels and that too many entities were created in a single level description, which made the text a little long and repetitive to read.

As for Tiny Backrooms, the main critiques were related to the repetition problem, which was also raised by GPT, and contradictions in the text. It was also said sometimes that the text could be a bit confusing, even if the ideas inside were interesting and creative.

Even though DeepSeek performed reasonably better in the first 4 affirmations, when it comes to the last two, which are about originality and how much the idea suits The Backrooms universe, Backrooms-Llama outperformed DeepSeek. That shows that DeepSeek is stronger when it comes to the most objective metrics like syntax, clarity, and consistency, but Backrooms-Llama actually seems to generate more creative, original, and universe-fitting levels. It is also interesting to point out that these results diverge from GPT's evaluation, which evaluated Deepseek's creativity and fidelity to the Backrooms universe higher than Backrooms-Llama. That shows that humans familiar with the Backrooms universe seem to think that Backrooms-Llama is actually stronger when it comes to the two last criteria.

5.3 Evaluation trough generated pictures

This section proposes a new evaluation method based on images for evaluating LLMs in descriptive tasks, such as generating Backroom level descriptions. In the context of this research, this was done by

randomly selecting 30 samples from the the test dataset and, with the output of each of the three strategies for these samples, generating images of the Backrooms level descriptions with the deep-learning text-image model Dall-E.

To evaluate the quality of the produced images and their correspondence to the generated text, each image was then passed to chatGPT 4o, which accepts both text and pictures in the same prompt, accompanied by the corresponding level description. With them, another 4-point Likert scale was also passed, and the following prompt: "*Evaluate this Backrooms level description and the picture that was generated based on it, by answering the following questionnaire using a 4-point Likert scale (Strongly Disagree - Disagree - Agree - Strongly Agree). Grade each one of the affirmations in the questionnaire with one of these grades and offer a brief justification.*"

Both the affirmations that were evaluated in the questionnaire and the results of this evaluation were detailed in Table 9.

	DeepSeek				**B-Llama**				**Tiny Backrooms**			
Criteria	++	+	-	–	++	+	-	–	++	+	-	–
1. The picture captured well what was described in the level	29	1	0	0	22	8	0	0	10	20	0	0
2. There are no elements in the picture that were not mentioned in the level description	21	9	0	0	22	8	0	0	0	8	22	0
3. The level description is clear enough so as to describe what should be in the picture	30	0	0	0	24	5	1	0	0	20	10	0
4. The picture looks like a Backrooms level and suits the Backrooms universe	26	4	0	0	26	4	0	0	19	9	2	0

Table 9 – Evaluation of image-to-text alignment for Deep Seek, Backrooms-Llama, and Tiny Backrooms. Columns: ++: Strongly Agree, +: Agree, -: Disagree, –: Strongly Disagree.

Just like in the evaluation described in the previous subsection, for all affirmations of the questionnaire, there was not a single 'Strongly Disagree' for either DeepSeek's levels of Backrooms-Llama's levels. DeepSeek's levels also did not receive any 'Disagree' grades, while Backrooms-Llama received a single 'Disagree' grade for affirmation (3). This was also due to the confusion the model makes with the level names. These results show that both models performed very similarly, with Deepseek performing slightly better, which indicates that both the descriptions generated by Backrooms-llama and Deepseek allowed Dall-E to generate pictures that represented the levels fairly.

Regarding Tiny Backrooms, the main problem actually originated from Dall-e. In many cases, the image model added elements which weren't present in the description, mainly humanoid figures for some reason. Dall-e's weights might have been updated since Deepseek and Backrooms-Llama results (Tiny Backrooms was evaluated many months after) and that could be the reason for this, which was also the cause for the negative grades for criteria 2. The redundancy of Tiny Backrooms was the main reason presented for the negative grades for affirmation 3. In the justifications, Dall-e said that the repetition of words and inconsistencies in the text made it harder for it to generate a clear consistent picture of the level. An example of generation that got bad grades was illustrated in Figure 16. But, in general, Tiny Backrooms was able to generate descriptions clearly enough for Dall-e to create a faithful consistent picture.

Level description

Backrooms level idea: A power grid where electric wires slither like snakes, whispering in code.: **level name:** level 879: the "electric grid system"
safety: unstable; entities present.
detailed level description: level 879, known as the "electric grid,"is an expansive network of gridlocked utility lines that stretch endlessly in all directions. the grid is continuously shifting, making it impossible for wanderers to encounter one another at any given time. the power grid does not create any useful items, making it a dangerous location for wanderers. as wanderers navigate through this grid, they may encounter various electrical boxes embedded in the walls, although no electricity has been produced. although the electricity level does not produce electricity, it does create a mysterious and unsettling phenomenon that induces paranoia among wanderers. the level is described as feeling like a vacuum cleaner than a typical grid environment, akin to the feeling of being watched.
the grid, known as "the grid,"is a complex network of interconnected utility poles that loop endlessly. these poles are connected to numerous pipes, some of which are connected to the outside world with cables. the interior of the grid is characterized by a constant sound of running electricity, with the sound of distant sirens echoing throughout the area.
despite the power grid's seemingly infinite size and its ability to produce electricity, the power grid does not create any useful items. the layout of the level is labyrinthine and convoluted, often leading to dead ends or dead ends. the wiring of the poles is erratic, with the most notable power grid being the one found at the bottom of a pile of abandoned electrical boxes. the area surrounding the grid is a void, devoid of natural illumination, with a constant and pervasive darkness that obscures the area. the origin of this darkness remains ambiguous, leading to the conclusion that it is an illusion created by the grid itself.
entities: entities within the level are not well-defined.
entrances and exits: entrances to level 879 are primarily through entrances found in other backrooms levels. there are unconfirmed reports of individuals attempting to leave the level through various means.
exits: exiting level 879 is more challenging; the only known exit is by completing certain tasks that align with the grid's stability. these tasks include: (organization, maintenance, maintenance, and repair)
colonies and outposts: there are no known bases, communities, or outposts within level 879. the level's instability and lack of resources discourage the establishment of any permanent settlements.

Evaluation


```
{
  "1": {
    "grade": "Agree",
    "justification": "The image strongly
        reflects the described environment,
        showing an expansive network of
        utility poles, cables, and electrical
         boxes stretching into the distance.
        The dark, labyrinthine layout and
        overwhelming presence of wiring align
         well with the concept of an endless
        electric grid."
  },
  "2": {
    "grade": "Disagree",
    "justification": "The image includes wet,
         reflective ground and puddles, as
        well as a more industrial, grounded
        setting, which were not explicitly
        mentioned in the description. These
        elements add realism but are not
        directly described."
  },
  "3": {
    "grade": "Agree",
    "justification": "The description
        provides enough detail about utility
        poles, endless wiring, darkness, and
        electrical boxes to guide the visual
        outcome. While some parts are
        repetitive or slightly contradictory,
         the core imagery is clear enough to
        produce a coherent scene."
  },
  "4": {
    "grade": "Strongly Agree",
    "justification": "The image strongly fits
         the Backrooms aesthetic, with its
        eerie, industrial, and oppressive
        atmosphere. The sense of isolation,
        darkness, and endless repetition of
        structures aligns perfectly with the
        Backrooms universe."
  }
}
```


Figure 16 – Example of level, level image and image evaluation from Tiny Backrooms.

An example of pictures generated by the output of each model is presented in Figure 17.

Figure 17 – "Backrooms level idea: Level 882, also known as "The Sunken Cathedral", consists of a gothic cathedral half-flooded, with ghostly hymns in the distance..- Example of picture generated from Backrooms-Llama, Deepseek and Tiny Backrooms descriptions, respectively.

It can be observed that the three images images are of similar styles, presenting a very dark flooded cathedral, which is the central theme of the ideas/descriptions. In the ones created based on the Deepseek generation and Backrooms-Llama, there are a lot of humanoid figures, which is something that is not necessarily aligned with The Backrooms, since, even if there are entities, humans and colonies there, it is mostly described as a place of vast empty rooms. This could be a Dall-e problem, since it has a tendency to add humans to the pictures it generates. But, in these case, it actually highlights one of the problems that Deepseek and Backrooms-Llama generations sometimes contain, which is the excess of entities/colonies in certain levels. In this description, illustrated by Figure 18, for example, Backrooms-Llama generated way too many entities, while in Figure 19, Deepseek generated more entities/colonies than usual.

...
Entities:

The primary entity present in Level-882 is the Church of Christ. While it is not hostile, it can become aggressive if provoked. It is rumored that the cathedral itself may be inhabited by entities, though this remains unconfirmed.
The level is also home to a variety of other entities that are not present during daylight hours. Some of these entities include:

- **The Choir:** A group of ghost-like figures that sing hymnal melodies during the night. They are known to wander the halls of Level -882 and can pose a threat if encountered.
- The Choir of Angels: A smaller group that sings hymnas during daytime, often accompanied by The Church Choir. Their presence is less frequent than that of The Chorus.

- Other entities such as Smilers, Skin-Stealers, Hounds, Wretches, M.E.G. agents, The Collector, Deathmoths, Smiler Hordes, Flesh Eaters, Shriekers, Sanguine, Vampires, Demons, Fallen, Noclips, etc. may also be present, but their numbers are unknown.

Figure 18 – Example of Backrooms-Llama generation containing too many entities.

Entities:
The Drowned Choir: Silhouettes seen in the upper galleries, their mouths moving in sync with the hymns. Turning a light their way reveals empty space.

The Bishop: A robed figure standing waist-deep in water, facing the altar. It does not react to presence but slowly raises its hands in benediction when unobserved.

The Penitent: A humanoid shape kneeling in the water, submerged up to its shoulders. It does not move but emits quiet sobbing that stops when approached.

Colonies and Outposts:
The Congregation: A group that gathers near the altar, holding silent vigils. They claim the hymns guide them but refuse to explain to where.

M.E.G. Outpost "Sanctuary": Researchers documenting the cathedral's decay. Their equipment picks up sonar readings of large shapes moving beneath the water.

The Bell Ringers: Wanderers who toll the cathedral's rusted bells at random intervals. They speak of "calling something back" but won't elaborate.

Figure 19 – Example of Deepseek generation containing too many entities/colonies.

5.4 Final Discussion of Results

This section presents a general discussion of the results obtained throughout this chapter, with the goal of comparing the three proposed strategies, highlighting their strengths and weaknesses, establishing trade-offs and connecting the findings with existing literature.

5.4.1 Comparison of the Proposed Strategies

Based on the three evaluation strategies, it is clear that DeepSeek consistently achieved the highest scores in the most objective criteria, such as syntax correctness, clarity, and consistency. As shown in the evaluations, it produced grammatically flawless outputs and maintained strong alignment with the input ideas across all test samples. This behavior is expected given its large scale and extensive pre-training on diverse corpora.

Backrooms-Llama, on the other hand, demonstrated a strong balance between performance and specialization. Although it presented occasional issues, such as inconsistencies in level naming and omission of certain structural sections, it performed competitively across most criteria. Notably, in the human evaluation, it outperformed DeepSeek in terms of creativity and alignment with the Backrooms universe, suggesting that domain-specific fine-tuning can lead to more stylistically appropriate and engaging outputs.

Tiny Backrooms, despite being the weakest among the three, still demonstrated the ability to generate coherent and structured outputs after fine-tuning. Its main limitations were related to repetition, capitalization, and occasional contradictions. It is important to emphasize that the pre-trained model alone was not able to generate properly structured Backrooms levels, indicating that domain-specific pre-training by itself was insufficient for the target task. The subsequent fine-tuning stage proved essential for teaching the model the expected document structure, formatting conventions, and style of Backrooms level descriptions. Nevertheless, considering its relatively small size and that it was trained from scratch on a custom dataset, its final performance highlights the feasibility of building lightweight domain-specific language models through a complete pre-training and fine-tuning pipeline.

One thing that is important to highlight is that the fine-tuning dataset is relatively small, which made the adaptation task considerably more challenging for both Backrooms-Llama and Tiny Backrooms. Nevertheless, the experiments clearly show that even this limited amount of domain-specific supervision was crucial for Tiny Backrooms to learn how to generate properly structured Backrooms levels. The current results suggest that larger fine-tuning datasets could further improve the performance of both models.

5.4.2 Trade-offs Between Strategies

The comparison between the three approaches reveals important trade-offs:

- **Scale vs. Specialization:** DeepSeek benefits from scale, resulting in highly polished and consistent outputs. However, Backrooms-Llama demonstrates that smaller, domain-adapted models can achieve comparable performance while offering greater stylistic alignment and creativity.
- **Generalization vs. Fidelity:** While DeepSeek generalizes well across tasks, it sometimes produces less original or overly verbose descriptions. In contrast, Backrooms-Llama shows stronger fidelity to the specific conventions of the Backrooms universe.
- **Efficiency vs. Quality:** Tiny Backrooms represents a more efficient alternative in terms of computational cost, but at the expense of output quality and consistency.

These trade-offs are aligned with findings in the literature, which suggest that large-scale pre-trained models excel in general-purpose tasks, while smaller fine-tuned models can outperform them in niche domains when sufficient domain-specific data is available.

5.4.3 Evaluation Methods: Strengths and Limitations

The three evaluation strategies employed in this, which are LLM judgment, human evaluation, and image-based evaluation provided complementary perspectives.

The use of LLMs as judges enabled large-scale and consistent evaluation, but tended to produce overly optimistic assessments, as observed by the absence of strongly negative scores. This suggests a potential bias toward agreement, which has also been noted in recent literature on automated evaluation methods.

Human evaluation, while more subjective and variable, provided valuable insights into aspects such as creativity and fidelity to the Backrooms universe. The stricter judgments indicate that human evaluators are more sensitive to subtle issues, particularly those related to domain fidelity.

The image-based evaluation represents a novel contribution of this work. By measuring the alignment between generated text and images, this method captures an additional dimension of quality: how well descriptions translate into visual representations. However, this approach introduces dependencies on the image generation model, which can introduce noise, as observed in the case of Tiny Backrooms, where external factors such as DALL-E behavior influenced the results.

5.4.4 Final Remarks

The results obtained in this work are consistent with trends observed in the literature on language models. In particular, they reinforce the idea that domain-specific fine-tuning on smaller models can significantly enhance performance in specialized tasks, or at least compete with the performance of larger models, specially when considering the tradeoffs.

Finally, further investigation into multimodal evaluation strategies could provide deeper insights into the relationship between textual and visual representations, potentially leading to more comprehensive evaluation frameworks.

In summary, this work demonstrates that while large-scale models such as DeepSeek achieve superior performance in general metrics, smaller domain-specific models such as Backrooms-Llama and Tiny Backrooms offer a compelling alternative by balancing efficiency, creativity, and fidelity to the target domain, with a much smaller cost, while offering an alternative for dealing with private data.

CAPÍTULO 6

Conclusion

This work proposed an approach to generate a smaller language model, when compared with famous chat LLMs, like GPT and Deepseek, that can be trained in a smaller and simpler hardware, for generating text related to a specific domain, which, in this case, was Backrooms level descriptions.

With this in mind, a Backrooms dataset was created and openly published and three different strategies for generating new Backrooms level descriptions with LMs were explored: The first one simply consisted of some prompt engineering for generating new level descriptions with DeepSeek-V3; In the second one, Llama 1B was fine-tuned for generating Backrooms level descriptions and named Backrooms-Llama. As for the third one, a vanilla Decoder-only was trained from scratch with supernatural, backrooms-related and general data, from a custom made dataset and then finetuned for the Backrooms level description generation task and named Tiny Backrooms.

The performance of these three strategies were evaluated with two different evaluation methods that used GPT-4o as judge: one in which only the output of the three models were evaluated and a second one in which pictures were created, based on each model's output, by text-image model Dall-E and then evaluated alongside their respective descriptions. Human evaluation was also used to analyze and compare the strategies.

A big problem that surfaced during the experiments, specially for Tiny Backrooms, was the size of the fine-tuning dataset. Unfortunately, after excluding obsolete or empty pages from the scraped data, there was a lot less data than expected at first, which resulted in a quite small dataset for such a complex textual task.

The results showed that even though Backrooms-Llama performed worse than Deepseek by making mistakes like confusing level names, committing a few spelling errors, and generating levels that contained some contradiction, in general, it got manly good evaluations and was able to generate coherent Backrooms level descriptions. Furthermore, it is interesting to point out that the levels created by Backrooms-Llama were deemed more original and fitting to the Backrooms universe than DeepSeek's generations in the human evaluation.

As for Tiny Backrooms, even though it performed the worst out of the three strategies and presented many problems like redundancies, inconsistencies, and contradictions on the text, it did learn how to generate structured Backroom levels with original interesting ideas, with a much smaller architecture than both other models and an entirely custom dataset. The results provide some indication that a larger fine-tuning dataset could lead to improved performance and present more competition to the other

two strategies. These are very promising results, which indicate that with a much more modest model than Deepseek (or other famous LLMs), it is possible to perform a relatively difficult and specific text generation task.

6.1 Main Contributions

The main contributions of this work are:

- ❏ Analyzing whether, for specific-domain generative creative tasks, it is possible to produce Deep Learning (DL) models with smaller architectures and lower financial cost in terms of hardware and training requirements that present satisfactory performance in relation to consecrated and wide LLMs;
- ❏ Investigating the use of Large Language Models for creative text generation in highly specific and detail-rich domains, using the Backrooms universe as a case study due to its complex world building, structured level conventions and surreal atmosphere;
- ❏ Proposing a fine-tuning pipeline for adapting SLMs to domain-specific creative text generation tasks, using the Backrooms universe as a case study through the development of Backrooms-Llama, which is a fine-tuned Llama 1B;
- ❏ Proposing a complete training pipeline for compact domain-specific language models, including pretraining from scratch and task-specific fine-tuning with custom datasets, validated through the Tiny Backrooms model for Backrooms level generation;;
- ❏ Exploring an alternative evaluation strategy based on image generation from textual outputs. In this setting, the descriptions produced by the models are used as inputs to a deep learning model that generates images, and performance is assessed through the analysis of the resulting images, offering an additional perspective on text quality.
- ❏ Creating and publishing a public Backrooms Datasets and releasing all training code developed in this work, as well as the produced model [1];

6.1.1 Challenges and Limitations

Several challenges were encountered throughout this work. One of the main difficulties relates to the nature and size of the created dataset. Scraped data comes with a lot of problems and even after cleaning it, some mistakes can end up in the final dataset. Also, the cleaning process ended up reducing the dataset considerably, which also presented itself as a problem.

Another challenge lies in evaluating creative text generation tasks. Metrics such as creativity and fidelity to a fictional universe are inherently subjective, making it difficult to establish fully objective evaluation criteria.

[1] Backrooms-llama repository, Tiny Backrooms repository

Additionally, the dependence on external models (e.g., ChatGPT and DALL-E) introduces potential sources of variability and reproducibility concerns, since these models don't give the same answer for a same prompt always and are constantly being updated.

6.2 Future Work

Future work may explore the application of the proposed methodologies to other creative and domain-specific text generation scenarios beyond the Backrooms universe, allowing a broader investigation of how SLMs behave in different structured and detail-rich contexts. Examples could include fantasy world-building, horror narratives, science fiction environments, role-playing game lore, or other fictional universes with well-defined conventions and stylistic constraints.

Another important direction involves expanding and refining the datasets used during both pre-training and fine-tuning. Larger, cleaner, and more diverse datasets could improve the coherence, creativity, and structural consistency of the generated texts, potentially reducing the performance gap observed between the proposed SLMs and larger proprietary LLMs such as DeepSeek. Future work may also investigate enriching the fine-tuning corpus with complementary sources, such as books, movies, games, or other media related to the target fictional universe. Additionally, synthetic datasets generated by powerful LLMs could be explored as a means of increasing data diversity and volume while preserving domain-specific characteristics.

Future studies could also evaluate additional open-source and proprietary LLMs, enabling a more comprehensive comparison between different architectures, parameter scales, and adaptation strategies for domain-specific creative generation. Similarly, the evaluation process itself could be expanded considerably. Human evaluation could involve a larger and more diverse group of participants, improving the statistical reliability and reducing potential subjectivity in the analyses.

Furthermore, future experiments may include multiple automatic evaluators instead of relying on a single LLM-based judge. Comparing evaluations produced by different models could help investigate consistency, biases, and limitations in LLM-as-a-judge methodologies for creative text generation tasks. Another interesting direction would be to investigate persona-based evaluation, where LLM judges are instructed to assume different profiles, such as experienced Backrooms enthusiasts, casual readers, or creative writing experts, allowing the analysis of how evaluator perspective influences the assessment of generated content.

Similarly, the image-based evaluation methodology could be extended by employing multiple text-to-image generation models instead of relying on a single system. Comparing image generators with different architectures and capabilities would help determine how robust the proposed evaluation methodology is with respect to the characteristics and biases of the underlying image generation model.

Finally, another promising direction would be exploring alternative training strategies, such as parameter-efficient fine-tuning techniques (e.g., LoRA), Knowledge Distillation, retrieval-augmented generation approaches, reinforcement learning techniques, or hybrid symbolic-generative methods. These strategies may help compact models achieve better performance while maintaining lower computational costs and improved accessibility for local deployment.

6.3 Published Papers, Datasets and Models

- JULIA, ROXANNE; JULIA, RITA ; ZANCHETTA DO NASCIMENTO, MARCELO ; RIBEIRO DE FARIA, ELAINE. Backrooms-Llama: A Specific Domain Small Language Model for Computational Storytelling. In: 18th International Conference on Agents and Artificial Intelligence, 2026, Marbella. Proceedings of the 18th International Conference on Agents and Artificial Intelligence, 2026. v. 3. p. 2376-2384. doi=10.5220/0014239800004052
- JULIA, ROXANNE. Backrooms Levels Dataset, Hugging Face, 2026. doi:10.57967/hf/8715
- JULIA, ROXANNE. Backrooms-Llama, Hugging Face, 2026. doi:10.57967/hf/8730

References

... Generating role-playing game quests with gpt language models. 2024.

BADSHAH, S.; SAJJAD, H. **Reference-Guided Verdict: LLMs-as-Judges in Automatic Evaluation of Free-Form Text**. 2024. Disponível em: <https://arxiv.org/abs/2408.09235>.

BANERJEE, S.; LAVIE, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: GOLDSTEIN, J. et al. (Ed.). **Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization**. Ann Arbor, Michigan: Association for Computational Linguistics, 2005. p. 65–72. Disponível em: <https://aclanthology.org/W05-0909>.

BAVARIAN, M. et al. Efficient training of language models to fill in the middle. **ArXiv**, abs/2207.14255, 2022. Disponível em: <https://api.semanticscholar.org/CorpusID:251135268>.

BENDER, E. M. et al. On the dangers of stochastic parrots: Can language models be too big? In: **Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)**. [S.l.]: ACM, 2021. p. 610–623.

BLACK, S. et al. Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow. In: . [s.n.], 2021. Disponível em: <https://api.semanticscholar.org/CorpusID:245758737>.

BOMMASANI, R. et al. On the opportunities and risks of foundation models. **ArXiv**, 2021. Disponível em: <https://crfm.stanford.edu/assets/report.pdf>.

BRECK, E. et al. The ml test score: A rubric for ml production readiness and technical debt reduction. In: **Proceedings of the 2017 IEEE International Conference on Big Data**. [S.l.: s.n.], 2017.

BROWNE, C. B. et al. A survey of monte carlo tree search methods. In: **IEEE Trans. Comput. Intell. AI Games**. [S.l.: s.n.], 2012. v. 4, n. 1, p. 1–43.

CAFFAGNI, D. et al. Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops**. [S.l.: s.n.], 2024. p. 1818–1826.

CARLINI, N. et al. Extracting training data from large language models. In: **USENIX Security Symposium**. [S.l.: s.n.], 2021.

CHANG, Y. et al. A survey on evaluation of large language models. **ACM Trans. Intell. Syst. Technol.**, Association for Computing Machinery, New York, NY, USA, v. 15, n. 3, mar. 2024. ISSN 2157-6904. Disponível em: <https://doi.org/10.1145/3641289>.

CHEN, Y. et al. Evaluating diversity in automatic poetry generation. In: AL-ONAIZAN, Y.; BANSAL, M.; CHEN, Y.-N. (Ed.). **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**. Miami, Florida, USA: Association for Computational Linguistics, 2024. p. 19671–19692. Disponível em: <https://aclanthology.org/2024.emnlp-main.1097/>.

CHOWDHERY, A. et al. Palm: scaling language modeling with pathways. **J. Mach. Learn. Res.**, JMLR.org, v. 24, n. 1, jan. 2023. ISSN 1532-4435.

DEEPSEEK-AI et al. **DeepSeek-V3 Technical Report**. 2025. Disponível em: <https://arxiv.org/abs/2412.19437>.

DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In: **Proc. NAACL-HLT**. [S.l.: s.n.], 2019. p. 4171–4186.

GAO, L. et al. The pile: An 800gb dataset of diverse text for language modeling. 2020.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [S.l.]: MIT Press, 2016.

GURURANGAN, S. et al. Don't stop pretraining: Adapt language models to domains and tasks. In: JURAFSKY, D. et al. (Ed.). **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**. Online: Association for Computational Linguistics, 2020. p. 8342–8360. Disponível em: <https://aclanthology.org/2020.acl-main.740/>.

HENDRYCKS, D. et al. Measuring massive multitask language understanding. In: **International Conference on Learning Representations (ICLR)**. [S.l.: s.n.], 2021.

HINTON, G. E.; VINYALS, O.; DEAN, J. Distilling the knowledge in a neural network. **ArXiv**, abs/1503.02531, 2015. Disponível em: <https://api.semanticscholar.org/CorpusID:7200347>.

HOU, Y. et al. Wikicontradict: A benchmark for evaluating llms on real-world knowledge conflicts from wikipedia. In: GLOBERSON, A. et al. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2024. v. 37, p. 109701–109747. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2024/file/c63819755591ea972f8570beffca6b1b-Paper-Datasets_and_Benchmarks_Track.pdf>.

HU, C.; ZHAO, Y.; LIU, J. **Game Generation via Large Language Models**. 2024. Disponível em: <https://arxiv.org/abs/2404.08706>.

HUBER et al. everaging the potential of large language models in education through playful and game-based learning. In: **Educ Psychol Rev 36, 25**. [S.l.: s.n.], 2024.

KAPLAN, J. et al. Scaling laws for neural language models. **ArXiv**, abs/2001.08361, 2020. Disponível em: <https://api.semanticscholar.org/CorpusID:210861095>.

KASHYAP, R.; KASHYAP, V.; C.P, N. Gpt-neo for commonsense reasoning-a theoretical and practical lens. **ArXiv**, abs/2211.15593, 2022. Disponível em: <https://api.semanticscholar.org/CorpusID:254043914>.

KASNECI, E. et al. Chatgpt for good? on opportunities and challenges of large language models for education. **Learning and Individual Differences**, v. 103, p. 102274, 2023. ISSN 1041-6080. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1041608023000195>.

LANZI, P. L.; LOIACONO, D. Chatgpt and other large language models as evolutionary engines for online interactive collaborative game design. In: **Proceedings of the Genetic and Evolutionary Computation Conference**. ACM, 2023. (GECCO '23). Disponível em: <http://dx.doi.org/10.1145/3583131.3590351>.

LEHMAN, J. et al. Evolution through large models. In: ____. **Handbook of Evolutionary Machine Learning**. Singapore: Springer Nature Singapore, 2024. p. 331–366. ISBN 978-981-99-3814-8. Disponível em: <https://doi.org/10.1007/978-981-99-3814-8_11>.

LIN, C.-Y. ROUGE: A package for automatic evaluation of summaries. In: **Text Summarization Branches Out**. Barcelona, Spain: Association for Computational Linguistics, 2004. p. 74–81. Disponível em: <https://aclanthology.org/W04-1013>.

LIN, T. et al. A survey of transformers. **AI Open**, v. 3, p. 111–132, 2022. ISSN 2666-6510. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2666651022000146>.

LIU, B. Comparative analysis of encoder-only, decoder-only, and encoder-decoder language models. 2024.

MARCO, G. et al. **Pron vs Prompt: Can Large Language Models already Challenge a World-Class Fiction Author at Creative Text Writing?** 2024. Disponível em: <https://arxiv.org/abs/2407.01119>.

MARCO, G.; RELLO, L.; GONZALO, J. Small language models can outperform humans in short creative writing: A study comparing SLMs with humans and LLMs. In: RAMBOW, O. et al. (Ed.). **Proceedings of the 31st International Conference on Computational Linguistics**. Abu Dhabi, UAE: Association for Computational Linguistics, 2025. p. 6552–6570. Disponível em: <https://aclanthology.org/2025.coling-main.437/>.

MART'iNEZ-CRUZ, R.; L'oPEZ-L'oPEZ, A. J.; PORTELA, J. Chatgpt vs state-of-the-art models: A benchmarking study in keyphrase generation task. **Applied Intelligence**, v. 55, n. 50, 2025. Disponível em: <https://doi.org/10.1007/s10489-024-05901-4>.

MCCLOSKEY, M.; COHEN, N. J. Catastrophic interference in connectionist networks: The sequential learning problem. In: BOWER, G. H. (Ed.). Academic Press, 1989, (Psychology of Learning and Motivation, v. 24). p. 109–165. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0079742108605368>.

MCDONALD, T.; EMAMI, A. Trace-of-thought prompting: Investigating prompt-based knowledge distillation through question decomposition. In: FU, X.; FLEISIG, E. (Ed.). **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)**. Bangkok, Thailand: Association for Computational Linguistics, 2024. p. 293–306. ISBN 979-8-89176-097-4. Disponível em: <https://aclanthology.org/2024.acl-srw.35/>.

PAPINENI, K. et al. Bleu: a method for automatic evaluation of machine translation. In: **Proceedings of the 40th Annual Meeting on Association for Computational Linguistics**. USA: Association for Computational Linguistics, 2002. (ACL '02), p. 311–318. Disponível em: <https://doi.org/10.3115/1073083.1073135>.

PERLIN, K. An image synthesizer. In: **Proc. SIGGRAPH**. [S.l.: s.n.], 1985. p. 287–296.

RADFORD, A. et al. Language models are unsupervised multitask learners. 2019.

____. Language models are unsupervised multitask learners. In: **OpenAI Tech. Rep.** [S.l.: s.n.], 2019.

RAFFEL, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer. In: **J. Mach. Learn. Res.** [S.l.: s.n.], 2020. v. 21, p. 1–67.

RIVIERE, G. T. M. et al. Gemma 2: Improving open language models at a practical size. **ArXiv**, abs/2408.00118, 2024. Disponível em: <https://api.semanticscholar.org/CorpusID:270843326>.

ROZI'eRE, B. et al. Code llama: Open foundation models for code. **ArXiv**, abs/2308.12950, 2023. Disponível em: <https://api.semanticscholar.org/CorpusID:261100919>.

SCHAUL, T. A video game description language for model-based or interactive learning. In: **2013 IEEE Conference on Computational Inteligence in Games (CIG)**. [S.l.: s.n.], 2013. p. 1–8.

SCHRYVER, G.-M. de. Generative AI and Lexicography: The Current State of the Art Using ChatGPT. **International Journal of Lexicography**, v. 36, n. 4, p. 355–387, 10 2023. ISSN 0950-3846. Disponível em: <https://doi.org/10.1093/ijl/ecad021>.

SHAKER, N.; TOGELIUS, J.; NELSON, M. J. **Procedural Content Generation in Games**. Cham: Springer, 2016. ISBN 978-3-319-42715-7.

SHIN, J. et al. Prompt Engineering or Fine-Tuning: An Empirical Assessment of LLMs for Code . In: **2025 IEEE/ACM 22nd International Conference on Mining Software Repositories (MSR)**. Los Alamitos, CA, USA: IEEE Computer Society, 2025. p. 490–502. Disponível em: <https://doi.ieeecomputersociety.org/10.1109/MSR66628.2025.00082>.

STEPHEN, S. J. 'the backrooms': exploring the unconscious together through collective meaning making. **Journal of Psychosocial Studies**, Policy Press, Bristol, UK, v. 15, n. 3, p. 201 – 208, 2022. Disponível em: <https://bristoluniversitypressdigital.com/view/journals/jps/15/3/article-p201.xml>.

TAN, S. et al. Judgebench: A benchmark for evaluating llm-based judges. **ArXiv**, abs/2410.12784, 2024. Disponível em: <https://api.semanticscholar.org/CorpusID:273374769>.

TAYLOR, R. et al. Galactica: A large language model for science. **ArXiv**, abs/2211.09085, 2022. Disponível em: <https://api.semanticscholar.org/CorpusID:253553203>.

TODD, G. et al. Gavel: generating games via evolution and language models. In: **Proceedings of the 38th International Conference on Neural Information Processing Systems**. Red Hook, NY, USA: Curran Associates Inc., 2024. (NIPS '24). ISBN 9798331314385.

TOUVRON, H. et al. Llama: Open and efficient foundation language models. **arXiv preprint arXiv:2302.13971**, 2023.

VASWANI, A. et al. Attention is all you need. In: **Advances in Neural Information Processing Systems (NeurIPS)**. [S.l.: s.n.], 2017. v. 30.

WANG, J. et al. Is chatgpt a good nlg evaluator? a preliminary study. In: . [S.l.: s.n.], 2023. p. 1–11.

WANG, Y. et al. A survey on chatgpt: Ai–generated contents, challenges, and solutions. **IEEE Open Journal of the Computer Society**, v. 4, p. 280–302, 2023.

Wikipedia contributors. **4chan — Wikipedia, The Free Encyclopedia**. 2024. [Online; accessed 25-November-2024]. Disponível em: <https://en.wikipedia.org/w/index.php?title=4chan&oldid=1258415260>.

____. **The Backrooms — Wikipedia, The Free Encyclopedia**. 2024. [Online; accessed 25-November-2024]. Disponível em: <https://en.wikipedia.org/w/index.php?title=The_Backrooms&oldid=1259073807>.

YAO, L. et al. Plan-and-write: towards better automatic storytelling. In: **Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence**. AAAI Press, 2019. (AAAI'19/IAAI'19/EAAI'19). ISBN 978-1-57735-809-1. Disponível em: <https://doi.org/10.1609/aaai.v33i01.33017378>.

YIN, P.; NEUBIG, G. A syntactic neural model for general-purpose code generation. In: **Proc. ACL**. [S.l.: s.n.], 2017. p. 440–450.

ZAWACKI, A. J. Glitches and ghosts: The digital uncanny in video games and creepypasta. **Horror Studies**, Intellect, v. 15, n. 1, p. 83–98, 2024. ISSN 2040-3283. Disponível em: <https://intellectdiscover.com/content/journals/10.1386/host_00082_1>.

ZHANG, P. et al. Tinyllama: An open-source small language model. **ArXiv**, abs/2401.02385, 2024. Disponível em: <https://api.semanticscholar.org/CorpusID:266755802>.

ZHANG, T. et al. Bertscore: Evaluating text generation with bert. In: **International Conference on Learning Representations (ICLR)**. [S.l.: s.n.], 2020.

ZHAO, W. X. et al. **A Survey of Large Language Models**. 2024. Disponível em: <https://arxiv.org/abs/2303.18223>.

ZHAO, Y. The state-of-art applications of nlp: Evidence from chatgpt. **Highlights in Science, Engineering and Technology**, v. 49, p. 237–243, May 2023. Disponível em: <https://drpress.org/ojs/index.php/HSET/article/view/8512>.