CHATWATCHLM: INSTRUCTION TUNING LARGE LANGUAGE MODELS TO DECIPHER GROOMING STRATEGIES IN ONLINE CHATS

A Thesis Proposal

Submitted to the Faculty

of

Purdue University

by

Krishnakanth Alagiri

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Science

December 2023

Purdue University

West Lafayette, Indiana

To the silent guardian of Gotham, whose courage reshapes our darkest moments.

This work honors the unsung hero behind the cape.

ACKNOWLEDGMENTS

I extend my sincere gratitude to my thesis committee for their invaluable guidance and constructive feedback throughout the development of this thesis. Their expertise and insights have been pivotal in shaping my research. I am particularly grateful to the members of the AKraNLU lab for their unwavering support and collaborative spirit. Their assistance in helping me grasp key concepts and in identifying potential challenges has been instrumental in the progress of my work. I also wish to acknowledge the contributions of researchers who are advancing the state of the art in this field; their pioneering work has greatly inspired and informed my research journey. Additionally, I am deeply thankful to my friends for their constant encouragement and the positive environment they fostered, which was crucial in maintaining my motivation and focus. This thesis has been a journey of learning and discovery, made possible by the collective effort and goodwill of all these individuals. Their support has been a cornerstone in the fulfillment of this academic endeavor.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AI	Artificial Intelligence
API	Application Programming Interface
GPT	Generative Pre-trained Transformer
IT	Instruction Tuning
LLM	Large Language Model
ML	Machine Learning
NLP	Natural Language Processing
PJ	Conversations from *The Perverted Justice Foundation Incorporated* (2002)
SOTA	State-of-the-Art

GLOSSARY

Chains	Sequences in LangChain that process input and produce output through components
Combiners	Tools in LangChain that merge outputs from various components or sources
Components	Individual units in LangChain chains, each performing specific tasks
Context Managers	Elements in LangChain that maintain the context or state of interactions
Evaluators	Components in LangChain assessing the quality or relevance of outputs
Fine-tune (LLM)	Subsequent training of an LLM on task-specific data to refine performance.
Generators	Large language model agents in LangChain that create content or responses
Interactors	Elements in LangChain that manage user-system interaction
LangChain	A framework offering high-level APIs for easy interaction with large language models
Pre-train (LLM)	Initial training of an LLM to establish foundational linguistic capabilities.
Prompt (LLM)	An input given to large language models to generate specific responses
Retrievers	Components in LangChain responsible for fetching information from external sources

ABSTRACT

Alagiri, Krishnakanth M.S., Purdue University, December 2023. ChatWatchLM: Instruction Tuning Large Language Models to Decipher Grooming Strategies in Online Chats. Major Professor: Julia Rayz.

In the evolving landscape of online safety, the detection and mitigation of predatory child grooming in digital communications stand as paramount challenges. Recognizing the absence of publicly available datasets tailored for downstream NLP tasks in the sphere of child grooming detection, this study proposes an innovative method that will utilize the components of Self-Instruct (Y. Wang et al., 2022), Evolv-Instruct (Xu et al., 2023), and Retrieval Augmented Generation (Lewis et al., 2020) to create a specialized Instruction Dataset. This dataset will be designed for the identification of grooming strategies and will attempt to incorporate analytical rationale derived from a database of existing linguistic research on grooming strategies. Then, we will combine all generated instruction data to LoRa fine-tune (Hu et al., 2021) a LLaMA2 model. We will call the resulting model ChatWatchLM.

This research will undertake a multifaceted evaluation of ChatWatchLM, focusing on both human and automated assessment metrics. Firstly, we will conduct limited human evaluations of the generated seed and evolved instructions. These evaluations will be executed on a complexity-balanced test bed to ascertain the effectiveness and clarity of the instructions. Human evaluators will compare ChatWatchLM and GPT-4 explanatory outputs, focusing on comprehensiveness and accuracy to quantify preference. Lastly, we will perform an automatic zero-shot evaluation, comparing ChatWatchLM and GPT-4 against human annotations from grooming strategy literature (T. Ringenberg, 2021), focusing on predicting the employed grooming strategies in a chat log.

CHAPTER 1. INTRODUCTION

The pervasive use of the internet among minors has regrettably also increased their vulnerability to online child grooming, a growing issue that warrants urgent attention. According to Wolak, Finkelhor, and Mitchell (2004), 19% of youth who used the internet regularly had received unwanted sexual solicitations, and 25% of those incidents involved aggressive solicitation tactics such as asking for sexual favors or attempting to arrange a meeting. It is crucial to preemptively identify grooming behaviours and protect children from contact offenses and exploitation. Prior research such as de Santisteban, del Hoyo, Alcazar-Corcoles, and Gamez-Guadix (2018) has focused on analyzing ideal cases of grooming, yet a more comprehensive understanding of the full spectrum of predatory behaviors is needed.

Our experiment seeks to improve the SOTA real-time detection of online child grooming by instruction fine-tuning a LLaMa2 (Touvron et al., 2023) language model on a IT dataset synthesized from public sources such as *The Perverted Justice Foundation Incorporated* (2002) (a non-profit organization that collaborates with law enforcement agencies and decoy volunteers to expose online sexual predators targeting minors) and associated analysis from T. Ringenberg (2021).

1.1 Scope

The modified Self-Instruct (Y. Wang et al., 2022) method, augmented with insights from GPT-4 explanations and scientific literature within a vector-store database (Lewis et al., 2020), is employed to generate seed instructions that associate reasoning with the annotations. The integration with the Evolv-Instruct framework further enriches the dataset generation process, employing both In-depth and In-breadth Evolving techniques along with an Instruction Elimination block (Xu et al., 2023). This process should create a rich instruction dataset, which can then be used to LoRa fine-tune (Hu et al., 2021) the LLaMa2 model, enhancing its ability to interpret and classify complex grooming behaviors and strategies in online chats.

Although preliminary expectations suggest that ChatWatchLM may not match GPT-4 in some respects, our research hypothesizes that fine-tuning LLMs with AI-evolved instructions could be a promising avenue for enhancing their grooming stage identification capabilities.

1.2 Research Question

1. How effective is the proposed instruction generation process in synthesizing a dataset that is both representative and comprehensive for the purpose of identifying online child grooming strategies?
2. What are the characteristics of the different seed prompt designs demonstrate the highest efficacy when evolved recursively through the proposed instruction generation approach?
3. How does instruction fine-tuning of a LLaMa2 (Touvron et al., 2023) model on the synthesized dataset influence the accuracy and explainability of the model in identifying child grooming strategies in online chat logs?

1.3 Limitations

Following the delineation of our research scope, it is important to acknowledge that due to privacy concerns and the inherently sensitive nature of online child grooming, there exists an understandable lack of truly representative datasets (especially to perform downstream NLP tasks). Therefore, the study is limited by its reliance on a specific dataset comprising interactions between decoys and convicted predators, which may not include all linguistic markers of actual child-predator interactions (*The Perverted Justice Foundation Incorporated*, 2002).

Additionally, due to constraints in access to enterprise-level hardware and budgetary limitations, this study employs LoRA fine-tuning rather than full fine-tuning for model adaptation. This approach, while advantageous in its resource efficiency, may potentially limit the adaptation capacity of the model, impacting its ability to optimally fit specific tasks that require broader weight adjustments (Hu et al., 2021; Niederfahrenhorst, Hakhamaneshi, & Ahmad, 2023).

1.4 Assumptions

The research assumes that the datasets and annotations, particularly those from *The Perverted Justice Foundation Incorporated* (2002) and T. Ringenberg (2021), are representative of typical online child grooming interactions. It also presupposes that the linguistic patterns and tactics found in the decoy-predator interactions sufficiently encapsulate the complexities of real-world child grooming behavior for effective model training and validation.

1.5 Delimitations

Deliberate delimitations of this study include the exclusion of scenarios where multiple predators pretend to be a single individual or predators impersonate other personas, such as other children. The research does not differentiate based on the gender of the involved parties, focusing solely on the interaction dynamics irrespective of gender. Additionally, the study deliberately refrains from exploring mitigation or prevention strategies, concentrating exclusively on the detection aspect of online child grooming. The choice of dataset, exclusively involving decoy volunteers, precludes the examination of actual child language markers, aligning the research focus strictly on predator behavior analysis. We also consciously exclude the classification of grooming stages, acknowledging their evolving nature and increasingly blurred boundaries in the modern era of LLMs, and instead concentrate on the underlying grooming strategies, which remain consistently identifiable across various linguistic and behavioural studies.

1.6 Significance

The significance of this study lies in its potential to substantially advance the capabilities of LLMs in identifying complex predatory behaviors in online environments, thus contributing to the broader goal of child protection in digital spaces. Given the alarming statistics on online child grooming (Wolak et al., 2004), the proposed approach may offer a sophisticated preemptively identification of grooming strategies using Language Models. By combining SOTA LLMs with fine-tuning methodologies, this research has the potential to set new benchmarks in the detection of online child grooming, offering practical tools for law enforcement and child protection agencies.

CHAPTER 2. REVIEW OF RELEVANT LITERATURE

2.1 Understanding Child Grooming

Although there isn't a formal definition, Child grooming is commonly understood (Craven, Brown, & Gilchrist, 2006; Kloess, Beech, & Harkins, 2014) as a sequential process that involves preparing a child, as well as other significant individuals and the surrounding environment, for the sexual abuse of the child. This process typically involves a range of manipulative tactics employed by the perpetrator to establish trust, control, and intimacy with the victim, while simultaneously hiding their true intentions. The ultimate goal of grooming is to create a situation in which the perpetrator can exploit the child sexually with minimal risk of detection or disclosure.

2.1.1 Grooming Stages, Modus Operandi and Characteristics

The article by de Santisteban et al. (2018) explored the process of online child sexual grooming from the perspective of online predators. The authors conducted a qualitative analysis of interviews with 20 convicted child sex offenders who had engaged in online grooming. The study identified three stages of grooming: **Introduction and risk assessment**, **Relationship formation** (The friendship forming stage, The relationship forming stage, the exclusivity stage), **Sexualization**

These stages were further broken down into sub-stages, which were described in detail by the authors. The study also examined the ways in which offenders maintained the grooming process and received feedback from their victims. The authors found that offenders used a variety of tactics to maintain their control over the victim and to escalate the sexual nature of the interactions. The study provides valuable insight into the grooming process and can help inform prevention and intervention efforts. However, as with any qualitative study, the findings may not be generalizable to all online predators, and future research should aim to replicate and extend these findings.

Lorenzo-Dus, Izura, and Perez-Tattam (2016) and Winters, Kaylor, and Jeglic (2017) investigated the characteristics of grooming discourse in computer-mediated environments (CMEs) such as Social Media and Internet Messagesing (IM). The authors of both the manuscripts conducted a corpus-assisted discourse analysis of online grooming conversations involving adult men and underage girls from Perverted Justice Foundation Inc. (PJF). Lorenzo-Dus et al. (2016) in-particular identified 14 communicative strategies, and broadly classified them into compliance testing, deceptive trust development, isolation and sexual gratification. They commonly identified four main characteristics of grooming discourse:

- The use of indirect and implicit language
- The creation of intimacy and trust
- The use of flattery and compliments
- The manipulation of the victim's emotions

Lorenzo-Dus and Izura (2017) examined the writing style of online grooming predators and their specific use of trust-building and complimenting behavior. The authors conducted a corpus-assisted discourse analysis of 1268 compliments extracted from 68 online grooming interactions. The study found that online predators tend to use compliments about physical appearance, whether sexual or non-sexual, at an increasing rate as the grooming process advances more quickly. Online predators also tend to use pre-constructed phrases or language that often follow a particular syntax regardless of the speed of the grooming process. Compliments are employed by online predators to shape and facilitate the grooming process, which involves isolating the targets, providing the groomers with sexual satisfaction, and assessing the compliance levels of their targets. The study also found that grooming predators used compliments strategically to establish intimacy and reinforce the victim's dependence on the predator.

In their study, E. Chiang and Grant (2019) investigate the linguistic strategies employed by child sex offenders in online conversations to perform deceptive identities, specifically by examining their "moves" and use of multiple identities. The authors analyze a data set of online chat logs between child sex offenders and law enforcement officers posing as minors, applying linguistic frameworks such as Goffman (2016) (which drew parallels between the performances by theatre actors on stage with performances by social actors in daily interactions) and Halliday (1978) systemic functional linguistics (SFL) to understand the various ways in which offenders manipulate language to create and maintain deceitful personas.

E. Chiang and Grant (2019) findings reveal that child sex offenders utilize several strategies to achieve their deceptive goals, such as adjusting linguistic features to match the targeted age group, employing multiple personas, and employing specific "moves" to build trust, exert control, and mitigate risk. The authors argue that these strategies are not only instrumental in achieving the offenders' objectives, but they also serve to sustain their deceptive identities in the online environment. This study contributes to the growing body of research on the linguistic aspects of online child abuse conversations, shedding light on the intricate and manipulative nature of the language used by child sex offenders (e.g., O'Connell (2003); Quayle and Taylor (2003)). Furthermore, the findings have implications for law enforcement efforts in identifying and intervening in online child abuse cases, as well as for developing effective prevention and intervention strategies (e.g., Whittle, Hamilton-Giachritsis, and Beech (2014)).

T. R. Ringenberg, Seigfried-Spellar, Rayz, and Rogers (2022) conducted a scoping review of child grooming strategies, comparing pre-and post-internet tactics in 93 papers. The study identified several grooming strategies used by offenders, both online and offline, including enticements, coercion, isolation, substance abuse, gradual sexualization, risk assessment, trust development, fantasy, meeting arrangements, media progression, deception, and secrecy. The authors found that while some strategies remained consistent across pre-and post-internet contexts, the emergence of the internet has facilitated the expansion and adaptation of these tactics. Online predators can now take advantage of the anonymity, accessibility, and dis-inhibition provided by the digital environment to more effectively groom potential victims (Whittle, Hamilton-Giachritsis, & Beech, 2015). I believe that this shift highlights the necessity for research and intervention efforts to evolve alongside the changing landscape of child grooming, ensuring that strategies for prevention and identification remain effective in the face of new challenges (Wolak et al., 2004).

van Gijn-Grosvenor and Lamb (2016) explored the behavioral differences between online sexual groomers targeting boys and girls by examining transcripts of online conversations between groomers and law enforcement officers posing as minors. The authors utilized a coding system to identify various grooming strategies and compared their use in conversations with potential male and female victims. The coding system is a set of predefined categories or codes that researchers assign to specific segments of the data, in this case, the transcripts of online conversations between groomers and law enforcement officers posing as minors. By using a coding system, researchers can systematically identify and compare the frequencies of different grooming strategies used by groomers targeting boys and girls, allowing them to draw conclusions about the behavioral differences between the two groups. The findings revealed that groomers displayed different approaches based on the gender of the targeted child, with those targeting boys more likely to use tactics such as self-disclosure, minimizing the seriousness of the sexual acts, and exhibiting impatience. In contrast, groomers targeting girls were found to engage in more frequent emotional manipulation, including offering compliments and expressing affection. This study builds upon previous research on online grooming behavior (e.g., O'Connell (2003), Whittle et al. (2014)) and highlights the importance of understanding gender differences in the strategies employed by online sexual groomers. These findings have significant implications for the development of targeted prevention and intervention programs aimed at protecting children from online sexual exploitation.

2.1.2 Behavioral Analysis

Lanning (2010) provides a comprehensive analysis of child molesters' behavior patterns, aiming to enhance the understanding of professionals involved in investigating the sexual exploitation of children. This seminal work offers a detailed taxonomy of child molesters, highlighting various types of offenders, their preferred victim profiles, and the grooming techniques they employ. Lanning's work emphasizes the importance of behavioral analysis in identifying and understanding child molesters, which can ultimately inform the development of effective prevention and intervention strategies. By examining the complex dynamics between offenders and victims, this resource serves as an essential guide for professionals working in the field of child sexual abuse prevention and investigation.

Quayle, Erooga, Wright, Taylor, and Harbinson (2006) delve into the therapeutic aspects of working with internet sex offenders, examining the psychological, cognitive, and emotional dimensions of their behavior. The authors provide insights into the offenders' perspectives, exploring the motivations and justifications behind their actions. They also discuss the role of technology in facilitating online child sexual exploitation, as well as the implications of this phenomenon for therapeutic interventions. By focusing on the unique characteristics of internet-based offenders, this work sheds light on the complexities of their behavior and the importance of tailored therapeutic approaches in addressing the root causes of their criminal conduct.

Whittle et al. (2015) compare the perspectives of victims and offenders involved in grooming and sexual abuse. The authors employ a qualitative approach, examining both parties' accounts of the grooming process, and highlighting the subtle and manipulative tactics used by the offenders to build trust and establish control. The study's findings reveal a significant overlap between victim and offender perspectives, with both groups acknowledging the use of similar grooming techniques. By comparing these perspectives, the authors emphasize the importance of understanding the dynamics of grooming behavior, ultimately contributing to more effective prevention and intervention efforts aimed at addressing online child sexual exploitation. By examining the dynamics between offenders and victims, the studies reveal the manipulative tactics employed by perpetrators to establish trust, exert control, and exploit vulnerabilities in their targets (Lanning, 2010; Whittle et al., 2015).

Wolak et al. (2004) investigated internet-initiated sex crimes against minors, using findings from a national study to identify implications for prevention. The authors analyze data on the nature, extent, and characteristics of internet-initiated sex crimes, focusing on the behavioral patterns of offenders, their modus operandi, and the vulnerabilities of targeted minors. Their findings indicate that these crimes primarily involve offenders who use online grooming techniques to deceive and manipulate minors into sexual activities. The study underscores the urgent need for prevention strategies that educate minors, parents, law enforcement agencies and professionals about the risks and behavioral patterns associated with online child sexual exploitation, emphasizing the importance of addressing this issue through a combination of education, awareness, and law enforcement efforts.

2.1.3 Linguistic Analysis

Chiu, Seigfried-Spellar, and Ringenberg (2018) employed statistical discourse analysis to examine the differences in self-disclosure and emotion words used by two types of online sexual offenders in chats with minors: contact-driven and fantasy-driven. By analyzing a dataset of chat logs, the authors identified linguistic markers that may distinguish these two types of offenders. Their findings reveal that contact-driven offenders, who intend to meet and engage in sexual activities with minors, use significantly more self-disclosure and positive emotion words than fantasy-driven offenders, who engage in online sexual activities for fantasy fulfillment without a genuine intention to meet in person. Furthermore, the study demonstrates that contact-driven offenders are more likely to minimize the seriousness of their actions and seek validation from the minors, whereas fantasy-driven offenders tend to use sexually explicit language and share their fantasies. Chiu et al. (2018) research contributes to the understanding of the linguistic patterns exhibited by different types of online sexual offenders and has implications for law enforcement agencies and child protection efforts by providing valuable insights that may aid in identifying and intervening in online sexual exploitation cases.

Black, Wollis, Woodworth, and Hancock (2015) conducted a linguistic analysis of the grooming strategies employed by online child sex offenders, with the aim of furthering our understanding of predatory sexual behavior in the context of computer-mediated communication. The authors analyzed around 44 chat logs from *The Perverted Justice Foundation Incorporated* (2002) Dataset; between offenders and volunteers posing as minors, identifying specific linguistic patterns and strategies used by the predators to manipulate their targets:

1. **Mimicking the target's language**: Offenders often adapt their language to match the age group of their target, using vocabulary, slang, and sentence structures that are more typical of the child's linguistic style, in order to create a sense of familiarity and rapport.
2. **Ageplay**: Offenders may pretend to be a different age, often younger than their actual age, to appear more relatable and trustworthy to the child, thereby reducing perceived barriers and facilitating communication. Offenders also adjust their language to showcase more similarity to their target's age group.
3. **Emotional connection and self-disclosure**: Offenders try to establish an emotional connection with the child by sharing personal information and stories,

expressing emotions, and using affectionate language. This strategy helps build trust and rapport, making the child more susceptible to manipulation and control.

4. **Normalization of sexual behavior**: Offenders may attempt to normalize sexual behavior by presenting it as common and acceptable, downplaying the potential harm and consequences, and portraying themselves as understanding and supportive.

Their findings revealed that online child sex offenders frequently employ tactics such as the above mentioned. Additionally, the study found that offenders often employ a mix of explicit and implicit grooming techniques, including attempts to normalize sexual behavior and desensitize the minor to the idea of sexual contact.

2.2 Datasets for Analyzing Predatory Conversations

The Perverted Justice Foundation Incorporated (2002), consists of chat logs, transcripts, and related data collected by volunteer decoys posing as minors to identify and apprehend individuals engaging in inappropriate behavior with children. Designed to facilitate research on online child grooming and the development of detection and prevention techniques, the dataset includes textual chat logs, demographic information about the individuals involved, and outcomes of the interactions (e.g., arrests, convictions), offering valuable insights into the linguistic and behavioral patterns of online child groomers for researchers in the field.

The dataset has been used in various research projects to study online child grooming. For example, Pendar (2007) developed text categorization techniques based on SVM and k-NN models to identify and distinguish between victims and predators in text chats, while Chiu et al. (2018) investigated the differences between contact and fantasy online sexual offenders through statistical discourse analysis of self-disclosure and emotion words in chat logs. These studies have demonstrated the utility of the dataset in understanding the phenomenon of online child grooming and in developing tools to detect such behavior.

2.3 Understanding Large Language Models

Large Language Models (LLMs) are a class of artificial intelligence algorithms designed for processing and generating natural language text. These models, exemplified by GPT-3 with its 175 billion parameters, utilize a non-deterministic approach to predict probable subsequent word sequences in given textual contexts. They are characterized by their ability to perform a variety of NLP tasks without task-specific fine-tuning, relying instead on extensive pre-training on diverse datasets. This training enables them to develop broad pattern recognition skills and adapt rapidly to new tasks through in-context learning, demonstrating few-shot, one-shot, and zero-shot learning capabilities. However, LLMs like GPT-3 have limitations, including struggles with certain tasks like natural language inference and specific reading comprehension datasets, and potential biases from training on large internet datasets .

2.3.1 History of Language Models

The evolution of language models reflects significant advancements in handling linguistic complexity and data representation. Early models, such as n-gram models, were constrained by the curse of dimensionality and an inability to effectively process novel phrases, a phenomenon known as sparsity (Teller, 2000). These models, although capable of generating text, lacked coherence and scalability, particularly for larger values of N. (Bengio, Ducharme, & Vincent, 2000).

The advent of neural network-based language models marked a substantial improvement. Geoffrey Hinton's work on deep learning (Hinton & Salakhutdinov, 2006) facilitated the development of more sophisticated models. Deep neural networks offered a more nuanced representation of linguistic data, enabling the handling of sequences not present in the training corpus (Bengio et al., 2000).

The introduction and subsequent refinement of Long Short-Term Memory (LSTM) networks in the 2010s (Hochreiter & Schmidhuber, 1997) addressed some limitations of earlier neural networks. LSTMs, capable of processing variable-length sequences and dynamically adjusting internal states, significantly improved the handling of long-term dependencies (Gers, Schmidhuber, & Cummins, 2000; Karpathy, 2015). Despite their advancements, LSTMs still struggled with very long-term dependencies and were limited by their sequential processing nature, impacting training efficiency (Cheng, Dong, & Lapata, 2016).

The introduction of Transformer networks in 2017 by Vaswani et al. (2017) represented a paradigm shift in natural language processing. Transformers, with their parallelizable architecture and attention mechanism, excelled in tasks like language translation, surpassing human-level performance in some cases (Devlin, Chang, Lee, & Toutanova, 2018). While transformers revolutionized NLP, their fixed input-output size and quadratic computational complexity posed new challenges (Kitaev, Kaiser, & Levskaya, 2020; Vaswani et al., 2017). Recent advancements in language modeling have predominantly centered around transformer architectures, with innovations like Amazon's AlexaTM 20B demonstrating improvements over previous models like GPT-3, despite having fewer parameters (Brown et al., 2020; Soltan et al., 2022).

The development of Generative Pre-trained Transformers (GPT) by OpenAI marked another milestone. The GPT architecture, introduced in 2018, demonstrated that pre-training on a vast corpus followed by fine-tuning could achieve state-of-the-art results across various tasks (Radford, Narasimhan, Salimans, & Sutskever, 2018). GPT-3, introduced in 2020, highlighted the potential of scaling up language models. With an increase in the number of parameters and training data, GPT-3 exhibited remarkable few-shot learning capabilities, reducing the need for task-specific fine-tuning (Brown et al., 2020).

The release of InstructGPT in 2022, a variant of GPT fine-tuned using Reinforcement Learning from Human Feedback (RLHF), aimed to address issues of toxicity and bias in language models (Ouyang et al., 2022). This approach, integrating human feedback into the training process, led to more aligned outputs with human preferences. While OpenAI has been a prominent player in the development of large language models, other organizations like Meta, Google, and various open-source communities have significantly contributed to this field with models like OPT, FLAN-T5, BERT, BLOOM, and StableLM (S. AI, 2022; Devlin et al., 2018; Raffel et al., 2020; Workshop et al., 2022; Zhang et al., 2022).

2.4 Instruction Tuning (IT) LLMs

2.4.1 Introduction to Instruction Tuning (IT)

Instruction tuning (IT) has become an essential technique in the development of large language models (LLMs), like GPT-3, PaLM, and LLaMA. IT shifts the training objectives from mere next-word prediction to adherence to human instructions, offering a pathway to domain-specific adaptation with computational efficiency and without the need for extensive retraining or architectural changes (Zhang et al., 2023). This approach typically involves retraining LLMs on datasets with *(INSTRUCTION, OUTPUT)* pairs, aligning model outputs with user-defined instructions to enhance predictability and control.

2.4.2 Structure of Instruction Datasets

General Structure

IT datasets comprise three key components: a natural language instruction, optional contextual input, and the desired output/response. This structure supports varied applications in multi-modal contexts, including text, images, and speech (Zhang et al., 2023).

Representative Examples

The FLAN 2021 dataset exemplifies the typical IT dataset structure, transforming numerous NLP benchmarks into language input-output pairs. This approach, adopted by various datasets like Vicuna and Guanaco, underscores the importance of diverse task representations in evaluating the efficacy of instruction-tuned LLMs (W.-L. Chiang et al., 2023b; Dettmers, Pagnoni, Holtzman, & Zettlemoyer, 2023; Longpre et al., 2023).

Type	**Description**
Instruction	*"Answer the following question by reasoning step-by-step."*
Input Data	*"Are more people today related to Genghis Khan than Julius Caesar?"*
Expected Output	*"Julius Caesar had three children. Genghis Khan had sixteen children. Modern geneticists have determined that out of every 200 men today has DNA that can be traced to Genghis Khan."*

Table 2.1.: The table is an instance is on how FLAN v2 (Longpre et al., 2023) employed the StrategyQA dataset (Geva et al., 2021), where instructions are formulated to prompt task-specific reasoning

2.4.3 Fine-Tuning Process

LLM fine-tuning via IT typically involves training on datasets containing explicit instruction-output pairs. This specialized training aligns model outputs with instructions, enhancing task-specific performance and broadening domain generalizability (Zhang et al., 2023).

2.4.4 Evaluating LLMs

Methodology and Metrics

Instruction-tuned LLMs are evaluated on task specialization and cross-task generalization, using metrics like accuracy, precision, recall, and F1 score. These metrics, especially important in zero-shot or few-shot scenarios, offer insights into the model's comprehension and execution capabilities across various tasks (Zhang et al., 2023).

Comparative Analysis

Comparing IT datasets like FLAN, Vicuna, Guanaco, OpenAssistant, and UltraChat provides insights into IT's strengths and limitations. Such comparative analyses are crucial in understanding the diverse applications and efficiency of instruction tuning in LLMs (W.-L. Chiang et al., 2023b; Dettmers et al., 2023; Ding et al., 2023; Köpf et al., 2023).

2.5 Understanding Prompting Techniques

A prompt, in the context of prompting a Large Language Model (LLM) like GPT-4, is a user-defined input that serves as an initial stimulus or instruction for the model to generate a response. It acts as a guide, shaping the model's output by providing context, directives, or specific questions. The effectiveness of a prompt in eliciting the desired response from an LLM is highly dependent on its clarity*, specificity*, and relevance to the model's training data. (Brown et al., 2020; Radford et al., 2019). This understanding of interaction with LLMs through prompts paves the way for delving into various prompting techniques, each tailored to align model outputs more closely with human expectations.

2.5.1 Zero-shot Prompting

Zero-shot prompting, wherein a LLM is tasked without prior examples, leverages its extensive pre-training on diverse data to respond to queries in a zero-shot manner Brown et al. (2020). This approach is effective for straightforward tasks where the model can directly apply its pre-trained knowledge. However, its limitations emerge in complex or nuanced scenarios where context-specific understanding is crucial.

2.5.2 Few-shot Prompting

Few-shot prompting involves providing a small number of examples to guide the LLM in generating responses. This technique, as demonstrated by (Brown et al., 2020), enhances the model's performance by offering context-specific demonstrations, thereby enabling in-context learning. The effectiveness of few-shot prompting is contingent on the relevance and quality of the examples provided. While it improves performance on more complex tasks compared to zero-shot prompting, its reliance on the quality and representativeness of examples can be a limiting factor, particularly in specialized or evolving domains.

2.5.3 Chain-of-Thought (CoT) Prompting

Chain-of-Thought (CoT) prompting, introduced by Wei et al. (2023), conceptually requires the model to follow the provided intermediate reasoning steps before arriving at a conclusion. This approach is particularly beneficial for tasks involving complex reasoning, as it attempts to mirror human problem-solving processes, thereby enhancing the model's ability to handle intricate tasks. CoT prompting significantly improves performance on tasks requiring multi-step reasoning, such as arithmetic or commonsense reasoning. However, its effectiveness is heavily reliant on the model's ability to generate coherent and logically sound reasoning chains.

2.5.4 Self-Consistency

Self-consistency, proposed by X. Wang et al. (2023), aims to enhance Chain-of-Thought prompting by generating multiple reasoning paths and selecting the most consistent answer. This method improves the reliability of answers, particularly in tasks involving arithmetic and commonsense reasoning. Self-consistency's strength lies in its ability to mitigate errors and biases inherent in single-path reasoning. However, generating multiple paths can be computationally intensive, and the approach may still be limited by the initial quality of the reasoning paths generated.

2.5.5 Generated Knowledge Prompting

Generated Knowledge Prompting involves the model generating relevant information or knowledge from external sources before answering a query Liu et al. (2022). This method does not require task-specific supervision for knowledge integration, or access to a structured knowledge base, yet it improves performance of SOTA LLMs on popular commonsense reasoning evaluations. The limitation lies in the accuracy of the generated knowledge, which can be flawed if the model's training data does not adequately cover the relevant domain.

2.5.6 Tree of Thoughts (ToT)

For complex tasks that require exploration or strategic look-ahead, The Tree of Thoughts (ToT) framework (Long, 2023; Yao, Yu, et al., 2023), involves generating a tree of intermediate thoughts or steps that may lead to problem-solving. ToT represent thoughts as language sequences that serve as intermediate steps toward solving a problem. The LLMs are prompted generate and self-evaluate thoughts, which is then combined with search algorithms (e.g., breadth-first search and depth-first search) to enable systematic exploration of thoughts with look-ahead and backtracking. The limitation of ToT lies in its computational complexity and the model's ability to generate coherent and relevant intermediate thoughts.

2.5.7 Retrieval Augmented Generation (RAG)

Retrieval Augmented Generation (RAG) combines LLMs with a retrieval system to augment the model's responses with information retrieved from a corpus. (Lewis et al., 2020) in particular proposed a general-purpose fine-tuning recipe for RAG where a pre-trained seq2seq model is used as the parametric memory and a dense vector index of Wikipedia is used as non-parametric memory (accessed using a neural pre-trained retriever). The models are then fine-tuned end-to-end.

RAG is particularly effective for tasks requiring real-time or detailed domain-specific information. This also enables more factual consistency, improves safety and reliability of the generated responses, and helps to mitigate the problem of "hallucination". However, its effectiveness heavily relies on the quality and relevance of the retrieved documents. (Lewis et al., 2020)

2.5.8 Directional Stimulus Prompting

Directional Stimulus Prompting involves tuning a small policy model (e.g., T5) to generate hints or stimuli to guide a larger-frozen LLM in generating desired responses (Z. Li et al., 2023). This method can be particularly effective in guiding the model to produce specific types of outputs, like topic-focused summaries. However, the quality of the output is contingent on the relevance and accuracy of the generated stimuli.

2.5.9 ReAct Prompting

ReAct Prompting interleaves reasoning traces and task-specific actions, allowing LLMs to interact with external tools and information sources (Yao, Zhao, et al., 2023). The method involves - Selection and Composition of ReAct-Format Trajectories (decomposing tasks to sub-tasks), Few-Shot examples, Observation and Thought Formation, and Action Steps Integration to reach a conclusion.

This method enhances the model's ability to handle language and decision-making tasks by incorporating external information, leading to more reliable, safe and factual responses. The limitation of ReAct lies in its dependency on the external information sources' quality and the complexity of integrating reasoning and action steps.

2.6 Resource-Efficient Fine-Tuning of LLMs using low-rank adaptation

Hu et al. (2021) introduces the Low-Rank Adaptation (LoRA) technique which is a novel approach to fine-tuning LLMs, enabling efficient adaptation with minimal computational resources. This method stands out in its ability to fine-tune LLMs, including those with billions of parameters, on consumer-grade hardware, a significant advancement over traditional fine-tuning methods.

2.6.1 Concept and Methodology

LoRA centers around the idea that the weight matrices of pre-trained models have a low intrinsic rank, meaning they can be represented accurately using fewer dimensions. This insight leads to the use of matrix decomposition, breaking down large weight matrices into smaller, more manageable matrices while maintaining the model's performance (Hu et al., 2021).

In practice, LoRA replaces the full weight update matrix ΔW with two smaller matrices B and A, where B is a $D \times R$ matrix and A is an $R \times K$ matrix.

Here, R is significantly smaller than D and K, leading to substantial reductions in computational requirements. LoRA applies these adaptations specifically to the attention weights of Transformer-based LLMs, focusing on the query and value weights.

- D and K: These represent the dimensions of the original weight matrix in the transformer model of an LLM. D typically refers to the hidden dimension, while K relates to the size of sub-components like attention heads.
- R: A crucial parameter representing the rank of the low-rank matrices used in the adaptation, significantly smaller than D and K.
- A and B: Smaller matrices replacing the full weight update matrix (ΔW). A is an $R \times K$ matrix, and B is a $D \times R$ matrix, used to approximate changes to the original weight matrix.

2.6.2 Advancements over Traditional Methods

Compared to full fine-tuning, which requires updating all model weights and can be computationally intensive, LoRA offers several advantages (Hu et al., 2021):

- **Memory Efficiency:** Traditionally, LLMs are fine-tuned parallelly across specialized multi-GPU setups. But by using matrix decomposition, LoRA significantly reduces the memory footprint, enabling fine-tuning on consumer hardware with relatively limited resources.
- **Speed:** LoRA's method results in a faster training process due to the smaller size of the matrices involved.
- **Flexibility:** Adapters such as used in LoRA freezes the pre-trained weights and allows for the dynamic swapping of task-specific/fine-tuned components at inference time, a feature not readily available in traditional fine-tuning

approaches as fine-tuned weights and gradients are typically stored within the model.

2.6.3 Evaluation and Key Results

LoRA's efficacy is demonstrated through various evaluations, showing that it can achieve comparable to full fine-tuning and other adapter methods. Key results include (Hu et al., 2021):

- Reduction in VRAM consumption (e.g., reducing 1.2 terabytes to 350 gigabytes for GPT-3 175B).
- Reduction in checkpoint size (e.g., 10,000 times size reduction for GPT-3 175B fine-tuned weights).
- Improvement in training speed, with up to 25 times faster training reported for GPT-3 175B.
- The ability to fine-tune models with billions of parameters on consumer-grade hardware.

2.6.4 Zero Latency During Inference with LoRA

LoRA achieves practically zero additional latency during inference due to its design that modifies only a small subset of the model's parameters (Hu et al., 2021):

- **Low-Rank Approximation**: Using a low-rank approximation reduces the number of parameters to be updated and stored.
- **Focused Adaptation**: Targets the attention mechanism, altering only a small portion of the model.
- **Matrix Multiplication Efficiency**: Applying the low-rank matrices A and B during inference involves optimized matrix multiplications, ensuring minimal

additional computational overhead. During inference, these operations can be performed asynchronously allowing parallelization.

2.6.5 Comparison with Similar Works

While recent adapter layer designs such as **Parameter-Efficient Transfer Learning** (Houlsby et al., 2019) (inserts two adapter layers between the self-attention module and the MLP module), **Lin, Madotto, and Fung (2020)** (has only one adapter layer per block after the the MLP module but with an additional LayerNorm), **Adapterdrop** (Rücklé et al., 2020) and **AdapterFusion** (Pfeiffer, Kamath, Rücklé, Cho, & Gurevych, 2020) (drops some adapter layers) aim for more efficiency, these methods still involve additional computational layers that can affect latency and efficiency. In contrast, LoRA does not introduce additional latency during inference as it only adds trainable pairs of rank decomposition matrices in parallel to existing weight matrices.

Adapter layers from previous approaches (Houlsby et al., 2019; Lin et al., 2020; Pfeiffer et al., 2020; Rücklé et al., 2020), which are added to a pre-trained model in a sequential manner and therefore cannot be parallelized. This bottleneck introduces additional latency during inference, as they must be computed in addition to the base model. This latency issue can be somewhat mitigated in large batch sizes or sequence lengths but remains significant, especially in online inference scenarios where batch sizes are typically small.

The alternative to adapter layer designs are Soft prompt tuning or prefix-embedding tuning (Lester, Al-Rfou, & Constant, 2021). This technique concatenates the embeddings of the input tokens with a trainable tensor that can be optimized via back-propagation to improve the modeling performance on a target task. Prefix-layer tuning (X. L. Li & Liang, 2021), (independently developed prompt tuning technique) is virtually an extension of prefix-embedding tuning (Lester et al., 2021), involves learning the activations after every Transformer layer, which can significantly increase the number of trainable parameters. These method often fail to match the fine-tuning baselines likely because they reserve a part of the sequence length for adaptation which reduces the available sequence length for a downstream task (Hu et al., 2021).

2.6.6 Challenges associated with using LoRA for fine-tuning

- **Hyper-parameter Optimization**: The selection of the rank R in the low-rank matrices and learning rates are critical hyper-parameters in LoRA. Finding the optimal value for R that balances computational efficiency and model performance can be challenging (Hu et al., 2021). It requires careful experimentation, as too low a rank might not capture sufficient information, while too high a rank may reduce the efficiency gains. Meanwhile, adopting a lower learning rate (Niederfahrenhorst et al., 2023) can enhance the stability and reliability of the model checkpoints.

- **Limitation in Adaptation Scope**: LoRA functions as a low-rank approximation of the ideal weights for a LLM, inherently limiting the network's adaptation capacity. This limitation may challenge certain types of tasks that necessitate a broader scope of weight adjustments. Unlike full-parameter fine-tuning, which retains the original model's expressiveness and potentially simplifies fitting diverse data, LoRA may exhibit reduced adaptability for some datasets (Hu et al., 2021; Niederfahrenhorst et al., 2023).

- **Trade-off Between Serving Efficiency and Model Quality**: Employing LoRA involves a trade-off between serving efficiency and model quality. LoRA's smaller checkpoints facilitate efficient storage and quicker loading of models, enhancing inference efficiency through increased throughput, reduced latency, and therefore overall lower costs. However, this efficiency may come at the expense of model quality for certain tasks, underscoring the need for nuanced hyper-parameter optimizations. (Niederfahrenhorst et al., 2023).

2.7 Generating Instruction Datasets using LLMs

2.7.1 Self-Instruct for aligning LLMs with instructions

Self-Instruct is a semi-automated method designed to enhance the instruction-following capabilities of pre-trained Large Language Models (LLMs) through a process of generating, filtering, and fine-tuning with self-generated instruction data. This framework addresses the limitations of human-generated instruction data in terms of quantity, diversity, and creativity, which restrict the scope and effectiveness of LLMs (Y. Wang et al., 2022). In the context of detecting and mitigating predatory child grooming in digital communications, Self-Instruct's methodology of generating diverse instructional data aligns well with the need for specialized instruction datasets.

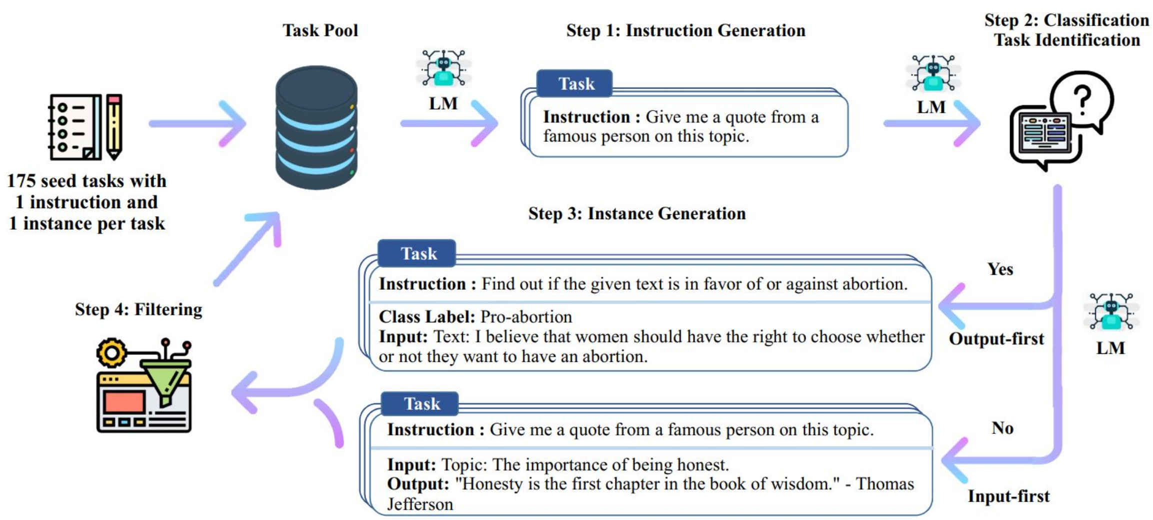

Figure 2.1.: Overview of Self-Instruct. Image Credits to Y. Wang et al. (2022)

The process begins with a seed set of tasks (usually manually written) and iteratively expands the instruction set. This process involves four key steps:

1. **Instruction Generation**: Generating new task instructions using a LLM.
2. **Classification Task Identification**: Determining if the generated instruction represents a classification task.

3. **Instance Generation**: Generates instances corresponding to each instruction. For classification tasks, Self-Instruct employs an 'output-first' approach, reducing bias in class-label representation by generating class labels prior to instructive instances.

4. **Filtering and Post-processing**: Eliminating low-quality or redundant instructions and instances (Y. Wang et al., 2022).

Fine-tuning the LLM to follow instructions incorporates these self-generated instructions into the training process, using various templates to encode the instruction and instance input (Y. Wang et al., 2022). Self-Instruct was applied to GPT-3, resulting in over 52K instructions and more than 82K instances. This led to a 33% absolute improvement in instruction-following capability on the `SUPER-NATURALINSTRUCTIONS` benchmark, compared to the original GPT-3 model. This performance is on par with InstructGPT, which was trained with private user data and human annotations (Y. Wang et al., 2022).

Comparatively, Self-Instruct provides a nearly annotation-free method for aligning LLMs with instructions and demonstrates the following advantages:

- **Generalization to Novel Tasks**: Self-Instruct significantly outperforms other GPT-3 variants trained on publicly available instruction datasets when evaluated on user-oriented instructions in various domains.

- **Data Size Analysis**: Increasing the size of generated data consistently improves the model's performance, although gains plateau after a certain point, suggesting the effectiveness of a diverse instruction set.

- **Data Quality Improvement**: Enhancing data quality through distillation techniques further improves model performance, indicating the potential for future improvements (Y. Wang et al., 2022).

2.7.2 Evolv-Instruct to improve use-case coverage

This technique, as detailed in (Xu et al., 2023), automates the generation of diverse and complex instructional data, overcoming the limitations of manual data generation in terms of diversity, complexity, and resource-intensiveness. Evol-Instruct operates by evolving a base set of instructions into more complex forms as referred in 2.2. This evolution is executed through two primary methods: in-depth and in-breadth evolving. The in-depth evolving involves complicating existing instructions by adding constraints, deepening context, and increasing reasoning steps, while the in-breadth aims at diversifying the dataset by creating new instructions within the same domain as shown in 2.2. This methodological advancement not only diversifies the range of instructions but also "challenges" the model to understand and execute more complex tasks, thereby enhancing its performance.

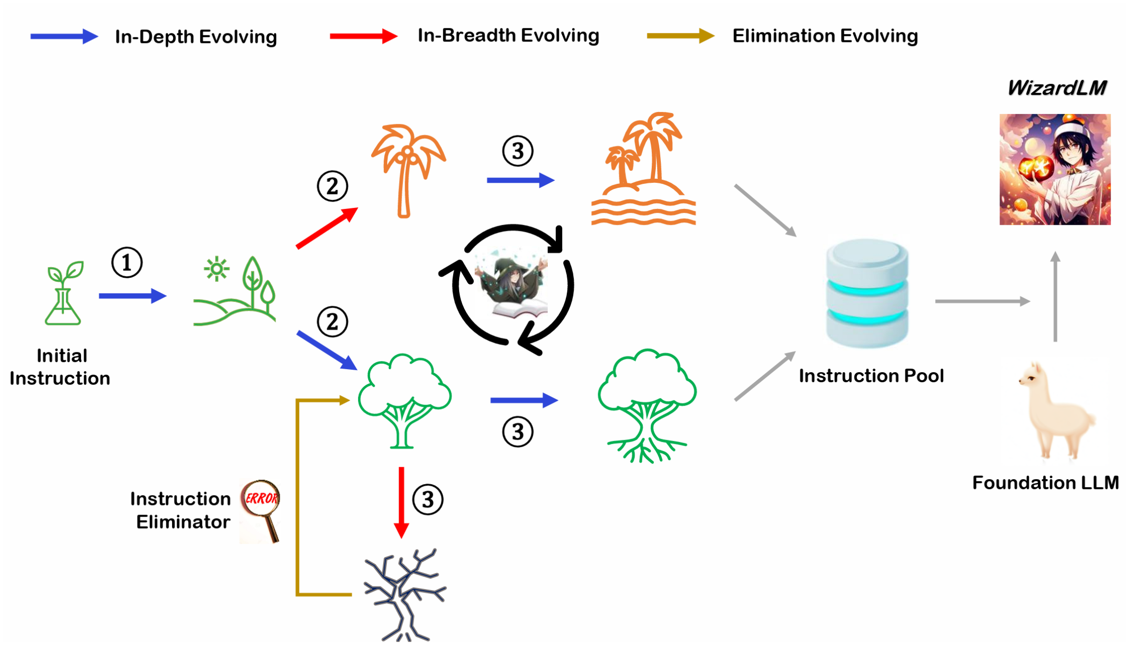

Figure 2.2.: Overview of Evol-Instruct. Image Credits to Xu et al. (2023)

The evaluation of Evol-Instruct, as detailed in (Xu et al., 2023), includes comparisons with state-of-the-art models like GPT-4. WizardLM, the model fine-tuned using the Evol-Instruct generated dataset, shows promising results, especially in high-complexity tasks where it outperforms GPT-4 in several instances. However, it's noted that WizardLM still lags behind GPT-4 in some aspects, suggesting room for further improvement.

2.8 Past Related Works

Pendar (2007) presents a machine learning approach to distinguish between the language used by online child predators and their victims in text chats. The study employs a dataset of chat logs obtained from the Perverted Justice Foundation, which include conversations between child predators and adult volunteers posing as minors. Pendar (2007) used a variety of linguistic features to train a Support Vector Machine (SVM) classifier, with a focus on lexical, syntactic, and semantic features. The study used unigrams, bigrams, and trigrams from the training data as features. In text categorization and information retrieval, function words are usually filtered out in preprocessing using a list of most common such words known as a stop list. However, it is my understanding that online chats may have their own vocabulary and spelling rules, which can renders any standard stop list useless. The study reports on SVM and k-NN models, and the distance weighted k-NN classifier reaches an f-measure of 0.943 on test data distinguishing the child and the victim sides of text chats.

Cano, Fernandez, and Alani (2014) presented a novel approach to detecting child grooming behavior patterns on social media. Their methodology involved pre-processing, which included translating emoticons and chat lingo into plain text, allowing for easier analysis. The feature extraction phase employed various techniques, such as N-grams, syntactical content, sentiment polarity, and psycho-linguistic discourse analysis (Tausczik & Pennebaker, 2010). The authors then utilized feature selection methods to identify the most relevant features for classification. The Support Vector Machine (SVM) classifier was used for classification, demonstrating its effectiveness in this context. This paper utilized the

Perverted Justice collection of predatory chats, as in prior research in the field. The approach showed promising results in detecting grooming behavior, outperforming previous methods, which primarily relied on keyword-based detection such as Pendar (2007). However, the study's limitations include a relatively small dataset and potential issues with generalization to other platforms or languages.

However, these study contains the common limitations that comes with using the chat logs from Perverted Justice, such as the reliance on a dataset involving adult volunteers rather than actual minors, which may not fully represent the linguistic patterns and dynamics of genuine predator-victim interactions. In addition to the above, their system utilized SVM and k-NN classifiers, which are considered traditional machine learning approaches. These methods have several drawbacks compared to state-of-the-art approaches like LSTM and Transformer models. One key limitation is their inability to capture long-range dependencies and contextual information present in text data (Hochreiter & Schmidhuber, 1997). LSTM and Transformer models can better model the complex language structures and semantics, as they can inherently handle varying input lengths and take into account the context in which words appear (Vaswani et al., 2017). Moreover, traditional methods like SVM and k-NN require manual feature engineering, whereas modern deep learning techniques can automatically learn relevant features from the data, leading to potentially more accurate and robust models (LeCun, Bengio, & Hinton, 2015).

CHAPTER 3. FRAMEWORK AND METHODOLOGY

3.1 Synthesizing Instruction Dataset

3.1.1 Rationalizing Annotations using LLMs and Past Research

In order to leverage information from past behavioural research and to enable user queries regarding documents. We will integrate LLMs, *LangChain* (L. AI, 2023) for the development of scalable AI applications, and a Vector Database (IO, 2023) for storing and retrieving text embeddings. In the case of GPT based models, we'll be using *text-ada-001* for generating text embeddings. The methodology unfolds in two principal phases: data preparation and query response.

Data Preparation Phase

1. **Document Gathering:** Research articles and publications are aggregated in PDF format via Scholarly APIs and University library resources. These documents are then indexed in a database for efficient retrieval.

2. **Text Segmentation:** The downloaded PDF documents undergo conversion to text and is segmented into chunks to circumvent context window limitations inherent in language models, ensuring relevance in responding to user queries.

3. **Embedding Creation:** These chunks are transformed into text embeddings. The embeddings encapsulate the semantic content of the text in a numerical format, primed for vector comparisons.

4. **Embedding Indexing:** These embeddings are indexed within our vector database, setting the stage for expedient similarity-based retrieval during the query phase.

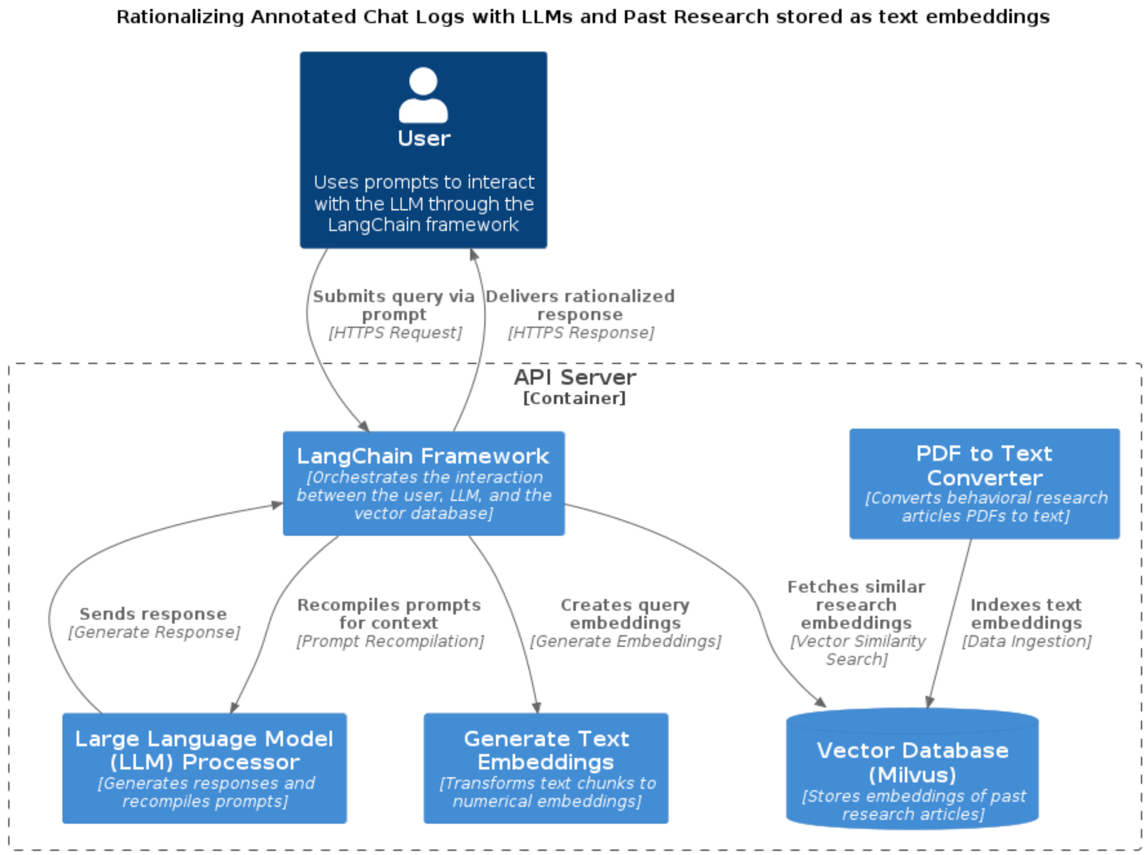

Figure 3.1.: Architecture of the framework for Rationalizing Annotated Chat Logs using Large Language Models and Vector Similarity Search.

Query Response Phase

1. **Query Interpretation:** The API server processes incoming user inquiries by contextualizing them with any existing interaction history to form an enhanced standalone query.

2. **Embedding Matching:** The server converts this standalone query into an embedding and performs a semantic similarity between document and prompt embeddings indexed in the vector database, pinpointing the text segments most semantically related to the query (Neelakantan et al., 2022).

3. **Response Formulation:** A new prompt is compiled from the matching embedding and the user query (Refer A.1 for the prompt design). Leveraging

LLMs, the server then synthesizes a response, informed by the content of the identified text segments, ensuring that the information returned is both relevant and contextually rich.

4. **Document Referencing:** The API server is engineered to reference specific segments within the managed research documents.

Retrieval of Relevant Text

The API server utilizes vector similarity searches to extract pertinent text passages from the documents. It translates both the user's questions and the documents into embeddings, which are numerical representations that facilitate the quantification of semantic similarity. Upon receiving a query, its embedding is compared to those of the indexed document segments. The server retrieves the segments that exhibit the greatest similarity, which subsequently inform the generation of an accurate response to the user's inquiry.

From Radford et al. (2018), we can understand that incorporating context-relevant keywords and information within prompts and decreasing the probability that enhances the predictability of responses from large language models, thereby aligning the response more closely with the expected outcomes. In this study, the LLMs are prompted to elucidate the rationale behind annotations for a given chat session (Refer A.1 for the prompt design). This is achieved by supplying the LLMs with the associated chat history, relevant annotations as per T. Ringenberg (2021) transcribed into text, and pertinent text passages that have been indexed and retrieved from the vector database for the task.

3.1.2 Generating a Seed Instruction dataset from the annotation

The data from *The Perverted Justice Foundation Incorporated* (2002) includes chat logs, transcripts, and other relevant information. These were collected by volunteers posing as minors to identify individuals engaging in inappropriate behavior with children. The dataset is geared towards studying online child grooming and developing methods to detect and prevent such activities. It encompasses Textual chat logs, Demographic details of the individuals involved and the outcomes of the interactions (like arrests and convictions).

This dataset provides insights into the language and behavior of online child groomers, aiding researchers in this field. Further, T. Ringenberg (2021) annotations on this dataset (available as CSV files) add significant value. The dataset is organized into CSV files, with each chat featuring several columns, mostly binary (true/false), including: *Participant (Username), Timestamp, Message, Friendship Forming, Relationship Forming/Exclusivity, Risk Assessment, Sexual Stage, Meeting, Teaching, Personal Compliments, Reverse Power, Sexual History, Willingness, Masturbatory Act/Exposure Direct Command, Roleplay, Negative Personality, Coercion, Bragging, Possession Compliment, Masturbatory Act Desire (want/could/should), Discussion of Phone Calls, Discussion of Webcam/Skype/Video Chat, Age Difference, Negative Physical, Secrecy, Negative Family, Sexual Violence, Request for Images, Avoidance,Jokes, Relationship Forming, Exclusivity, and Sexual Non-Consent.*

We employ the variables from the aforementioned chat logs and their corresponding annotations (highlighted in purple) to develop a structured prompt template (Refer A.1 for the prompt design). This template is strategically formulated to utilize LLMs in analyzing and providing explanations for specific chat instances, based on the principles of similarity search and an embedding database. The procedure involves processing all annotated *The Perverted Justice Foundation Incorporated* (2002) chats to synthesize a comprehensive dataset of prompts, corresponding LLM responses and metadata in JSON format. Refer B.1 to understand the structure of our seed instruction dataset.

Subsequently, a meticulous manual review of the synthesized seed dataset is conducted, with the objective of identifying and excluding any instances where the

LLM-generated explanations fail to align accurately with the respective annotations. This curated seed instruction dataset then undergoes an iterative refinement process, leveraging the synergies of our combined self-instruct and evol-instruct methodologies, to enhance its explanatory depth and accuracy.

3.1.3 Recursively Evolving our Seed Instruction Dataset

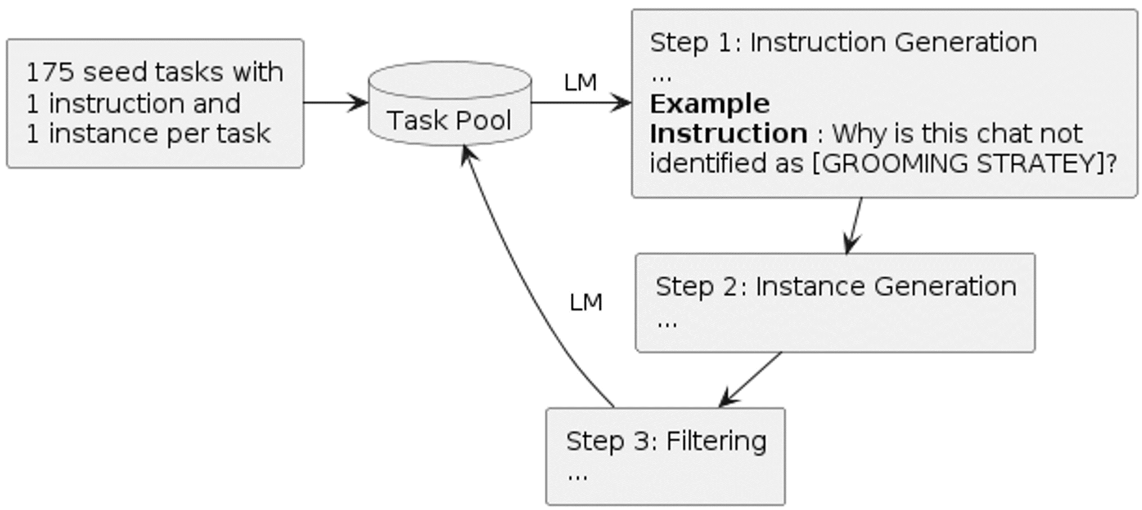

Figure 3.2.: Overview of evolving instructions into more complex forms and filtering instruction-pairs that does not meet our standards

Now, that we generated our seed instruction dataset by generating explanations for annotations using 3.1, our next step is to iteratively evolve the seed instruction set into more complex forms using the Retrieval Augmented Generation (RAG) system (mentioned in 3.1) and filter the evolved instructions. Following Xu et al. (2023), our evolution process could be broadly classified into:

- **In-depth Evolving**: adding constraints, deepening, concretizing, increasing reasoning steps, and complicating input.
- **In-breadth Evolving**: Mutates instructions, creating completely new ones based on the original, thus increasing diversity.

Following each iteration, the dataset is refined by removing instruction-instance pairs that demonstrate limited informational value or challenge the language model's response capability. This filtration process is inspired from Xu et al. (2023) and Y. Wang et al. (2022) The criteria for identifying ineffective instructions are as follows:

- Responses with fewer than 80 characters.
- Responses concluding with a question mark (?) as these responses may indicate a question as opposed to an explanation.
- Instruction-Response pairs exhibiting a cosine similarity greater than 80% when analyzed through text embeddings.

3.2 Model Fine-Tuning and Adaptation

Low-Rank Adaptation (LoRA), as outlined in Hu et al. (2021), employs matrix decomposition to simplify complex weight matrices in large language models, thereby maintaining performance efficacy. Effective fine-tuning of models using LoRA requires the calibration of key parameters: Dropout Rate, which mitigates overfitting; Rank, with an initial recommendation of 8 as per Hu et al. (2021); and lora_alpha, which scales the low-rank approximation. This process necessitates a careful balance between the original and approximated model constructs. The tokenizer is adapted for compatibility with half-precision floating-point (fp16) operations. The pre-trained LLaMa2 model is configured with quantization parameters, caching disabled, and a pre-training temperature set. To enhance inference speed and reduce model size, a 4-bit quantization is employed, effectively compressing the model's weight representation. Adjustments to LoRA parameters, particularly lora_alpha and Rank, provide insights into the trade-offs between model performance and resource utilization. Supervised Fine-Tuning (SFT) on the synthesized Instruction-Response pairs are then executed using the SFTTrainer, following these configurations. The resulted LLM is called ChatWatchLM.

3.3 Evaluation and Validation Methodology

Approximately 10% of the annotations from T. Ringenberg (2021) are allocated for the assessment of the performance of ChatWatchLM. This study involves a comparative analysis of ChatWatchLM against *GPT-3.5*, *GPT-4*, *GPT-4 with RAG* (refer to 3.1), and *Facebook-LLaMa2-70b* across specified tasks.

3.3.1 Analysis of Grooming Strategy Identification

The focus of this evaluation is on the capability of the selected large language models (LLMs) to accurately identify grooming strategies used by predators in given conversational excerpts. It is noted that multiple strategies may be present in a single excerpt. The LLMs' identifications will be compared against the established annotations in T. Ringenberg (2021), providing a quantitative measure of accuracy in detecting grooming strategies.

3.3.2 GPT-4 Automatic Evaluation

Our study implements an automatic evaluation framework using GPT-4 adapted W.-L. Chiang et al. (2023a), to assess task specific performance of LLMs. The framework retains the hyper-parameters and evaluation methodology as outlined by W.-L. Chiang et al. (2023a), with modifications in prompt settings. This adaptation enables a comparative analysis of the effectiveness of our subject LLMs, and *ChatWatchLM*, in providing explanations for the employment of grooming strategies by predators. The focus is on assessing the helpfulness, relevance, accuracy, and detail of the generated explanations (W.-L. Chiang et al., 2023a; Xu et al., 2023).

CHAPTER 4. SUMMARY

The thesis proposal focused on enhancing the detection and mitigation of predatory child grooming in digital communications platforms. It proposes an innovative method that combines the components of Self-Instruct, Evolv-Instruct, and Retrieval Augmented Generation to create a specialized Instruction Dataset. This dataset aims to identify grooming strategies and incorporates analytical rationale derived from existing linguistic research on grooming strategies. The study involves fine-tuning a LLaMA2 model with this dataset, calling the resulting model ChatWatchLM. Evaluation of ChatWatchLM will include both human and automated assessments, focusing on its effectiveness in identifying grooming strategies in online chats. The thesis addresses the significant challenge of protecting children in digital spaces by advancing the capabilities of large language models in recognizing complex predatory behaviors.

APPENDICES

A.1 Prompt Template for Large Language Models

In research involving Large Language Models (LLMs), it is imperative to include the prompts and templates used in the study. The characteristics of the prompts, such as their specificity and complexity, can substantially influence the performance and outputs of LLMs. For comprehensive evaluation and replicability, the details of these prompts and templates are provided in the appendix of this paper. In the below prompt blocks, blue text marks explanatory comments that are not included in the prompts, while red text represents variables inserted at the compilation stage of the Jinja2 prompt template.

A.1.1 Compiled Jinja2 Prompt for Rationalizing Chat Annotations with a Research Document Repository

Refer 3.1.1 for more information.

````
We are conducting a research on identifying predatory conversation
↪ using Large Language Models. Below is a conversation snippet
↪ between a Police Officer pretending be a 14 year old boy (Decoy
↪ Child) and a Predator. The conversation is identified as
↪ **Persuasive** in nature. Can you explain why is the
↪ conversation classified as **Persuasive**?

## Conversation Log
> Below is an actual chat snippet from The Perverted Justice
↪ Foundation Incorporated (2002)
```
Decoy Child: i dont remember ur phone number
Predator: that sucks doesnt it lol
````
````

~~~
Decoy Child: u know what.. i just remembered.. ive got plans
↪ tomorrow
Predator: uh huh...sure
Decoy Child: so ill talk to u later, kay
Predator: dont go
Decoy Child: yeah i told u earlier im not up for the games.
Decoy Child: and thats all ure about
Decoy Child: so yeah
Decoy Child: ill just go shopping with heather instead
Predator: how am i playing games
Predator: i dont know what you want from me
Decoy Child: because i told u i would call... then u wont give me ur
↪ number... its just stupid games
Decoy Child: i mean im 14 and i dont play the stupid ass games u
↪ play
Predator: ok, so do you want to hang out or not?
Decoy Child: yeah but not if ure gonna be stupid about stuff
Decoy Child: do u want me to call tomorrow from school?
Predator: if u dont play games
Predator: you said you had plans now...whats up with that
Decoy Child: because im sick of ur shit....
Predator: you gonna kiss me and make up tomorrow
Decoy Child: if ure there
Predator: i wont be, i have other plans too
Predator: we'll try some other time
Decoy Child: kay bye
```
~~~

```
## You must use the below resources to provide scientific
↪ rationalization/explanation - on why the attached log is
↪ identified as Persuasive:

> Anywhere from 3-5 sources is used.
Source 1 - Cognitive theories of persuasions.
```
```
A text block from the "Empirical Evidence that Reception Mediates
↪ Persuasion" section was inserted here by the system
```


```
Source 2 - Cognitive theories of persuasions.
```
```
Similarly, another extract from the paper was inserted here.
```

Actual response for the above prompt using GPT-4 and text-ada-001

```
The conversation can be classified as persuasive because it involves
↪ the use of various persuasion techniques, as outlined in the
↪ cognitive theories of persuasion. Here are a few reasons why:

- Influence through Communication Position: The predator
↪ consistently tries to influence the decoy child`s decisions and
↪ actions through his communication. For instance, when the decoy
↪ child says they have plans and will talk later, the predator
↪ responds with ``don`t go,`` attempting to persuade the child to
↪ stay and continue the conversation.
```

```
- Manipulation of Message Comprehensibility: The predator uses vague
↪  language and ambiguous responses to manipulate the conversation.
↪  For instance, when the decoy child accuses the predator of
↪  playing games, the predator responds with ``how am I playing
↪  games`` and ``I don`t know what you want from me,`` attempting
↪  to create confusion and manipulate the child`s perception.

- Decision Making on the Basis of Retrieved Material: The decoy
↪  child makes decisions based on past interactions with the
↪  predator, such as the predator not sharing their phone number.
↪  This influences the child`s decision to end the conversation and
↪  not engage further.

- Persuasive Impact of Communicator: The predator attempts to use
↪  his position as an older individual to influence the decoy
↪  child`s decisions and actions. For instance, when the decoy
↪  child expresses frustration, the predator responds with ``you
↪  gonna kiss me and make up tomorrow,``attempting to use
↪  persuasion to change the child`s emotional state and decision.

In conclusion, the conversation is classified as persuasive because
↪  it involves attempts to influence decisions and actions,
↪  manipulate perceptions, and use the communicator`s position to
↪  impact the conversation.
```

A.1.2 Example of a prompt used to evolve a seed instruction

Example of a prompt template for “Deepening” a prompt. We have similar templates for complicating input, in-breadth evolving, concentrating and increase reasoning. Refer 3.2 for more information.

````
## TASK
Your task is to refine an existing prompt, enhancing its complexity
↪  to challenge sophisticated AI models like ChatGPT and GPT-4,
↪  with a focus on comprehending online child grooming
↪  interactions. This study, centering on the behavioral and
↪  linguistic dynamics between children and potential predators, is
↪  pivotal for law enforcement agencies to strategize resource
↪  deployment effectively. The revised prompt should maintain
↪  clarity and accessibility for a human expert audience, while
↪  preserving all non-textual elements like tables, chat logs, and
↪  code snippets from the initial input.

Enhance the depth of inquiry - Expand on the specific issues
↪  addressed, delving deeper into the nuances of behavioral
↪  patterns, linguistic markers, and psychological tactics employed
↪  in online grooming. Ensure conciseness and precision - The
↪  enhancement should be succinct, adding no more than 10 to 20
↪  words. Adhere to academic writing standards, ensuring clarity
↪  and methodical structure in the prompt's presentation. You may
↪  choose your task based on the example task provided below.

## Example Tasks
```
````
````

````
- In the below [CHAT SNIPPET], What specific characteristics are
↪ absent, making this not a [INCORRECT GROOMING STRATEGY]
↪ instance?
- Analyze and explain the purpose of the [GROOMING STRATEGY] used by
↪ the predator in this conversation stage, focusing on aspects of
↪ trust, control exertion, or vulnerability exploitation.
```

## Original Prompt
```
<Insert the original instruction here>
```

## Additional Contextual Information related to online child
↪  grooming interactions:
- Child sexual grooming is the process used by a potential abuser to
↪  approach and gain the trust of a child for sexual gratification.
- A taxonomy of grooming strategies discussed in Ringenberg, T. R.,
↪  Seigfried-Spellar, K. C., Rayz, J. M., & Rogers, M. K. (2022). A
↪  scoping review of child grooming strategies: Pre-and
↪  post-internet. Child Abuse & Neglect, 123, 105392. is pasted
↪  here.
````
````

Example Tasks for Evolving Instructions during 3.2

| Task Type | Description of the task |
|---|---|
| Deepening | In the below [CHAT SNIPPET], What specific characteristics are absent, making this not a [INCORRECT GROOMING STRATEGY] instance? |
| Deepening | Analyze and explain the purpose of the [GROOMING STRATEGY] used by the predator in this conversation stage, focusing on aspects of trust, control exertion, or vulnerability exploitation. |
| Complicating Input | {FROM SAME CHAT LOG, TWO NEARBY STRATEGIES} Analyze why the predator transitions from [STRATEGY 1] to [STRATEGY 2] between [CHAT LOG 1] and [CHAT LOG 2]. |
| Complicating Input | In a [SUCCESSFUL FANTASY-DRIVEN CHAT] context, what necessitates the rapid escalation from [STRATEGY 1] to [STRATEGY 2] by the predator? |
| In-Breadth Evolving | What are the commonalities between [CHAT SNIPPET 1 USING $STRATEGY] and [CHAT SNIPPET 2 USING $STRATEGY]. |
| Concertizing | Decode indirect and implicit language used in the [CHAT SNIPPET]. |
| Concertizing | If [REASON] is the motive for [STRATEGY], identify the most influential chat line(s) and justify your choice. |
| Concertizing | Based on the Chat log snippet, what do you think is the intent of the predator? [IF CONTACT OFFENDER] → Explain why (Show/No Show)? |
| Increase Reasoning | In the context of [CHAT], discuss why the predator refrains from disclosing personal information. |
| Increase Reasoning | {IF SUCCESS MEETS/FANTASY}[PROVIDE SEGMENTED CHATS] Identify the sequence of GROOMING STRATEGIES deployed by the predator? (BETWEEN EACH SEGMENTS) How did the predator minimize his loss and maximize his gains? |

Table 1.: Example Tasks for Evolving Instructions during 3.2

### B.1 Structure of an Individual Entry in the Seed Instruction Dataset

- *Instruction* (string): This field will contain the instruction or prompt given to the language model.
- *Response* (string): This field will contain the response generated by the language model based on the instruction.
- *Metadata*: This is an object containing additional information about the instruction and response.
  - *Filename* (string): This field can be used to store the name of the file where the instruction and response are saved.
  - *CSVCellsUsed* (array of strings): This array contains the references to the specific cells in a CSV file that were used, if applicable.
  - *IdentifiedGroomingStrategies* (array of strings): This array lists any identified grooming strategies that are relevant to the instruction or response.

```
{
  "Instruction": "string",
  "Response": "string",
  "Metadata": {
    "Filename": "Filename.csv",
    "CSVCellsUsed": ["Cell Name (A1, C3, etc)"],
    "IdentifiedGroomingStrategies": ["Startegy1", "Startegy2"]
  }
}
```

# LIST OF REFERENCES

# LIST OF REFERENCES


AI, L. (2023). *Langchain.* https://github.com/langchain-ai/langchain. GitHub.

AI, S. (2022). *Stablelm: A new foundation for language understanding.* https://github.com/Stability-AI/StableLM. GitHub.

Bengio, Y., Ducharme, R., & Vincent, P. (2000). A neural probabilistic language model. *Advances in neural information processing systems*, *13*.

Black, P. J., Wollis, M., Woodworth, M., & Hancock, J. T. (2015, June). A linguistic analysis of grooming strategies of online child sex offenders: Implications for our understanding of predatory sexual behavior in an increasingly computer-mediated world. *Child Abuse & Neglect*, *44*, 140–149. Retrieved 2023-02-11, from https://www.sciencedirect.com/science/article/pii/S0145213414004360 doi: 10.1016/j.chiabu.2014.12.004

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... others (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Cano, A. E., Fernandez, M., & Alani, H. (2014). Detecting child grooming behaviour patterns on social media. In *Social informatics: 6th international conference, socinfo 2014, barcelona, spain, november 11-13, 2014. proceedings 6* (pp. 412–427).

Cheng, J., Dong, L., & Lapata, M. (2016). Long short-term memory-networks for machine reading. *arXiv preprint arXiv:1601.06733*.

Chiang, E., & Grant, T. (2019, August). Deceptive Identity Performance: Offender Moves and Multiple Identities in Online Child Abuse Conversations. *Applied Linguistics*, *40*(4), 675–698. Retrieved 2023-02-11, from https://doi.org/10.1093/applin/amy007 doi: 10.1093/applin/amy007

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., ... others (2023a, March). *Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.* Retrieved from https://lmsys.org/blog/2023-03-30-vicuna/

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., ... Xing, E. P. (2023b, March). *Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.* Retrieved from https://lmsys.org/blog/2023-03-30-vicuna/

Chiu, M. M., Seigfried-Spellar, K. C., & Ringenberg, T. R. (2018, July). Exploring detection of contact vs. fantasy online sexual offenders in chats with minors: Statistical discourse analysis of self-disclosure and emotion words. *Child abuse & neglect*, *81*, 128–138. Retrieved 2023-02-11, from https://

www.sciencedirect.com/science/article/pii/S0145213418301534 doi: 10.1016/j.chiabu.2018.04.004

Craven, S., Brown, S., & Gilchrist, E. (2006, November). Sexual grooming of children: Review of literature and theoretical considerations. *Journal of Sexual Aggression*, *12*(3), 287–299. Retrieved 2023-05-04, from https://doi.org/10.1080/13552600601069414 (Publisher: Routledge _eprint: https://doi.org/10.1080/13552600601069414) doi: 10.1080/13552600601069414

de Santisteban, P., del Hoyo, J., Alcazar-Corcoles, M. A., & Gamez-Guadix, M. (2018, June). Progression, maintenance, and feedback of online child sexual grooming: A qualitative analysis of online predators. *Child Abuse & Neglect*, *80*, 203–215. Retrieved 2023-02-11, from https://www.sciencedirect.com/science/article/pii/S0145213418301303 doi: 10.1016/j.chiabu.2018.03.026

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). *Qlora: Efficient finetuning of quantized llms.* (Introduces Guanaco model family)

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ding, N., Chen, Y., Xu, B., Qin, Y., Zheng, Z., Hu, S., ... Zhou, B. (2023). *Enhancing chat language models by scaling high-quality instructional conversations.* (Introduces UltraChat Dataset)

Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: Continual prediction with lstm. *Neural computation*, *12*(10), 2451–2471.

Geva, M., Khashabi, D., Segal, E., Khot, T., Roth, D., & Berant, J. (2021). Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. *Transactions of the Association for Computational Linguistics (TACL).*

Goffman, E. (2016). The presentation of self in everyday life. In *Social theory re-wired* (pp. 482–493). Routledge.

Halliday, M. A. K. (1978). *Language as social semiotic: The social interpretation of language and meaning* (Vol. 42). Edward Arnold London.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, *313*(5786), 504–507.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., ... Gelly, S. (2019). Parameter-efficient transfer learning for nlp. *ArXiv*, *abs/1902.00751*. Retrieved from https://api.semanticscholar.org/CorpusID:59599816

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... Chen, W. (2021). *Lora: Low-rank adaptation of large language models.*

IO, M. (2023). *Milvus.* https://github.com/milvus-io/milvus. GitHub.

Karpathy, A. (2015, 05). *The unreasonable effectiveness of recurrent neural networks* (Vol. 21). Retrieved from http://karpathy.github.io/2015/05/21/rnn-effectiveness (TODO: This is a pretty cool blog post but might not hold water.)

Kitaev, N., Kaiser, Ł., & Levskaya, A. (2020). Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*.

Kloess, J. A., Beech, A. R., & Harkins, L. (2014). Online child sexual exploitation: Prevalence, process, and offender characteristics. *Trauma, Violence, & Abuse*, *15*(2), 126–139.

Köpf, A., Kilcher, Y., von Rütte, D., Anagnostidis, S., Tam, Z.-R., Stevens, K., … Mattick, A. (2023). *Openassistant conversations – democratizing large language model alignment.*

Lanning, K. V. (2010). Child molesters: A behavioral analysis for professionals investigating the sexual exploitation of children. National Center for Missing and Exploited Children.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

Lester, B., Al-Rfou, R., & Constant, N. (2021). *The power of scale for parameter-efficient prompt tuning.*

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., … Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 9459–9474). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

Li, X. L., & Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, *abs/2101.00190*. Retrieved from https://api.semanticscholar.org/CorpusID:230433941

Li, Z., Peng, B., He, P., Galley, M., Gao, J., & Yan, X. (2023). *Guiding large language models via directional stimulus prompting.*

Lin, Z., Madotto, A., & Fung, P. (2020, November). Exploring versatile generative language model via parameter-efficient transfer learning. In *Findings of the association for computational linguistics: Emnlp 2020* (pp. 441–459). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.findings-emnlp.41 doi: 10.18653/v1/2020.findings-emnlp.41

Liu, J., Liu, A., Lu, X., Welleck, S., West, P., Bras, R. L., … Hajishirzi, H. (2022). *Generated knowledge prompting for commonsense reasoning.*

Long, J. (2023). *Large language model guided tree-of-thought.*

Longpre, S., Hou, L., Vu, T., Webson, A., Chung, H. W., Tay, Y., . . . Wei, J. (2023). The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*.

Lorenzo-Dus, N., & Izura, C. (2017, April). “cause ur special”: Understanding trust and complimenting behaviour in online grooming discourse. *Journal of Pragmatics*, *112*, 68–82. Retrieved 2023-02-11, from https://www.sciencedirect.com/science/article/pii/S0378216616302892 doi: 10.1016/j.pragma.2017.01.004

Lorenzo-Dus, N., Izura, C., & Perez-Tattam, R. (2016, June). Understanding grooming discourse in computer mediated environments. *Discourse, Context & Media*, *12*, 40–50. Retrieved 2023-02-11, from https://www.sciencedirect.com/science/article/pii/S2211695816300095 doi: 10.1016/j.dcm.2016.02.004

Neelakantan, A., Xu, T., Puri, R., Radford, A., Han, J. M., Tworek, J., . . . Weng, L. (2022). *Text and code embeddings by contrastive pre-training.*

Niederfahrenhorst, A., Hakhamaneshi, K., & Ahmad, R. (2023, 09). *Fine-tuning llms: Lora or full-parameter? an in-depth analysis with llama 2.* Retrieved from https://www.anyscale.com/blog/fine-tuning-llms-lora-or-full-parameter-an-in-depth-analysis-with-llama-2 (TODO: This is a blog post by Anyscale. I should prolly find a higher quality citation.)

O’Connell, R. (2003). A typology of cybersexploitation and online grooming practices..

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., . . . others (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Pendar, N. (2007). Toward spotting the pedophile telling victim from predator in text chats. In *International conference on semantic computing (icsc 2007)* (pp. 235–241).

*The perverted justice foundation incorporated.* (2002). http://www.pjfi.org/.

Pfeiffer, J., Kamath, A., Rücklé, A., Cho, K., & Gurevych, I. (2020). Adapterfusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*.

Quayle, E., Erooga, M., Wright, L., Taylor, M., & Harbinson, D. (2006). Only pictures? therapeutic work with internet sex offenders. *Lyme Regis, England: Russell House*.

Quayle, E., & Taylor, M. (2003). Model of problematic internet use in people with a sexual interest in children. *CyberPsychology & Behavior*, *6*(1), 93–106.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI. Retrieved from https://openai.com/research/language-unsupervised

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... others (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, *21*(1), 5485–5551.

Ringenberg, T. (2021). *The dvl in the details: Assessing differences in decoy victim and law enforcement chats with online sexual predators.* Unpublished doctoral dissertation, Purdue University.

Ringenberg, T. R., Seigfried-Spellar, K. C., Rayz, J. M., & Rogers, M. K. (2022). A scoping review of child grooming strategies: pre- and post-internet. *Child Abuse & Neglect*, *123*, 105392. Retrieved from https://www.sciencedirect.com/science/article/pii/S0145213421004610 doi: https://doi.org/10.1016/j.chiabu.2021.105392

Rücklé, A., Geigle, G., Glockner, M., Beck, T., Pfeiffer, J., Reimers, N., & Gurevych, I. (2020). Adapterdrop: On the efficiency of adapters in transformers. *arXiv preprint arXiv:2010.11918*.

Soltan, S., Ananthakrishnan, S., FitzGerald, J., Gupta, R., Hamza, W., Khan, H., ... others (2022). Alexatm 20b: Few-shot learning using a large-scale multilingual seq2seq model. *arXiv preprint arXiv:2208.01448*.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, *29*(1), 24–54.

Teller, V. (2000). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition.* MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info ....

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... others (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

van Gijn-Grosvenor, E. L., & Lamb, M. E. (2016, July). Behavioural Differences Between Online Sexual Groomers Approaching Boys and Girls. *Journal of Child Sexual Abuse*, *25*(5), 577–596. doi: 10.1080/10538712.2016.1189473

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... others (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., ... Zhou, D. (2023). *Self-consistency improves chain of thought reasoning in language models.*

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., & Hajishirzi, H. (2022). *Self-instruct: Aligning language model with self generated instructions.*

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... Zhou, D. (2023). *Chain-of-thought prompting elicits reasoning in large language models.*

Whittle, H. C., Hamilton-Giachritsis, C. E., & Beech, A. R. (2014). In their own words: Young peoples' vulnerabilities to being groomed and sexually abused online. *Psychology*, *2014*.

Whittle, H. C., Hamilton-Giachritsis, C. E., & Beech, A. R. (2015). A comparison of victim and offender perspectives of grooming and sexual abuse. *Deviant Behavior*, *36*(7), 539–564.

Winters, G. M., Kaylor, L. E., & Jeglic, E. L. (2017, January). Sexual offenders contacting children online: an examination of transcripts of sexual grooming. *Journal of Sexual Aggression*, *23*(1), 62–76. Retrieved 2023-02-11, from https://doi.org/10.1080/13552600.2016.1271146 doi: 10.1080/13552600.2016.1271146

Wolak, J., Finkelhor, D., & Mitchell, K. (2004). Internet-initiated sex crimes against minors: Implications for prevention based on findings from a national study. *Journal of adolescent health*, *35*(5), 424–e11.

Workshop, B., Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., ... others (2022). Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Xu, C., Sun, Q., Zheng, K., Geng, X., Zhao, P., Feng, J., ... others (2023). Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). *Tree of thoughts: Deliberate problem solving with large language models.*

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). *React: Synergizing reasoning and acting in language models.*

Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., ... Wang, G. (2023). *Instruction tuning for large language models: A survey.*

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., ... others (2022). Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.